Similarity-Based Rough Sets with Annotation Using Deep Learning

Dávid Nagy, Tamás Mihálydeák, Tamás Kádek

Similarity-based Rough Sets with Annotation Using Deep Learning

Dávid Nagy, Tamás Mihálydeák, and Tamás Kádek

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

Abstract In the authors' previous research the possible usage of correlation clustering in rough set theory was investigated. Correlation clustering is based on a tolerance relation that represents the similarity among objects. Its result is a partition which can be treated as the system of base sets. However, singleton clusters represent very little information about the similarity. If the singleton clusters are discarded, then the approximation space received from the partition is partial. In this way, the approximation space focuses on the similarity (represented by a tolerance relation) itself and it is different from the covering type approximation space relying on the tolerance relation. In this paper, the authors examine how the partiality can be decreased by inserting the members of some singletons into base sets and how this annotation affects the approximations. This process can be performed by the user of system. However, in the case of a huge number of objects, the annotation can take a tremendous amount of time. This paper shows an alternative solution to the issue using neural networks.

Keywords: Rough set theory · Correlation clustering · Set approximation

1 Introduction

In our previous work, we examined whether the clusters, generated by correlation clustering, can be understood as a system of base sets. Correlation clustering is a clustering method in data mining that is based on a tolerance relation. Its result is a partition. The groups, defined by this partition, contain similar objects. In [12], we showed that it is worth to generate the system of base sets from the partition. In this way, the base sets contain objects that are typically similar to each other and they are also pairwise disjoint. The proposed approximation space is different from the tolerance-based covering approximation spaces. There can be some clusters that have only one member. These singletons represent very little information regarding the similarity. This is the reason why they are not treated as base sets. Without them, the approximation space becomes partial. In practice, partiality can cause issues in logical systems. That is why its degree should be minimized. In [13] we showed a possible way to decrease it by allowing the user to insert a member of a singleton into a base set. We called this process annotation. Its main problem is that it needs to be performed manually which takes a lot of time if there are a huge number of data points. In this paper, we propose an improved version of the annotation which performs the process using a neural network. Thus the annotation can be done automatically. So, the main problem of manual annotation is mitigated. The structure of the paper is the following: A theoretical background about the classical rough set theory comes first. In section 4 we present our previous work and in section 3 we define correlation clustering mathematically. Then, we show why decreasing partiality is important in logical systems. In section 6 the annotation process is described. Finally, we conclude our results.

2 Theoretical Background

In general, a set is a collection of objects which is uniquely identified by its members. It means that if one would like to decide, whether an object belongs to a certain set, then a precise answer can be given (yes/no). A good example is the set of numbers that are divisible by 3 because it can be decided if an arbitrary number is divisible by 3 or not. Of course, it is required that one knows how to use the modulo operation. This fact can be considered as a background knowledge and it allows us to decide if a number belongs to the given set. Naturally, it is not necessary to know how to use the modulo operation for each number. Some second graders may not be able to divide numbers greater than 100. They would not be able to decide if 142 is divisible by 3 because they lack the required background knowledge. For them, 142 is neither divisible nor indivisible by 3. So there is uncertainty (vagueness) based on their knowledge. Rough set theory was proposed by Zdisław Pawlak in 1982 [14]. The theory offers a possible way to treat vagueness caused by some background knowledge. In data sciences, each object can be characterized by a set of attribute values. If two objects have the same known attribute values, then these objects cannot really be distinguished. The indiscernibility generated this way, gives the mathematical basis of rough set theory.

Definition 1. *The ordered 5–tuple* $\langle U, \mathfrak{B}, \mathfrak{D}_{\mathfrak{B}}, \mathsf{l}, \mathsf{u} \rangle$ *is a general approximation space if*

1. *U is a nonempty set;*
2. *$\mathfrak{B} \subseteq 2^U \setminus \emptyset$, $\mathfrak{B} \neq \emptyset$ ($\mathfrak{B}$ is the set of base sets);*
3. *$\mathfrak{D}_{\mathfrak{B}}$ is the set of definable sets and it is given by the following inductive definition:*
   (a) *$\emptyset \in \mathfrak{D}_{\mathfrak{B}}$;*
   (b) *$\mathfrak{B} \subseteq \mathfrak{D}_{\mathfrak{B}}$;*
   (c) *if $D_1, D_2 \in \mathfrak{D}_{\mathfrak{B}}$, then $D_1 \cup D_2 \in \mathfrak{D}_{\mathfrak{B}}$*
4. *$\langle \mathsf{l}, \mathsf{u} \rangle$ is a Pawlakian approximation pair i.e.*
   (a) *$Dom(\mathsf{l}) = Dom(\mathsf{u}) = 2^U$*
   (b) *$\mathsf{l}(S) = \bigcup\{B \mid B \in \mathfrak{B}$ and $B \subseteq S\}$;*
   (c) *$\mathsf{u}(S) = \bigcup\{B \mid B \in \mathfrak{B}$ and $B \cap S \neq \emptyset\}$.*

The system of base sets represents the background knowledge or its limit. 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.

Definition 2. *A general approximation space is Pawlakian [16,15] if $\mathfrak{B}$ is a partition of U.*

The indiscernibility modeled by an equivalence relation represents the limit of our knowledge embedded in an information system (or background knowledge). It has also an effect on the membership relation. In certain situations, it makes our judgment of the membership relation uncertain – thus making the set vague – as a decision about a given object affects the decision about all the other objects that are indiscernible from the given object. In practice, indiscernibility can be too strict as the attribute values of the objects must be completely the same. In these situations, the similarity among the objects can be enough to consider. Over the years, many new approximation spaces have been developed as the generalization of the original Pawlakian space [11]. The main difference between these kinds of approximation spaces (with a Pawlakian approximation pair) lies in the definition of the base sets (members of $\mathfrak{B}$).

Definition 3.
Different types of general approximation spaces $\langle U, \mathfrak{B}, \mathfrak{D}_{\mathfrak{B}}, \mathsf{l}, \mathsf{u} \rangle$ are as follows:

1. *A general approximation space is a covering approximation space [17] generated by a tolerance relation $\mathcal{R}$ if $\mathfrak{B} = \{[u]_{\mathcal{R}} \mid u \in U\}$, where $[u]_{\mathcal{R}} = \{u' \mid u \mathcal{R} u'\}$.*
2. *A general approximation space is a covering approximation space if $\bigcup \mathfrak{B} = U$.*
3. *A general approximation space is a partial approximation space if $\bigcup \mathfrak{B} \neq U$.*

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 in other groups. In many cases, the similarity is based on the attribute values of the objects. Although there are some cases when these values are not numbers, we can still say something about their similarity or dissimilarity. From the mathematical point of view, similarity can be modeled by a tolerance relation. Correlation clustering

is a clustering technique based on a tolerance relation [5,6,18]. Bansal et al. defined correlation clustering for complete weighted graphs. Here $G = (V, E)$ is a graph and function $w : E \to \{+1, -1\}$ is the weight of edges. Weight $+1$ and -1 denotes the similarity/dissimilarity of the nodes of the edges. We always treat a node similar to itself. This graph defines a relation: xRy iff $w\big((x, y)\big) = +1$ or $x = y$. It is obvious, that this relation R is tolerance relation: it is reflexive and symmetric. Let p denote a clustering (partition) on this graph and let $p(x)$ be the set of vertices that are in the same cluster as x. In a partition p we call an edge (x, y) a conflict if $w\big((x, y)\big) = +1$ and $x \notin p(y)$ or $w\big((x, y)\big) = -1$ and $x \in p(y)$. The cost function is the number of these disagreements. Solving the correlation clustering is minimizing its cost function.

It is easy to check that we cannot necessarily find a perfect partition for a graph. Consider the simplest case, given three objects x, y and z, and x is similar to both y and z, but y and z are dissimilar. The number of partitions can be given by the Bell number [1], which grows exponentially. So the optimal partition cannot be determined in a 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, since the number of partitions grows exponentially, it is an NP-hard problem.

In the original definition, a weight of an edge could be only $+1$ or -1. Naturally the function w can be the following as well: $w : E \to [-1, 1]$. If $w\big((x, y)\big) > 0$, then x and y are similar. If $w\big((x, y)\big) < 0$, then x and y are dissimilar and if the weight is 0, then they are neutral. In a partition p we call an edge (x, y) a conflict if $w\big((x, y)\big) > 0$ and $x \notin p(y)$ or $w\big((x, y)\big) < 0$ and $x \in p(y)$. A natural cost function is one in which an edge of weight w incurs a cost of $|w|$ when it is clustered improperly and a cost of 0 when it is correct.

4 Similarity-based Rough Sets

When we would like to define the base sets, we use the background knowledge embedded in an information system. The base sets represent background knowledge (or its limit). In a Pawlakian system, we can say that two objects are indiscernible if all of their known attribute values are identical. The indiscernibility relation defines an equivalence relation. In some cases, it is enough to treat the similar objects in the same way. From the mathematical point of view, similarity can be described by a tolerance relation. Some covering systems are based on a tolerance relation. In these covering spaces, a base set contains objects that are similar to a distinguished member. This means that the similarity to a given element is considered and it generates the system of base sets. Using correlation clustering, we obtain a (quasi- optimal) partition of the universe (see in [2,3,4]). The clusters contain such elements which are typically similar to each other and not just to a distinguished member. In our previous research, we investigated whether the partition can be understood as a system of base sets (see in [12]). By our experiments, it is worth to generate a partition with correlation clustering. The base sets, generated from the partition, have several good properties:

– the similarity of objects relying on their properties (and not the similarity to a distinguished object) plays a crucial role in the definition of base sets;
– the system of base sets consists of disjoint sets, so the lower and upper approximation 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 $y \in U$ object which is in the same cluster as $x \in \mathsf{l}(S)$. If $x \in \mathsf{u}(S)$, then we can say, that every $y \in U$ object which is in the same cluster as $x \in \mathsf{u}(S)$.
– only the necessary number of base sets appears (in applications we have to use an acceptable number of base sets);
– the size of base sets is not too small, or too big.

In the case of singleton clusters, their members cannot be considered as similar to any other objects without increasing the value of the cost function (see in section 3). Therefore, they represent very little information about the similarity. This is the reason why these objects can be treated as outliers. In machine learning, outliers can impair the decisions and result in more inaccurate results. Singleton clusters, therefore, are not considered as base sets. Thus, the approximation space becomes partial (the union of the base sets does not cover the universe).

5 Partiality in Logical Systems

Classical first-order logic gives the necessary tools to prove the soundness of the inference chains. But what inferences could be derived from the background knowledge when it appears in a vague (rough) structure, and what kind of logical systems need to be used to verify the correctness of the information gained from a rough-set-based framework? In this section, we briefly introduce the rough-set-based semantics of first-order logic (or at least we will show one approach), then we will emphasize the threats hiding in partiality.

The semantics of classical first-order logic is based on set theory. The semantic meaning of the predicate symbols is often defined with the help of a positivity domain which is determined by an interpretation of the logical language. The positivity domain of a unary predicate is a subset of a given universe. It contains those objects for which the predicate is said to be true. Predicates with higher arity can be defined similarly using some Cartesian product of the universe as base set [9]. Since an approximation space gives the ability to create the lower or upper approximation of sets, it can be used to approximate the positivity domain of predicates.

In a reasonable logical system, the positivity and negativity domains of the predicates must be disjoint. From this point of view, the use of the lower approximation can represent our certain knowledge (supposing that the lower approximation of a set S is a subset of S). In these circumstances, it is a legitimate expectation that the derived results in the approximated system, if there is any, must coincide with the results we could receive from the crisp (approximation free) world. These expectations can be satisfied by a three-valued logic system where the relationship between an object and a unary predicate can be the following:

– the object certainly belongs to the positivity domain of the predicate, or

– the object certainly belongs to the negativity domain of the predicate, or
– it cannot be determined whether the object belongs to the positivity or negativity domain (the object is it in the border).

The cases above are usually represented with the truth values 1, 0, $\frac{1}{2}$ respectively, so the result is a three-valued logic system [8]. The way how these systems extend the semantics of the logical connectives is crucial. A widely accepted principle to define the existential and the universal quantifiers so that they generalize the zero-order connectives: the disjunction and the conjunction. A partial approximation space requires partial logic system [10]. It gives us the ability to distinguish situations where we cannot say anything certain about the above-mentioned relationship between an object and a predicate:

– we do not know anything about the object (it is missing from the approximation),
– we do not know how the object is related to the predicate (the object is in the border).

The partiality causes the appearance of the truth value gap (usually denoted by 2 which extends the three-valued system). It is also widely accepted, that the connectives are defined so that the truth value gap is inherited.

The pessimistic scenario says that missing knowledge can refute the conclusions derived from our available knowledge. Keeping in mind how we defined the goal of the logic system, we have to adopt this pessimistic approach. In other words, from our viewpoint, it is better to say nothing than to say something unsure.

The disadvantage of the pessimistic approach is that, if we respect all the earlier mentioned widely accepted properties of the partial three-valued logic system, it makes the quantification useless in the case of partial approximation space. For example, to evaluate a universally quantified formula, we need to evaluate the subformula substituted all the objects of the universe in place of the bound variable (with the help of modified assignments). Since the approximation space is partial, at least one evaluation of the subformula will cause truth value gap. An often-used solution is to modify the semantics of the quantifiers so that truth value gap appears only if all subformula evaluations raise truth value gap [7] but it also voids the pessimistic approach. The second approach is to avoid partial approximation spaces.

6 Similarity-based Rough Sets with Annotation

Sometimes it can happen that an object does not belong to a base set (non-singleton cluster) because the system could not consider it similar to any other objects based on the background information. This does not mean that this object is only similar to itself, but without proper information (maybe due to noisy data) the system could not insert it into any base set (non-singleton cluster) to decrease the number of conflicts. Correlation clustering is based on a tolerance relation that represents similarity. The degree of similarity is between -1 and 1. It can also happen that some relevant information is lost when we map the difference of two objects to $[-1, 1]$. In [13] we proposed a possible way to handle this situation. The users can use their knowledge to help the system by

inserting the members of some singletons into base sets (non-singleton clusters). With the help of this manual annotation, the users can put their knowledge into the system. It also decreases the partiality by decreasing the number of singletons. One of the issues with this approach is that it assumes that the user has some background knowledge. It must also be performed manually, so it cannot be used in the case of a huge number of points because it requires too much time.

Artificial neural networks (ANN) are inspired by the biological neural networks in machine learning. They can be used to perform classification. The annotation process is a classification problem as we need to find a proper cluster to an object. Here, the cluster IDs can be treated as class labels. Given the specifics of an approximation space, the deep learning algorithms can help the user in the annotation process. Each layer of the neural network can identify the main properties of the base sets. Based on these characteristics, it can perform an automated comparison and offer options accordingly. In this way, annotation can be executed completely automatically which means it can also be used in the case of huge data. Naturally, during the annotation, not every object must be inserted into a base set. The neural network identifies if an object is an outlier and discards it. Fig. 1 shows how the similarity-based rough sets approximation is constructed from the original data.

In a real-world application, it can happen that an attribute value of an object is missing. This means that it can be unknown, unassigned, or inapplicable (e.g. maiden name of a male). Handling these data is usually a difficult task. In many cases, these values are imputed. It is common to replace them with the mean or the most frequent value. Typically, this gives a rather good result in many situations. In early-stage diabetes, it is not unusual that the patient has only an elevated blood sugar level. If this value is missing for a patient, then it should not be replaced by the mean because the mean is usually the normal blood sugar level. After the substitution, this patient can be treated as healthy. This type of substitution does not consider the information of an object itself but the information of a collection of objects, therefore it can lead to a false conclusion. In this paper, we propose another method to handle missing data. If an object has a missing attribute value, then it cannot be treated as similar to any other objects, so this entity becomes a member of a singleton. As mentioned earlier, such a cluster cannot be treated as a base set. However, with the annotation, it can be placed into a base set. The neural network should only consider the non-missing attribute values and based on them it should find the appropriate base set.

In machine learning, it is very common to combine clustering with classification. Classification always requires class labels. After clustering the data, the cluster IDs can be treated as these class labels. In our approximation space, the non-singleton clusters are treated as base sets. If there is a new object for which we need to find the appropriate base set, then this object can be considered as a singleton. If it is a singleton, then the annotation can be applied to find the fitting base set.

7 Conclusion and Future Work

In [12] the authors introduced a partial approximation space relying on a tolerance relation that represents similarity. The novelty of this approximation space is that the

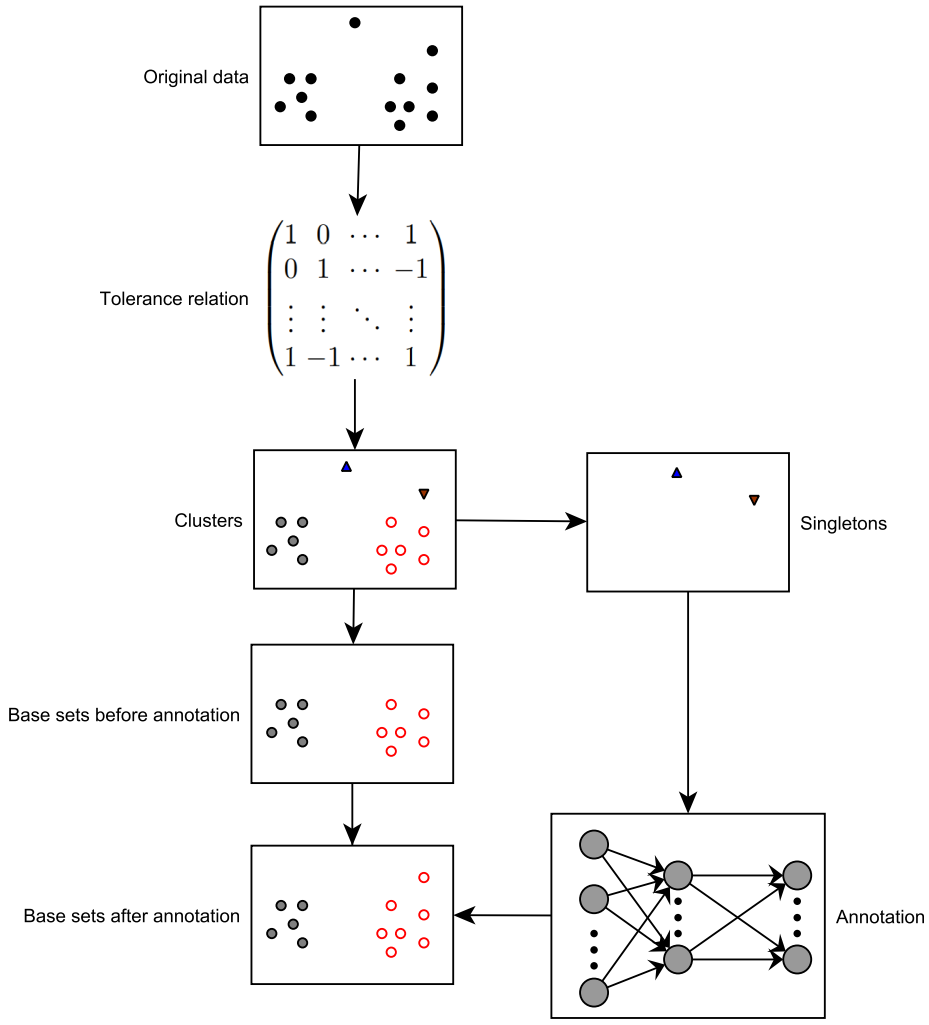

Figure 1. The main steps of the similarity-based rough sets approximation space

systems of base sets are the result of correlation clustering. Thus the similarity is taken into consideration generally. Singleton clusters have no real information in the approximation process, these clusters cannot be taken as base sets, therefore the approximation space is partial. In the present paper, a new possibility is proposed to embed some information into the approximation space. A neural network may decide the status of a member of a singleton cluster. It can be put into a base set, and the approximation of a set changes according to the new system of base sets. This possibility is crucial in practical applications because it decreases the degree of partiality. Neural networks can also be used when we need to decide to which base set a new object belongs. In machine learning, it is common to combine clustering with classification. In our proposed system, a neural network decides and puts the new objects into the chosen base set. This is especially promising for a large number of new objects as we do not need to perform correlation clustering, which is an NP-hard problem, for each object.

Acknowledgement

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. Aigner, M.: Enumeration via ballot numbers. Discrete Mathematics **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: Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, pp. 315–324. Springer Berlin Heidelberg (2013). https://doi.org/10.1109/TKDE.2007.1061
3. Aszalós, L., Mihálydeák, T.: Rough classification based on correlation clustering. In: Rough Sets and Knowledge Technology, pp. 399–410. Springer (2014). https://doi.org/10.1007/978-3-319-11740-9_37
4. Aszalós, L., Mihálydeák, T.: Correlation clustering by contraction. In: Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on. pp. 425–434. IEEE (2015)
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. Kádek, T., Mihálydeák, T.: Some Fundamental Laws of Partial First-Order Logic Based on Set Approximations, pp. 47–58. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-08644-6_5, http://dx.doi.org/10.1007/978-3-319-08644-6_5
8. Mihálydeák, T.: First-order logic based on set approximation: A partial three-valued approach. In: 2014 IEEE 44th International Symposium on Multiple-Valued Logic. pp. 132–137 (May 2014). https://doi.org/10.1109/ISMVL.2014.31
9. Mihálydeák, T.: Partial first-order logical semantics based on approximations of sets. Nonclassical Modal and Predicate Logics pp. 85–90 (2011)

10. Mihálydeák, T.: Partial first-order logic relying on optimistic, pessimistic and average partial membership functions. In: Pasi, G., Montero, J., Ciucci, D. (eds.) Proceedings of the 8th conference of the European Society for Fuzzy Logic and Technology. pp. 334–339 (2013)
11. Mihálydeák, T.: Logic on similarity based rough sets. In: Nguyen, H.S., Ha, Q.T., Li, T., Przybyła-Kasperek, M. (eds.) Rough Sets. pp. 270–283. Springer International Publishing, Cham (2018)
12. 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, https://doi.org/10.1007/978-3-319-60840-2_7
13. 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.) Rough Sets. pp. 88–100. Springer International Publishing, Cham (2018)
14. Pawlak, Z.: Rough sets. International Journal of Parallel Programming **11**(5), 341–356 (1982)
15. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information sciences **177**(1), 3–27 (2007)
16. Pawlak, Z., et al.: Rough sets: Theoretical aspects of reasoning about data. Systern Theory, Knowledge Engineering and Problem Solving, Kluwer Academic Publishers, Dordrecht, 199l **9** (1991)
17. Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundamenta Informaticae **27**(2), 245–253 (1996)
18. Zimek, A.: Correlation clustering. ACM SIGKDD Explorations Newsletter **11**(1), 53–54 (2009)