Different Types of Search Algorithms for Rough Sets

Dávid Nagy, Tamás Mihálydeák, László Aszalós

Abstract:

Based on the available information in many cases it can happen that two objects cannot be distinguished. If a set of data is given and in this set two objects have the same attribute values, then these two objects are called indiscernible. This indiscernibility has an effect on the membership relation, because in some cases it makes our judgment uncertain about a given object. The uncertainty appears because if something about an object is needed to be stated, then all the objects that are indiscernible from the given object must be taken into consideration. The indiscernibility relation is an equivalence relation which represents background knowledge embedded in an information system. In a Pawlakian system this relation is used in set approximation. Correlation clustering is a clustering technique which generates a partition using search algorithms. In the authors' previous research the possible usage of the correlation clustering in rough set theory was investigated. In this paper the authors show how different types of search algorithms affect the set approximation.

Keywords: rough set theory, set approximation, data mining

Introduction

In many computer science applications objects are stored in a database. Each of these objects has a unique ID and other attributes. The ID of an object is only a technical tool and it does not represent any information about the object itself. In practice, two objects can only be distinguished if they differ in at least one known attribute value. If we want to decide whether an object belongs to an arbitrary set, based on the available data, then our decision affects the decision about all the objects that are indiscernible from the given object.

In this case if we would like to check whether an object is in an arbitrary set then the following three possibility appear:

- it is sure that the object is in the set if all the objects, that are indiscernible from the given object, are in the set;
- the object may be in the set if there some objects that are in the set and are indiscernible from the given object;
- it is sure that the object is not in the set if all the objects, that are indiscernible from the given object, are not in the set.

So the indiscernibility makes a set vague. The relation and the set theory based on it was developed by professor Pawlak.

From the theoretical point of view a Pawlakian approximation space (see in [1, 3, 2]) can be characterized by an ordered pair $\langle U, \mathcal{R} \rangle$ where U is a nonempty set of objects and $\mathcal{R}$ is an equivalence relation on U. In order to approximate an arbitrary subset S of U the following tools have to be introduced:

- *the set of base sets*: $\mathfrak{B} = \{B \mid B \subseteq U, \text{ and } x, y \in B \text{ if } x\mathcal{R}y\}$, the partition of U generated by the equivalence relation $\mathcal{R}$;
- *the set of definable sets*: $\mathfrak{D}_{\mathfrak{B}}$ is an extension of $\mathfrak{B}$, and it is given by the following inductive definition:
  1. $\mathfrak{B} \subseteq \mathfrak{D}_{\mathfrak{B}}$;
  2. $\emptyset \in \mathfrak{D}_{\mathfrak{B}}$;
  3. if $D_1, D_2 \in \mathfrak{D}_{\mathfrak{B}}$, then $D_1 \cup D_2 \in \mathfrak{D}_{\mathfrak{B}}$.
- *the functions* l, u form a Pawlakian approximation pair $\langle \mathsf{l}, \mathsf{u} \rangle$, i.e.
  1. $Dom(\mathsf{l}) = Dom(\mathsf{u}) = 2^U$
  2. $\mathsf{l}(S) = \bigcup \{B \mid B \in \mathfrak{B} \text{ and } B \subseteq S\}$;

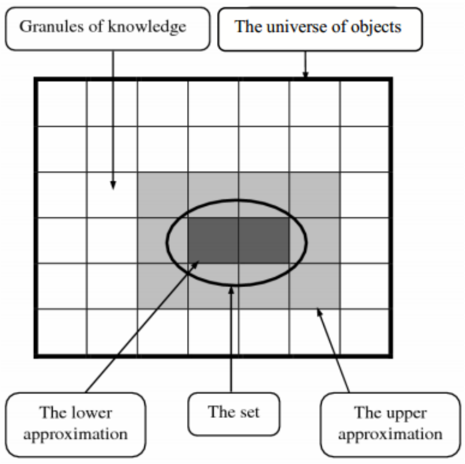

Figure 1: A rough set

3. $\mathsf{u}(S) = \bigcup\{B \mid B \in \mathfrak{B} \text{ and } B \cap S \neq \emptyset\}$.

U is the set of objects. $\mathfrak{B}$ is the system of base sets which represents the background knowledge. $\mathfrak{D}_{\mathfrak{B}}$ is the set of definable sets which defines how the base sets can be used in the set approximation. The functions l and u give the lower and upper approximation of a set. The lower approximation contains objects that surely belong to the set, and the upper approximation contains objects that possibly belong to the set.

In Fig 1 a set and its lower and upper approximation can be seen. Each rectangle denotes a base set.

Correlation clustering

Data mining is the process of discovering patterns and hidden information in large data sets. The goal of a data mining process is to extract information from a data set and transform it into an understandable structure for further use. Clustering is a data mining technique in which the goal is to group the objects so that the objects in the same group are more similar to each other than to those in other groups. In many cases the similarity is based on the attribute values of the objects. In most of them, some kind of distance is used to define the similarity. However, sometimes only nominal data are given. In this particular case distance can be meaningless. For example, what is the distance between a male and a female? In this case a similarity relation can be used, which is a tolerance relation. If this relation holds for two objects, we can say that they are similar. If this relation does not hold then they are dissimilar. It is easy to prove that this relation is reflexive and symmetric. The transitivity, however, does not hold necessarily. Correlation clustering is a clustering technique based on a tolerance relation (see in [5, 6, 7]).

Let V a set of objects and T the similarity relation. The task is to find an $R \subseteq V \times V$ equivalence relation which is *closest* to the tolerance relation.

A (partial) tolerance relation T (see in [10, 9]) can be represented by a matrix M. Let matrix $M = (m_{ij})$ be the matrix of the partial relation T of similarity:

$$m_{ij} = \begin{cases} 1 & i \text{ and } j \text{ are similar} \\ -1 & i \text{ and } j \text{ are different} \\ 0 & \text{otherwise} \end{cases}$$

A relation is called partial if there exist two elements (i, j) such that $m_{ij} = 0$. It means that if we have an arbitrary relation $R \subseteq V \times V$ we have two sets of pairs. Let R_{true} be the set of those pairs of elements for which the R holds and R_{false} be the one for which R does not hold. If R is partial, then $R_{true} \cup R_{false} \subseteq V \times V$. If R is total then $R_{true} \cup R_{false} = V \times V$.

A partition of a set S is a function $p : S \to \mathbb{N}$. Objects $x, y \in S$ are in the same cluster at partitioning p, if $p(x) = p(y)$. We treat the following two cases conflicts for any $x, y \in V$:

- $(x, y) \in T$ but $p(x) \neq p(y)$
- $(x, y) \notin T$ but $p(x) = p(y)$

The goal is to minimize the number of these conflicts. If their number is 0, the partition is called *perfect*. Given the T and R we call this value the distance of the two relations. The partition given this way, generates an equivalence relation. This relation can be considered as the closest to the tolerance relation.

The number of partitions can be given by the Bell number (see in [8]) which grows exponentially. For more than 15 objects, we cannot achieve the optimal partition by exhaustive search. In a practical case, a search algorithm can be used which can give a quasi optimal partition.

Similarity based rough sets

In practical applications, indiscernibility relation is too strong. Therefore, Pawlakian approximation spaces have been generalized using tolerance relations (symmetric and reflexive), which are similarity relations. Covering-based approximation spaces (see [11]) generalize Pawlakian approximation spaces in two points:

1. $\mathcal{R}$ is a tolerance relation;
2. $\mathfrak{B} = \{[x] \mid [x] \subseteq U, x \in U$ and $y \in [x]$ if $x\mathcal{R}y\}$, where $[x] = \big\{y \mid y \in U, x\mathcal{R}y\big\}$.

The definitions of definable sets and approximation pairs are the same as before. In these covering systems each object generates a base set.

Correlation clustering defines a partition. The clusters contain objects that are typically similar to each other. In our previous work (see in [4]) we showed that this partition can be understood as the system of base sets. The approximation space given this way has several good properties. The most important one is that it focuses on the similarity (the tolerance relation) itself and it is different from the covering type approximation space relying on the tolerance relation.

In reasonable time, correlation clustering can only be solved using search algorithms. However, each algorithm can provide different clusters. So the system of base sets can also be different. It is a natural question to ask, how the search algorithms can affect the structure of the base sets.

Acknowledgements

This work was supported by the construction EFOP-3.6.3-VEKOP-16-2017-00002. The project was co-financed by the Hungarian Government and the European Social Fund.

References

[1] Pawlak, Z.: Rough sets. International Journal of Parallel Programming 11(5), 341–356 (1982)

[2] Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information sciences 177(1), 3–27 (2007)

[3] Pawlak, Z., et al.: Rough sets: Theoretical aspects of reasoning about data. Systern Theory, Knowledge Engineering and Problem Solving, Kluwer Academic Publishers, Dordrecht, 1991 9 (1991)

[4] Nagy, D., Mihálydeák, T., Aszalós, L.: Similarity Based Rough Sets, pp. 94–107. Springer International Publishing, Cham (2017), https://doi.org/10.1007/978–3–319–60840-2_7

[5] Bansal, N., Blum, A., Chawla, S.: Correlation clustering. Machine Learning 56(1-3), 89–113 (2004)

[6] Becker, H.: A survey of correlation clustering. Advanced Topics in Computational Learning Theory pp. 1–10 (2005)

[7] Zimek, A.: Correlation clustering. ACM SIGKDD Explorations Newsletter 11(1), 53–54 (2009)

[8] Aigner, M.: Enumeration via ballot numbers. Discrete Mathematics 308(12), 2544 – 2563 (2008), http://www.sciencedirect.com/science/article/pii/S0012365X07004542

[9] Mani, A.: Choice inclusive general rough semantics. Information Sciences 181(6), 1097–1115 (2011)

[10] Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundamenta Informaticae 27(2), 245–253 (1996)

[11] Yao, Y., Yao, B.: Covering based rough set approximations. Information Sciences 200, 91–107 (2012), http://www.sciencedirect.com/science/article/pii/S0020025512001934