Similarity Based Granules

Dávid Nagy(✉), Tamás Mihálydeák, and Tamás Kádek

Department of Computer Science, Faculty of Informatics, University of Debrecen,
Egyetem tér 1, Debrecen 4010, Hungary
{nagy.david,kadek.tamas}@inf.unideb.hu,
mihalydeak.tamas@unideb.hu

Abstract. In the authors' previous research, a possible usage of the correlation clustering in rough set theory was investigated. Correlation clustering is based on a tolerance relation and its output is a partition. The system of granules can be derived from the partition and as a result, a new approximation space appears. This space focuses on the similarity (represented by a tolerance relation) itself and it is different from the covering type approximation space relying on a tolerance relation. In real-world applications, the number of objects is very high. So it can be effective only if a portion of the data points is used. Previously we provided a method that chooses the necessary number of objects that represent the data set. These members are called representatives and it can be useful to apply them in the approximation of an arbitrary set. A new approximation pair can be defined based on the representatives. In this paper, some very important properties are checked for this approximation pair and the system of granules.

Keywords: Rough set theory · Correlation clustering · Set approximation · Representatives · Granules

1 Introduction

Nowadays a huge amount of data is stored in databases. The stored data is usually represented by objects with (maybe different) properties. Properties are handled in two steps: attributes and the corresponding attribute values. Generally, a finite number of attributes and a finite number of the corresponding attribute values are used. Usually, there are more objects than attribute values. Therefore, more than one object may have the same attribute values (not considering the IDs), so they are indiscernible based on the available knowledge. Naturally, indiscernible objects have to be treated in the same way. Pawlak's original system of rough sets shows the consequences of indiscernibility. In many practical cases, not only indiscernible objects have to be treated in the same way, but objects with the same attribute values of some (and not all) attributes. This is one of the theoretical bases of the generalizations of Pawlak's original theory. Some objects have to be treated in the same way. In rough set theory the objects, that are treated in the same way, belong to a base set. In our previous study, we

R. Bello et al. (Eds.): IJCRS 2020, LNAI 12179, pp. 35–47, 2020.
https://doi.org/10.1007/978-3-030-52705-1_3

examined whether the partition, generated by correlation clustering, can be considered as the system of base sets in an application. Correlation clustering is a clustering method in data mining which creates a partition of the input data set based on a tolerance relation (representing similarity). The clusters gained this way contain similar objects. In our previous paper [11,12] we showed that it is worth to generate the system of base sets from the partition. This way, the base sets contain objects that are typically similar to each other and the generated approximation space (similarity based rough sets) possesses several very useful properties. Informally, in granular computing a granule contains objects which have to be treated in the same way. Granules play—as the most fundamental concept—a crucial role in granular computing. It means that granules (and not objects belonging to them) are in the focus of investigations. The clusters generated by the correlation clustering can be considered as granules. In order to use granules, one has to give their names. In order to preserve the connection between a granule and its objects, the name of the granule can be an object belonging to the granule. This object can represent the given granule. In a very general case to choose representatives is not a trivial problem. In the case of a system relying on an indiscernible relation any object of a granule can be its name, can represent the corresponding granule. When similarity (represented by a tolerance relation) is used to get granules, then the method of correlation clustering gives a possibility to define representatives [5,10]. In [10] a new approximation pair is proposed that is completely based on the representatives. Professor Mihir Chakraborty proposed some very important properties of granules (presented at the International Workshop on Modern and Unconventional Approaches to Reasoning and Computing in 2017). In this paper, we examined these properties along with some other ones for our introduced granules. We also show that the clusters gained from the correlation clustering satisfy all the minimal properties of the granules. Therefore, the clusters can be really treated as granules. The structure of the paper is the following: we begin by introducing the theoretical background of rough set theory. In Sect. 3 correlation clustering is defined. In Sect. 4 we present our previously introduced approximation space. In Sect. 5 we show the definition of the approximation pairs that are based on the representatives. After this, we show which of the defined properties hold for the proposed approximation pair. Finally, we conclude the results.

2 Theoretical Background

From the granular point of view a Pawlakian approximation space [13–15] is an ordered 5-tuple $\langle U, \mathfrak{G}, \mathfrak{D}, \mathsf{l}, \mathsf{u} \rangle$ generated by an equivalence relation $\mathcal{R}$ (which represents indiscernibility), where:

- $U \neq \emptyset$ is the universe of objects
- $\mathfrak{G}$ is the set of granules for which the following properties hold:
  - $\mathfrak{G} \neq \emptyset$
  - if $G \in \mathfrak{G}$ then $G \subseteq U$ and $G \neq \emptyset$

- $\mathfrak{G} = \{G \mid G \subseteq U,$ and $x, y \in G$ if $x \mathcal{R} y\}$

– $\mathfrak{D}$ is the set of definable sets which can be given by the following inductive definition:
  1. $\mathfrak{G} \subseteq \mathfrak{D}$;
  2. $\emptyset \in \mathfrak{D}$;
  3. if $D_1, D_2 \in \mathfrak{D}$, then $D_1 \cup D_2 \in \mathfrak{D}$.
– The functions l, u form a Pawlakian approximation pair $\langle \mathsf{l}, \mathsf{u} \rangle$ if the followings are true for an arbitrary set $S \subseteq U$:
  1. $Dom(\mathsf{l}) = Dom(\mathsf{u}) = 2^U$
  2. $\mathsf{l}(S) = \bigcup \{G \mid G \in \mathfrak{G}$ and $G \subseteq S\}$;
  3. $\mathsf{u}(S) = \bigcup \{G \mid G \in \mathfrak{G}$ and $G \cap S \neq \emptyset\}$.

3 Correlation Clustering

Cluster analysis is an unsupervised learning method in data mining. The goal is to group the objects so that the objects in the same group are more similar to each other than to those which are in other groups. In many cases, the similarity is based on the attribute values of the objects. Although there are some cases when the properties of objects can be difficult to be quantified, but something about their similarity or dissimilarity can still be said. For example, let's consider the humans. We cannot describe someone's looks using only a number, but we can make simple statements on whether two people are similar or dissimilar. These opinions are dependent on the person making the statements. Someone can say that two people are similar while others treat them as dissimilar. If we want to formulate the similarity and dissimilarity using mathematics, we need a tolerance relation (i.e. a reflexive and symmetric 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. This relation is reflexive because every object is similar to itself. It is also symmetric because if some object is similar to another one, then the similarity is equivalent the other way round. However transitivity does not necessarily hold. If we take a human and a mouse, then due to their inner structure they are considered similar. This is the reason mice are used in many drug experiments. A human and a mannequin are also similar, this time according to their shape. This is why these dolls are used in display windows. However, a mouse and a mannequin are dissimilar (except that both are similar to the same object). Correlation clustering is a clustering technique based on a tolerance relation [6,7,17].

The task is to find an $R \subseteq U \times U$ equivalence relation which is *closest* to the tolerance relation. A (partial) tolerance relation $\mathcal{R}$ [8,16] can be represented by a matrix M. Let matrix $M = (m_{ij})$ be the matrix of the partial relation $\mathcal{R}$ of similarity: $m_{ij} = 1$ if objects i and j are similar, $m_{ij} = -1$ if objects i and j are dissimilar, and $m_{ij} = 0$ otherwise.

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 $R \subseteq U \times U$ we have two sets of pairs. Let R_{true} be the set of those pairs of elements for which R holds and

R_{false} be the one for which R does not hold. If R is partial, then $R_{true} \cup R_{false}$ is a proper subset of $U \times U$. If R is total, then $R_{true} \cup R_{false} = U \times U$.

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 $p(x) = p(y)$. For a conflict one of the following two cases holds:

- Two dissimilar objects end up in the same cluster
- Two similar objects end up in different clusters

The cost function is the number of these disagreements. The formal definition can be seen in [11]. For a relation, the partition with the minimal cost function value is called *optimal*. Solving a correlation clustering problem is equivalent to minimising its cost function for the fixed relation. If the cost function's value is 0, the partition is called *perfect*. Given the $\mathcal{R}$ and R we call the value f the distance of the two relations. With this definition, the partition generates an equivalence relation. This relation can be considered to be the closest to the tolerance relation.

It is easy to check that we cannot necessarily find a perfect partition for an arbitrary similarity relation. Consider the simplest such case, given three objects A, B and C, and A is similar to both B and C, but B and C are dissimilar. In this situation, the following 5 partitions can be given:

$$\{\{A,B,C\},\{\{A,B\},\{C\}\},\{\{A,C\},\{B\}\},\{\{B,C\},\{A\}\},\{\{A\},\{B\},\{C\}\}\}.$$

It is easy to see that in every of one them there is at least 1 conflict. The number of partitions can be given by the Bell number [1], which grows exponentially. So the optimal partition cannot be determined in reasonable time. In a practical case a quasi optimal partition can be sufficient, so a search algorithm can be used.

The main advantage of the correlation clustering is that the number of clusters does not need to be specified in advance like in many clustering algorithms, and this number is optimal based on the similarity. However, as the number of partitions grows exponentially it is an NP-hard problem.

4 Similarity Based Granules

The system of granules is based on the background knowledge embedded in an information system. The granules represent the background knowledge (or its limit). In the Pawlakian systems, two objects are treated as indiscernible if all of their known attribute values are the same. The indiscernibility property can be represented by an equivalence relation. In practical applications not only the indiscernible objects must be handled in the same way but also those that are similar to each other based on some property. (Irrelevant differences for the purpose of the given applications should not be taken into account.) Some covering approximation spaces use tolerance relations, which represent similarity, instead of equivalence relations, but the usage of these relations is very special.

It emphasizes the similarity to a given object and not the similarity of objects 'in general'. This means that a granule contains objects which are similar to a distinguished object. In these systems, each object generates a granule. With correlation clustering, a quasi-optimal partition of the universe can be obtained [2–4]. The members of a partition are called clusters. They contain objects that are typically similar to each other and not just to a distinguished member. In our previous research, we investigated if the partition can be understood as a system of granules [9,11,12]. According to our results, it is worth to generate a partition with correlation clustering. Singleton clusters represent very little information (its member is only similar to itself). Without increasing the number of conflicts its member cannot be considered similar to any objects. So, they always require an individual decision. By deleting the singletons, a partial system of granules can be defined. The formal definition of the proposed approximation space (similarity based rough sets) can be seen in the following definition.

Definition 1. *Similarity based rough set approximation space can be represented by an ordered 6-tuple* $\langle U, \mathfrak{G}, \mathfrak{D}, \mathsf{l}, \mathsf{u}, \mathfrak{S} \rangle$ *based on a tolerance relation (representing similarity)* $\mathcal{R}$*. Let* p *be the partition gained from the correlation clustering (based on* $\mathcal{R}$*).*

- *the definition of* $U, \mathfrak{D}, \mathsf{l}$ *and* u *are the same as in the Pawlakian space.*
- $\mathfrak{S}$ *denotes the set of the singleton members.*
- $\mathfrak{G} = \{G \mid G \subseteq U \setminus \mathfrak{S},$ *and* $x, y \in G$ *if* $p(x) = p(y)\}$

The introduced approximation space has some useful features:

- the similarity of objects relying on their properties (and not the similarity to a distinguished object) plays an important role in the definition of granules;
- the system of granules consists of disjoint sets, so the lower and upper approximations are closed in the following sense: Let S be a set and $x \in U$. If $x \in \mathsf{l}(S)$, then we can say, that every object $y \in U$ which is in the same cluster as x is in $\mathsf{l}(S)$. If $x \in \mathsf{u}(S)$, then we can say, that every object $y \in U$ which is in the same cluster as x is in $\mathsf{u}(S)$.
- the number of clusters is not predefined because the algorithm finds the optimal number. This way, only the necessary number of granules appear (in applications we have to use an acceptable number of granules);
- the size of the granules is not too small, nor too big.

The amount of daily produced data is unbelievable. There are around 2.5 quintillion bytes of data created each day at our current pace and it is only accelerating with the growth of the Internet of Things (IoT). In data sciences, it is extremely important that certain methods can be used for a large amount of data. Due to the exploding volume and speed of data growth, the resource need and execution time of the algorithms show an increasing trend. In data mining to mitigate this problem, it is common to use samples. There are numerous ways to choose a part of the input dataset which can be treated as a sample. In every method, it is crucial that the chosen objects must represent the entire population. In this case, representativeness means that the specific properties are

as similar in the sample as in the entire set. Without this property, important information might be disregarded. Imagine that a product is needed to be sold, for example, a toy to a group of children. In almost every group of youngsters, there is at least one child whose decision has the most influence on the group's life. In this case, one child is enough to be found and convinced to buy the toy. The rest of the group will follow them. This child can be treated as the representative of the group. It means that in the computations only this child should be considered instead of the whole group. In a pawlakian system, any object can be the representative of a certain granule. In the covering systems (based on a tolerance relation) the representatives are obvious in each granule. In the similarity based rough set approximation space, the situation is not that simple. In each granule, we need to choose an object that is the most similar in the set. Naturally, it can happen that the entire granule cannot be represented by only one member. In [5] we proposed an algorithm that produces the necessary number of representatives for each granule. The algorithm assigns a rank value to each object. This value shows how much the given object represents the granule.

Definition 2. *The object with the highest rank value is called primary representative. If there is more than one object with the same rank, then the primary representative is chosen randomly.*

Generally speaking, we can say that a granule represents a property. A represented property can be characterized by attributes and the corresponding attribute values. For example, the property 'being red apple' can be characterized by color and fruit type as attributes and by red and apple as corresponding attribute values. If P is a property, then P can be an intension of a granule G. The granule itself is a set of objects that possess the property described by its intension. In our system, a granule contains objects that are typically similar to each other. Every granule has a primary representative which represents the entire granule the most. In an information system, every object has attributes and attributes values. The list of these attribute values describes a certain property.

Definition 3. *The intension of a granule is the property described by its primary representative.*

5 Approximation Based on Representatives

In the classical sense, the lower approximation of a set S is the union of those granules that are subsets of S. In order to get these granules, every object in each granule must be considered. It can be a time-consuming task if the number of points is high. The effectiveness of the representatives lies in situations when the number of objects is very large. It can be practical to use the strength of representatives in the approximation process. For each granule, let us consider only its representatives. Let $G \in \mathfrak{G}$ be a granule, and $REP(G)$ be the set of its representatives such that $REP(G) \subseteq G$ and $REP(G) \neq \emptyset$ for all $G \in \mathfrak{G}$ (and so $\emptyset \notin \mathfrak{G}$). The approximation pair are defined as the following:

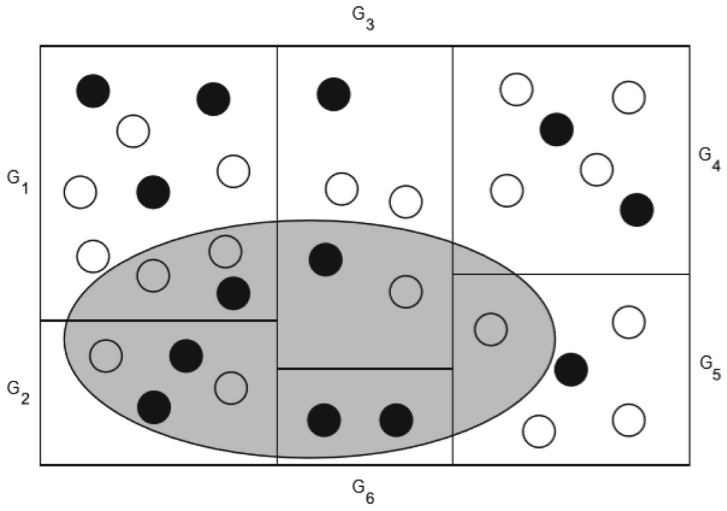

Fig. 1. Approximation based on representatives

- $\mathsf{l}_\mathsf{r}(S) = \bigcup\{G \mid G \in \mathfrak{G} \text{ and } REP(G) \subseteq S\}$ (and so $\mathsf{l}_\mathsf{r}(S) \in \mathfrak{D}$);
- $\mathsf{u}_\mathsf{r}(S) = \bigcup\{G \mid G \in \mathfrak{G} \text{ and } REP(G) \cap S \neq \emptyset\}$ (and so $\mathsf{u}_\mathsf{r}(S) \in \mathfrak{D}$).

This way, the lower approximation of a set S becomes the union of those granules for which every representative is a member of S. A granule belongs to the upper approximation if at least one of its representatives is in the set S. Naturally, the certainty of the lower approximation might be lost, but as the number of points is increasing, it can be very useful.

In Fig. 1 a simple example is provided for the method. The granules are denoted by solid-line rectangles, and the set we wish to approximate (S) is denoted by a grey ellipse. For each granule, the black circles symbolise the representatives.

The approximation of the set S is the following based on the representatives:

- $\mathsf{l}_\mathsf{r}(S) = G_2 \cup G_6$
- $\mathsf{u}_\mathsf{r}(S) = G_1 \cup G_2 \cup G_3 \cup G_6$

The approximation of the set S is the following based on the classical approximation pair:

- $\mathsf{l}(S) = G_2 \cup G_6$
- $\mathsf{u}(S) = G_1 \cup G_2 \cup G_3 \cup G_5 \cup G_6$

The lower approximation is the same in both cases. The upper approximation differs in one granule (G_5). When there is a huge number of points and there are several sets to be approximated, we recommend approximation using representatives. In this case, the method can reduce the run-time of the approximation significantly. Determining the approximation with the classical functions 32 objects

needed to be considered. Using the proposed method, only 13 of them had to be tested, so almost 60% of the original points were discarded. Of course, with 32 to 13 points is not a significant change, but in the case of millions of objects, it can be very useful. Working with only the representatives, we can always save time and resources because we can be sure that the number of representatives is less than that of U. Proving this is very straightforward. Naturally, there cannot be more representatives than objects in the universe. Their numbers cannot be equal either because it could only happen if every object were a representative which implies that every cluster were singleton. Using these system is pointless because the system of granules is empty (every singleton cluster is discarded).

6 Properties of Granules

In this section, we examine the following properties (we call them as axioms) of granules (by Prof. Mihir Chakraborty):

I $\forall G \in \mathfrak{G} : G \neq \emptyset$
II $\forall G \in \mathfrak{G} : \exists a \in U$ such that G may be associated with a. Notation: G_a
III if $b \in G_a$ then $a \in G_b$
IV $\forall G \in \mathfrak{G} : \mathsf{l}_\mathsf{r}(G) = G$
V $\forall G \in \mathfrak{G} : \mathsf{u}_\mathsf{r}(G) = G$
VI $\forall G \in \mathfrak{G} : \mathsf{l}_\mathsf{r}(\mathsf{l}_\mathsf{r}(G)) = \mathsf{l}_\mathsf{r}(G)$
VII $\forall G \in \mathfrak{G} : \mathsf{u}_\mathsf{r}(\mathsf{u}_\mathsf{r}(G)) = \mathsf{u}_\mathsf{r}(G)$
VIII $\forall G \in \mathfrak{G} : \mathsf{u}_\mathsf{r}(\mathsf{l}_\mathsf{r}(G)) = \mathsf{l}_\mathsf{r}(G)$
IX $\forall G \in \mathfrak{G} : \mathsf{l}_\mathsf{r}(\mathsf{u}_\mathsf{r}(G)) = \mathsf{u}_\mathsf{r}(G)$
X $\mathsf{l}_\mathsf{r}(G)$ and $\mathsf{u}_\mathsf{r}(G)$ are duals

Theorem 1. *In $\langle U, \mathfrak{G}, \mathfrak{D}, \mathsf{l}, \mathsf{u}\rangle$ (classical Pawlakian approximation space), all of the aforementioned axioms hold.*

Theorem 2. *All the existing granules admit Axiom I, II and IV.*

Theorem 3. *In $\langle U, \mathfrak{G}, \mathfrak{D}, \mathsf{l}_\mathsf{r}, \mathsf{u}_\mathsf{r}, \mathfrak{S}\rangle$ (similarity based rough sets approximation space based on the representatives) all of the aforementioned axioms hold except for the duality property.*

Proof (Axiom I). This axiom trivially holds because in the similarity based rough sets approximation space every granule contains at least 2 objects.

Proof (Axiom II). The axiom holds as every granule has at least one representative. We can associate the granule with one of the representatives of the granule.

Proof (Axiom III). If representative b is in the granule of representative a, then it could only happen if $G_a = G_b$. Let us suppose that $G_a \neq G_b$. From Axiom II we know that $b \in G_b$. So if representative b is in G_a, then $G_a \cap G_b = \{b\}$ which means that G_a and G_b are not disjoint. This is a contradiction, therefore G_a and G_b must be the same set.

Proof (Axiom IV). $\mathsf{l}_\mathsf{r}(G) = \bigcup\{G' \mid G' \in \mathfrak{G} \text{ and } \forall x \in REP(G') : x \in G\}$. The granules are pairwise disjoint, so there is no granule whose representatives is a member of G (other than G itself). Naturally, every representative of G is the member of G. Therefore, the set $\{G' \mid G' \in \mathfrak{G} \text{ and } \forall x \in REP(G') : x \in G\}$ contains only G from which $\mathsf{l}_\mathsf{r}(G) = G$ follows.

Proof (Axiom V). The proof of the fifth axiom is very similar to the proof of the fourth axiom. $\mathsf{u}_\mathsf{r}(G) = \bigcup\{G' \mid G' \in \mathfrak{G} \text{ and } \exists x \in REP(G') : x \in G\}$. The granules are pairwise disjoint, so there is no granule whose representatives is a member of G (other than G itself). If $\forall x \in REP(G) : x \in G$ is true, then $\exists x \in REP(G) : x \in G$ will be also true. Therefore, the set $\{G' \mid G' \in \mathfrak{G} \text{ and } \forall x \in REP(G') : x \in G\}$ contains only G from which $\mathsf{u}_\mathsf{r}(G) = G$ follows.

Proof (Axiom VI–IX). If Axiom 4 and 5 hold, then Axiom VI–IX follow.

Proof (Axiom X). The duality property holds if the following two equalities hold for any granule G ($\complement$ denotes the complement operator):

1. $\mathsf{l}_\mathsf{r}(G) = \mathsf{u}_\mathsf{r}(G^\complement)^\complement$
2. $\mathsf{u}_\mathsf{r}(G) = \mathsf{l}_\mathsf{r}(G^\complement)^\complement$

Let $U = \{a, b, c, d, e\}$, $\mathfrak{G} = \{G_1, G_2\}$, $G_1 = \{a, b\}$, $G_2 = \{c, d\}$, $REP(G_1) = \{a\}$, $REP(G_2) = \{c\}$. In this example, $\mathsf{l}_\mathsf{r}(G_1) = \{a, b\}$ and $G_1^\complement = \{c, d, e\}$. From this $\mathsf{u}_\mathsf{r}(G_1^\complement) = \{c, d\}$ follows. However, $\mathsf{u}_\mathsf{r}(G_1^\complement)^\complement = \{a, b, e\} \neq \{a, b\}$. Therefore the duality property does not hold.

6.1 Properties of Approximation Pairs

In the previous section, the axioms only focused on the granules. In this section, we examine some additional properties of the proposed approximation pair. Here, the properties to be checked are based on definable and arbitrary sets not only granules. The most essential features of approximation pairs are specified as follows.

Monotonicity
l and u are said to be monotone if $S \subseteq S'$ then $\mathsf{l}(S) \subseteq \mathsf{l}(S')$ and $\mathsf{u}(S) \subseteq \mathsf{u}(S')$
Weak approximation property
$\forall S \in 2^U : \mathsf{l}(S) \subseteq \mathsf{u}(S)$
Strong approximation property
$\forall S \in 2^U : \mathsf{l}(S) \subseteq S \subseteq \mathsf{u}(S)$
Normality of l
$\mathsf{l}(\emptyset) = \emptyset$
Normality of u
$\mathsf{u}(\emptyset) = \emptyset$

Theorem 4. *In $\langle U, \mathfrak{G}, \mathfrak{D}, \mathsf{l}_\mathsf{r}, \mathsf{u}_\mathsf{r}, \mathfrak{S}\rangle$ (similarity based rough sets approximation space based on the representatives), the monotonicity, the weak approximation property and the normality of l_r and u_r hold and the strong approximation property does not hold.*

Proof (Monotonicity). Let S and S' be two arbitrary set such that $S \subset S'$ which means that there is an object x which is a member of S' but not a member of S. The following cases can be true for x:

1. $x \in \mathfrak{S}$, then $\mathsf{l}_\mathsf{r}(S) = \mathsf{l}_\mathsf{r}(S')$ and $\mathsf{u}_\mathsf{r}(S) = \mathsf{u}_\mathsf{r}(S')$
2. x is a non-representative, then $\mathsf{l}_\mathsf{r}(S) = \mathsf{l}_\mathsf{r}(S')$ and $\mathsf{u}_\mathsf{r}(S) = \mathsf{u}_\mathsf{r}(S')$
3. x is a representative of a granule G, then the following cases can happen:
   (a) if $\neg\exists y(y \in REP(G) \wedge x \neq y \wedge y \in S)$, then $\mathsf{l}_\mathsf{r}(S) = \mathsf{l}_\mathsf{r}(S')$ and $\mathsf{u}_\mathsf{r}(S) \subset \mathsf{u}_\mathsf{r}(S')$
   (b) if $\exists y(y \in REP(G) \wedge x \neq y \wedge y \in S)$, then $\mathsf{l}_\mathsf{r}(S) = \mathsf{l}_\mathsf{r}(S')$ and $\mathsf{u}_\mathsf{r}(S) = \mathsf{u}_\mathsf{r}(S')$
   (c) if $\forall y(y \in REP(G) \wedge x \neq y \rightarrow y \notin S)$, then $\mathsf{l}_\mathsf{r}(S) \subset \mathsf{l}_\mathsf{r}(S')$ and $\mathsf{u}_\mathsf{r}(S) \subset \mathsf{u}_\mathsf{r}(S')$

In every case, we found that $\mathsf{l}_\mathsf{r}(S) \subseteq \mathsf{l}_\mathsf{r}(S')$ and $\mathsf{u}_\mathsf{r}(S) \subseteq \mathsf{u}_\mathsf{r}(S')$, therefore the monotonicity holds.

Proof (Weak approximation property). Let S be an arbitrary set and let us assume that there is a granule G such that $G \subseteq \mathsf{l}_\mathsf{r}(S)$ but $G \nsubseteq \mathsf{u}_\mathsf{r}(S)$. Due to the definition of the lower approximation, we know that $\forall x \in REP(G) : x \in S$ is true, so $\exists x \in REP(G) : x \in S$ is also true. This implies that $G \subseteq \mathsf{u}_\mathsf{r}(S)$. We reached a contradiction, therefore the weak approximation property holds.

Proof (Strong approximation property). Let $U = \{a, b, c\}$ be the universe, $G = \{a, b, c\}$ a granule, $\mathfrak{G} = \{G\}$ be the system of granules, $S = \{a, b\}$ be the set to be approximated and $REP(G) = \{b\}$ be the representatives of G. In this case $\mathsf{l}_\mathsf{r}(S) = G = \{a, b, c\}$ which means that $\mathsf{l}_\mathsf{r}(S) \nsubseteq S$. So the strong approximation property does not hold.

Proof (Normality of l_r and u_r). The empty set does not have a representative. Therefore the condition in the definition of the lower and upper approximation is false for every granule. This implies that $\mathsf{l}_\mathsf{r}(\emptyset) = \mathsf{u}_\mathsf{r}(\emptyset) = \emptyset$.

Theorem 5. *Let $G \in \mathfrak{G}$ and $D \in \mathfrak{D}$. If $a \in G$ and $a \in D$ then $G \subseteq D$.*

Proof. If $a \in D$ then there exists a $G' \in \mathfrak{G}$ such that $a \in G'$ and $G' \subseteq D$. The members of $\mathfrak{G}$ are pairwise disjoint, so it is true for all $G_1, G_2 \in \mathfrak{G}$ that $G_1 \cap G_2 \neq \emptyset$ only if $G_1 = G_2$. Therefore $G = G'$ hence $a \in G$ and $a \in G'$. Earlier we have found that $G' \subseteq D$ and so $G \subseteq D$.

Theorem 6. $\mathsf{l}_\mathsf{r}(D) \subseteq D$ *for all* $D \in \mathfrak{D}$.

Proof. We indirectly suppose, that there exists a $D \in \mathfrak{D}$ so that $\mathsf{l}_\mathsf{r}(D) \nsubseteq D$. Therefore there exists an $a \in \mathsf{l}_\mathsf{r}(D)$ so that $a \notin D$. If $a \in \mathsf{l}_\mathsf{r}(D)$ then there exists a $G \in \mathfrak{G}$ where $REP(G) \subseteq D$ such that $a \in G$. $REP(G) \neq \emptyset$ so there exists a $b \in REP(G)$ and so $b \in D$. Because $REP(G) \subseteq G$ it is also true that $b \in G$. Based on Theorem 5, if $b \in G$ and $b \in D$ then $G \subseteq D$. Because of $a \in G$ the $a \in D$ contradiction appears.

Theorem 7. $\mathsf{u}_\mathsf{r}(D) \subseteq D$ *for all* $D \in \mathfrak{D}$.

Proof. We indirectly suppose, that there exists a $D \in \mathfrak{D}$ so that $\mathsf{u}_\mathsf{r}(D) \not\subseteq D$. Therefore there exists an $a \in \mathsf{u}_\mathsf{r}(D)$ so that $a \notin D$. If $a \in \mathsf{u}_\mathsf{r}(D)$ then there exists a $G \in \mathfrak{G}$ where $REP(G) \cap D \neq \emptyset$ such that $a \in G$. So there exists a $b \in REP(G) \cap D$ so obviously $b \in REP(G)$ and $b \in D$. Because $REP(G) \subseteq G$ it is also true that $b \in G$. Based on Theorem 5, if $b \in G$ and $b \in D$ then $G \subseteq D$. Because of $a \in G$ the $a \in D$ contradiction appears.

Definition 4 (Weak approximation pair). *An approximation pair* $\langle \mathsf{l}, \mathsf{u} \rangle$ *is a weak approximation pair on* U *if:*

- l *and* u *are monotone (monotonicity)*
- $\mathsf{u}(\emptyset) = \emptyset$ *(normality of* u*)*
- *if* $D \in \mathfrak{D}$*, then* $\mathsf{l}(D) = D$ *(granularity of* $\mathfrak{D}$*)*
- *if* $\forall S \in 2^U : \mathsf{l}(S) \subseteq \mathsf{u}(S)$ *(weak approximation property)*

Theorem 8. $\langle \mathsf{l}_\mathsf{r}, \mathsf{u}_\mathsf{r} \rangle$ *is a weak approximation pair.*

Proof. Previously we proved that l_r and u_r are monotone and the normality of u_r and the weak approximation property hold. We need to prove that the granularity of $\mathfrak{D}$ also holds. From Theorem 6 we know that $\mathsf{l}_\mathsf{r}(D) \subseteq D$ for any definable set. We just need to prove that $D \subseteq \mathsf{l}_\mathsf{r}(D)$ for any definable set. Let's indirectly suppose that $D \not\subseteq \mathsf{l}_\mathsf{r}(D)$. It means that there is a granule G' such that $G' \subseteq D$ but $G' \not\subseteq \mathsf{l}_\mathsf{r}(D)$. Therefore, there must be a representative member r of G' such that $r \notin D$. By definition $r \in G'$. If $G' \subseteq D$, then every member of G' is a member of D. However $r \in G'$ but $r \notin D$, therefore G' cannot be a subset of D. This contradicts our original assumption. So $D \subseteq \mathsf{l}_\mathsf{r}(D)$.

7 Conclusion

In [11,12] the authors introduced a partial approximation space relying on the tolerance relation (representing similarity). The genuine novelty of this new approximation space is the way in which the system of base sets is defined: it is the result of correlation clustering, and so the similarity is taken into consideration generally. In granular computing, a granule is a collection of objects that are treated in the same way. In correlation clustering, a cluster contains entities that are typically similar to each other. In this case, the objects that are in the same cluster are treated in the same way. Therefore, we can treat the clusters and so the base sets as granules. In data sciences, it is very common to use only a subset of the original dataset instead of the entire collection. The members of this subset can be called as representatives. A very important criterion is that these objects must have the same properties as the whole data set. In [5,10] we provided a possible way to choose the necessary number of representatives of a set. We also introduced a new approximation pair which is based on the representatives. In this paper, we examined some essential properties of

granules (proposed by Prof. Mihir Chakraborty). We showed that the system of granules generated by the correlation clustering satisfies all the minimal properties of the granules. Therefore, the clusters can be really treated as granules. We also proved that the introduced approximation pair is a weak approximation pair.

Acknowledgement. This work was supported by the construction EFOP-3.6.3-VEKOP-16-2017-00002. The project was supported by the European Union, co-financed by the European Social Fund.

References

1. Aigner, M.: Enumeration via ballot numbers. Discret. Math. **308**(12), 2544–2563 (2008). https://doi.org/10.1016/j.disc.2007.06.012. http://www.sciencedirect.com/science/article/pii/S0012365X07004542
2. Aszalós, L., Mihálydeák, T.: Rough clustering generated by correlation clustering. In: Ciucci, D., Inuiguchi, M., Yao, Y., Ślęzak, D., Wang, G. (eds.) RSFDGrC 2013. LNCS (LNAI), vol. 8170, pp. 315–324. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41218-9_34
3. Aszalós, L., Mihálydeák, T.: Rough classification based on correlation clustering. In: Miao, D., Pedrycz, W., Ślęzak, D., Peters, G., Hu, Q., Wang, R. (eds.) RSKT 2014. LNCS (LNAI), vol. 8818, pp. 399–410. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11740-9_37
4. Aszalós, L., Mihálydeák, T.: Correlation clustering by contraction. In: 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 425–434. IEEE (2015)
5. Aszalós, L., Nagy, D.: Iterative set approximations based on tolerance relation. In: Mihálydeák, T., et al. (eds.) IJCRS 2019. LNCS (LNAI), vol. 11499, pp. 78–90. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22815-6_7
6. Bansal, N., Blum, A., Chawla, S.: Correlation clustering. Mach. Learn. **56**(1–3), 89–113 (2004)
7. Becker, H.: A survey of correlation clustering. In: Advanced Topics in Computational Learning Theory, pp. 1–10 (2005)
8. Mani, A.: Choice inclusive general rough semantics. Inf. Sci. **181**(6), 1097–1115 (2011)
9. Mihálydeák, T.: Logic on similarity based rough sets. In: Nguyen, H.S., Ha, Q.-T., Li, T., Przybyła-Kasperek, M. (eds.) IJCRS 2018. LNCS (LNAI), vol. 11103, pp. 270–283. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99368-3_21
10. Nagy, D., Aszalós, L.: Approximation based on representatives. In: Mihálydeák, T., et al. (eds.) IJCRS 2019. LNCS (LNAI), vol. 11499, pp. 91–101. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22815-6_8
11. Nagy, D., Mihálydeák, T., Aszalós, L.: Similarity based rough sets. In: Polkowski, L., Yao, Y., Artiemjew, P., Ciucci, D., Liu, D., Ślęzak, D., Zielosko, B. (eds.) IJCRS 2017. LNCS (LNAI), vol. 10314, pp. 94–107. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60840-2_7
12. Nagy, D., Mihálydeák, T., Aszalós, L.: Similarity based rough sets with annotation. In: Nguyen, H.S., Ha, Q.-T., Li, T., Przybyła-Kasperek, M. (eds.) IJCRS 2018. LNCS (LNAI), vol. 11103, pp. 88–100. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99368-3_7

13. Pawlak, Z.: Rough sets. Int. J. Parallel Prog. **11**(5), 341–356 (1982)
14. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Inf. Sci. **177**(1), 3–27 (2007)
15. Pawlak, Z., et al.: Rough sets: theoretical aspects of reasoning about data. In: System Theory, Knowledge Engineering and Problem Solving, vol. 9. Kluwer Academic Publishers, Dordrecht (1991)
16. Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundamenta Informaticae **27**(2), 245–253 (1996)
17. Zimek, A.: Correlation clustering. ACM SIGKDD Explor. Newslett. **11**(1), 53–54 (2009)