On Indiscernibility Relations for Missing Attribute Values

Rafał Latkowski

Warsaw University, Institute of Computer Science,
ul. Banacha 2, 02-097 Warszawa, Poland,
R.Latkowski@mimuw.edu.pl

Abstract. The indiscernibility relation is a fundamental concept of the rough set theory. The original definition of the indiscernibility relation, thus the rough set theory, does not capture the situation where some of the attribute values are missing. This paper tries to enhance former works by proposing an individual treatment of missing values at the attribute or value level. The main assumption of the theses presented in this paper considers that not all missing values are semantically equal. We propose two different approaches to create an individual indiscernibility relation for a particular information system. The first relation assumes variable, but fixed semantics of missing attribute values in different columns. The second relation assumes different semantics of missing attribute values, although this variability is limited with expressive power of formulas utilizing descriptors.

1 Introduction

The indiscernibility relation is a fundamental concept of the rough set theory. The original definition of the indiscernibility relation, thus the rough set theory, does not capture the situation where some of the attribute values are missing. The problem of missing values handling within the rough set framework has been already faced in literature, e.g., by Grzymała [3], Słowiński [7] and Stefanowski [9]. The proposed approaches consider alternative definitions of the indiscernibility relation, which reflect various semantics of missing attribute values. The main difficulty of applied alternatives for the indiscernibility relation arise from semantics fixed in advance of all missing values in whole information system, what was identified in, e.g., Stefanowski [9]. This paper tries to enhance former works by proposing an individual treatment of missing values at the attribute or value level.

The main assumption of the theses presented in this paper considers that not all missing values are semantically equal. Among a number taxonomies (see, e.g., [1, 5, 9]) for missing attribute value semantics, the two main types of missing values can be determined: the existential null as an unknown value of considered property, called also "missing" semantics and the placeholder null as an inapplicable value for considered property, what is similar to the "absent" missing value semantics. These two main types of missing attribute values possibly can be even mixed together in one database column, in a way precluding the distinguishing of one type from another. The different meanings of missing attribute values obviously have an impact on the concept of the indiscernibility relation and in consequence on the concept of certain and approximate decision rules. We expect the decision rules induced with help of an indiscernibility relation customized to a particular decision system to perform better in terms of knowledge discovery and classification accuracy.

In this paper we propose two different approaches to create an individual indiscernibility relation for a particular information system. The first relation assumes variable, but fixed semantics of missing attribute values in different columns. The second relation assumes different semantics of missing attribute values, although this variability is limited with expressive power of formulas utilizing descriptors.

2 Preliminaries

The concept of discernibility or similarity is very essential not only for the rough set theory, but also for all other aspects of reasoning. Its importance arise from the fact, that almost every other concept utilized in reasoning and especially in machine learning depends on the similarity or discernibility. For example, if the reasoning process is carried out based on some objects then it is necessary to know which objects are discernible and which are not. The other example is the decision rule matching. Before applying a decision rule, it has to be compared to an object, in order to determine does the object is somehow similar enough to the decision rule. Also a decision rule should be identically applicable to objects that are indiscernible. The semantic of the indiscernibility relation impact on soundness of reasoning.

The indiscernibility relation is formulated on objects belonging to an *information systems* (see, e.g., [4, 6]). An information system is a pair $\mathbb{A} = (U, A)$ where U is a finite set of objects and A is a finite set of attributes. Attributes $a_i \in A$ are functions $a_i : U \to V_i$, where V_i is a domain of attribute a_i. In a presence of missing data we may consider the attributes $a_i \in A$ as functions $a_i : U \to V_i^*$, where $V_i^* = V_i \cup \{*\}$ and $* \notin V_i$. The special symbol "$*$" denotes absence of regular attribute value and if $a_i(x) = *$ we say that a_i is not defined on x. In the relational databases exists a similar notion — "NULL" that represents missing attribute value in a database record (see, e.g., [2, 5]).

The classic indiscernibility relation is formulated for information systems where all attribute values are present (cf. [4, 6]).

$$\mathrm{IND}_{\mathbb{A}} = \{(x, y) \in U \times U : \forall_{a \in A}\; a(x) = a(y)\} \tag{1}$$

The above relation is an equivalence relation, which usually does not hold for other indiscernibility relations. Due to the natural extension of the equality relation to the additional symbol $*$, where $* = *$ and no other domain value is equal to the $*$, we obtain a natural extension of the classic indiscernibility relation IND to the case where some attribute values are missing. The $IND_{\mathbb{A}}$ is the smallest indiscernibility relation, making a common assumption, that identical objects should be indiscernible.

It is known fact, that missing value handling by IND relation decrease the correctness of inductive reasoning. To overcome this problem some other indiscernibility relations were proposed for an alternative missing values handling within the rough set framework. However, none of them is universally the best nor always correct. The two most important are *symmetrical similarity* relation and *unsymmetrical similarity* relation.

$$\mathrm{SS}_{\mathbb{A}} = \{(x,y) \in U \times U : \forall_{a \in A}\ a(x) = a(y) \vee a(x) = * \vee a(y) = *\} \tag{2}$$

$$\mathrm{US}_{\mathbb{A}} = \{(x,y) \in U \times U : \forall_{a \in A}\ a(x) = a(y) \vee a(x) = *\} \tag{3}$$

It is easy to observe, that for any information system $\mathbb{A}$ following property holds (cf. [9]):

$$\mathrm{IND}_{\mathbb{A}} \subseteq \mathrm{US}_{\mathbb{A}} \subseteq \mathrm{SS}_{\mathbb{A}}. \tag{4}$$

Moreover, any indiscernibility relation that does not join two different domain values is bounded by above property between the classic indiscernibility relation and the symmetrical similarity relation.

3 Flexible Indiscernibility Relations

The main difficulty of applied alternatives for the indiscernibility relation arise from semantics fixed in advance of all missing values in whole information system. It was already observed, that presented above indiscernibility relations SS and US have some deficiencies in creating relevant and big enough upper and lower approximations of considered concept (see, e.g., [9]). To overcome this problem also some other approaches were proposed, where the additional numerical tuning of the indiscernibility relation is made (see, e.g., [9, 10]). The common problem of all these approaches is lack of algorithm that selects parameters or shapes of fuzzy membership functions optimal (in some sense) for the considered information system.

In this paper we try to find another way to provide a flexible indiscernibility relation by using logical approach. It means that the indiscernibility relation should be expressed as a logical formula without any additional numeric parameters. Although at this stage of research we do not provide ready algorithm for finding such a formula, we believe that indiscernibility relation based on logical formula would be easier for automatic generation or induction using boolean reasoning.

We propose two different approaches to create an individual indiscernibility relation for a particular information system. The first relation assumes variable, but fixed semantics of missing attribute values in different columns. The second relation assumes different semantics of missing attribute values, although this variability is limited with expressive power of formulas utilizing descriptors.

3.1 Attribute Limited Indiscernibility Relation

The attribute limited indiscernibility relation or ALIR for short allows utilizing different missing value semantics for each attribute. To better explain the application area of such a relation let take an example of information system $\mathbb{A} = (U, \{c, w, p, ec, t\})$, containing descriptions of motorcycles and bicycles. Motorcycles and bicycles both have color (c), weight (w) and price (p). However, the engine capacity (ec) is a property, which does not make any sense in case of bicycles. In such an example the missing values in color, weight and price can be treated as existential nulls using "missing" semantics, while missing values in engine capacity can be treated as placeholder null using "absent" semantics. The simplest ALIR formula representing the above example can be:

$$\begin{aligned}\mathrm{AL}_{\mathbb{A}}(x,y) = {} & (c(x) = c(y) \vee c(x) = * \vee c(y) = *) \\ & \wedge\ (w(x) = w(y) \vee w(x) = * \vee w(y) = *) \\ & \wedge\ (p(x) = p(y) \vee p(x) = * \vee p(y) = *) \\ & \wedge\ ec(x) = ec(y) \wedge t(x) = t(y), \end{aligned} \tag{5}$$

where $t : u \rightarrow \{b, m\}$ is an attribute describing type of object: for motorcycle $t(u) = m$ and for bicycle $t(u) = b$. Let also assume for simplicity that type attribute t does not contain missing values. The relation $\mathrm{AL}_{\mathbb{A}}$ implements for attributes c,w and p the existential missing value semantics, while for attribute ec the placeholder missing value semantic. It is easy to observe, that for any information system $\mathbb{B}$ following property holds:

$$\mathrm{IND}_{\mathbb{B}} \subseteq \mathrm{AL}_{\mathbb{B}} \subseteq \mathrm{SS}_{\mathbb{B}}. \tag{6}$$

However, relation $\mathrm{AL}_{\mathbb{B}}$ is not comparable with relation $\mathrm{US}_{\mathbb{B}}$.

3.2 Descriptor Limited Indiscernibility Relation

The descriptor limited indiscernibility relation or DLIR for short, gives more flexibility than ALIRs. In this case the relation in not limited to fixed missing value semantics for an attribute, but the relation can be described with any propositional logic formula over descriptors from information system. Continuing the above example, let's consider that the values of engine capacity can be also missing in case of motorcycles, what should be treated as existential null rather than placeholder null. The simplest DLIR formula representing such a relation can be:

$$\begin{aligned}\mathrm{DL}_{\mathbb{A}}(x,y) = {} & (c(x) = c(y) \vee c(x) = * \vee c(y) = *) \\ & \wedge\ (w(x) = w(y) \vee w(x) = * \vee w(y) = *) \\ & \wedge\ (p(x) = p(y) \vee p(x) = * \vee p(y) = *) \\ & \wedge\ (ec(x) = ec(y) \vee (t(x) = m \wedge t(y) = m \wedge \\ & \quad (ec(x) = * \vee ec(y) = *))) \\ & \wedge\ (t(x) = t(y)). \end{aligned} \tag{7}$$

The relation $\mathrm{DL}_{\mathbb{A}}$ implements for attributes c,w and p the existential missing value semantics as well as for the attribute ecc, when both objects are motorcycles. If one or two of the considered objects are bicycles, then relation $\mathrm{DL}_{\mathbb{A}}$ implements for attribute ec the placeholder missing value semantic. It is easy to observe, that for any information system $\mathbb{B}$ following property holds:

$$\mathrm{IND}_{\mathbb{B}} \subseteq \mathrm{AL}_{\mathbb{B}} \subseteq \mathrm{DL}_{\mathbb{B}} \subseteq \mathrm{SS}_{\mathbb{B}}. \tag{8}$$

However, relation $\mathrm{DL}_{\mathbb{B}}$ is not comparable with relation $\mathrm{US}_{\mathbb{B}}$ and above property does not hold for any attribute and descriptor limited indiscernibility relations.

3.3 Free Indiscernibility Relation

There exists the possibility to create a free indiscernibility relation which would not be limited to the attribute nor descriptor expressive power. Such a relation gives an opportunity to capture all possible relationships between objects considered in information system and semantics of missing attribute values. However, exceeding the limits of expressive power of propositional formulae language over descriptors precludes usability of such a relation. Without the description of indiscernibility relation formulated in language easily decidable we are not able to apply such relation correctly to new, unseen objects. If the relation does not contain any (decidable) description, then the only way to characterize it is the enumeration of elements in relation, e.g., in form of relation matrix. If the matrix does not contain unseen objects, than we are not able to determine whether the particular new object is in the relation with any other or is not.

4 Conclusions and Further Work

The presented two approaches for constructing flexible and customizable for a considered information system indiscernibility relations provide a foundation for considering the problem of fitting an indiscernibility relation to an information system. The flexibility in selecting any indiscernibility relation between classic indiscernibility relation and symmetrical similarity is limited by some assumptions. This property provides an opportunity to efficiently search for optimal solution for this problem. The goal of introducing these relations is improvement in reasoning from data with missing attribute values.

The attribute limited indiscernibility relation is simpler in its construction and is limited by much stronger assumption. From one point of view this gives less flexibility, but from the other it should be very easy to construct an algorithm that search for such a relation. In this case the cardinality search space is equal to $2^{\mathrm{card}(A)}$.

The descriptor limited indiscernibility relation is more complex as it is limited by weaker assumption. This gives a lot of flexibility, but makes the searching for an optimal (somehow) relation more difficult. Perhaps the efficient algorithms that search for descriptor limited indiscernibility relation will be searching only in a special family of such relations, to keep the computations in reasonable time.

To get a complete solution for this problem we have to do some further work. The most important issue, apart from the optimality criterion, is the construction of an algorithm that searches for an optimal indiscernibility criterion. Such an algorithm has to meet some computational requirements in order to be applicable for real classification problems. The other issue is related with classifier induction. Most of the rough set concepts, such as lower and upper approximations or reducts, are naturally extensible to the case with an arbitrary indiscernibility relation. However, even if the classifier induction algorithms, in particular decision rule induction algorithms (see, e.g., [4, 8]), are easily extensible to the case with an arbitrary indiscernibility relation, their current implementations are not. Apart from that, there are several decision rule induction algorithms that implicitly utilizes classic indiscernibility relation or symmetrical similarity. Therefore it is necessary to extend decision rule induction algorithms to the case with an arbitrary indiscernibility relation.

Acknowledgments

The author would like to thank professor Andrzej Skowron for his support in this research. The research has been supported by the grant 3T11C00226 from Ministry of Scientific Research and Information Technology of the Republic of Poland.

References

1. Candan, K.S., Grant, J., Subrahmanian, V.S.: A unified treatment of null values using constraints. Information Sciences **98** (1997) 99–156
2. Codd, E.F.: Understanding relations (installment #7). FDT - Bulletin of ACM SIGMOD **7** (1975) 23–28
3. Grzymała-Busse, J.W.: On the unknown attribute values in learning from examples. In: Proc. of Int. Symp. on Methodologies for Intelligent Systems. (1991) 368–377
4. Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: A tutorial. In Pal, S.K., Skowron, A., eds.: Rough Fuzzy Hybridization. A New Trend in Decision Making, Singapore, Springer (1999) 3–98
5. Lipski, W.J.: On semantic issues connected with incomplete information databases. ACM Transactions on Database Systems **4** (1979) 262–296
6. Pawlak, Z.: Rough sets: Theoretical aspects of reasoning about data. Kluwer, Dordrecht (1991)
7. Słowiński, R., Stefanowski, J.: Rough classification in incomplete information systems. Math. Computing Modelling **12** (1989) 1347–1357
8. Stefanowski, J.: On rough set based approaches to induction of decision rules. In Polkowski, L., Skowron, A., eds.: Rough Sets in Knowledge Discovery 1: Methodology and Applications, Physica-Verlag (1998) 500–529
9. Stefanowski, J., Tsoukiàs, A.: On the extension of rough sets under incomplete information. In Zhong, N., Skowron, A., Ohsuga, S., eds.: Proceedings of the RSFDGrC '99. LNCS 1711, Springer (1999) 73–81
10. Stefanowski, J., Tsoukiàs, A.: Incomplete information tables and rough classification. International Journal of Computational Intelligence **17** (2001) 545–566