An approach to solution of constrained clustering problems using the constraint programming paradigm and the multiset theory

A A Zuenko¹, O V Fridman¹, O N Zuenko¹ and O G Zhuravleva²

¹ Institute of Informatics and Mathematical Modeling – Subdivision of the Federal Research Centre "Kola Science Centre of the Russian Academy of Sciences (IIMM KSC RAS) 24A, Fersman str., Apatity, Murmansk region, 184209, Russia
² Mining Institute – Subdivision of the Federal Research Centre "Kola Science Centre of the Russian Academy of Sciences" (MI KSC RAS) 24, Fersman str., Apatity, Murmansk region, 184209, Russia

E-mail: zuenko@iimm.ru

Abstract. An integrated approach is proposed to solve the Constrained Clustering problems which are oriented on taking the opinion of a group of experts into account. A cluster and the objects to be clustered are proposed to be represented as multisets, and the distance between these to be determined by metrics in the Petrovsky multisets space. The approach is implemented within the Constraint Programming paradigm. In so doing, the significant complexity is in organizing effective processing the qualitative constraints, namely, the rules for assigning the objects to one or different clusters. The qualitative constraints are proposed to be represented and processed as table constraints of a new type, i.e., as the smart-tables of D-type. The main attention is confined to the problem of how to reduce the amount of constraints and how to simplify the Constrained Clustering problem. It is proposed to generate constraints for some pairs of objects rather than for all of them, with the generation based on a priori interval estimation of the optimal value of the clustering criterion. To do that, a modified method of multisets hierarchical clustering has been proposed. The approach proposed allows a global optimum to be found for the Constrained Clustering problems considered.

1. Introduction
The shortcoming of most “classical” clustering methods is that the background knowledge from the subject domain is not taken into account. Placing two objects into one cluster only on the basis of the metric may appear to be an incorrect operation in terms of semantics. At present, to solve this problem, an approach referred to as Constrained Clustering [1, 2], is proposed and being developed. In this approach analysis is made not only of the distance between the objects but also of the constraints on combinations of the objects placed / not placed into one and the same cluster (different clusters). That is why the Constrained Clustering problem is also referred to as a semi-supervised clustering.

A typical approach to the solution of the Constrained Clustering problems involves modifying the well known local search methods (for instance, k-means method), taking into account the user constraints, but the present approach allows us to find only a local optimum [3-5]. The studies...
presented in the paper are actual because we need the systematic search methods that would help find a global optimum in high-dimensional spaces.

A clustering problem, with the partition diameter minimization criterion, is taken as an example, and the paper proposes an integrated approach which can be applied in development of systematic search methods to be used in solving the Constrained Clustering problems for group decision making. The approach proposed is implemented within the Constraint Programming paradigm [6, 7] and is based on the representation of the initial objects as multisets, and on the use of the Petrovsky multisets space metrics [8]. In contrast to other studies on group classification [9], the proposed approach is aimed at finding the global optimum and allows one to analyze additional expert constraints.

2. Materials and methods
To solve the problem within the constraint programming paradigm, we should specify a model of the problem, and formulate the mechanisms of constraints inference and search.

A model of the problem
Let it is necessary to partition $n$ objects $O = \{o_1, ..., o_n\}$ into $k$ clusters, so that the partition diameter is minimal among all the possible partitions. Remind that the partition diameter is the maximum diameter for all the partition clusters. The cluster diameter is the maximum distance between any two points belonging to this cluster. The model allows finding a partition, provided that the exact number $k$ of the final clusters is not specified, only interval $k \in [k_{\min}, k_{\max}]$.

The model presented in [2] was used as a basic model to solve the Constrained Clustering problem. The variables $G = \{G_1, ..., G_n\}$ are used to specify the points (objects) of the clusters. The domain of each variable is a set of indices of all the possible clusters $\{1, ..., k_{\max}\}$. Assignment $G_i = c$ means that point $o_i$ is placed into cluster $c$.

In formulating the Constrained Clustering problem, a number of constraints are specified in addition to a matrix of the distances between the objects $[d_{ij}]$. Let’s enumerate the constraints mentioned in [2].

$Precede([G_1, ..., G_n],[1..k_{\max}])$ is a constraint to be specified to avoid symmetrical solutions: only one complete assignment of the variable values from $G = \{G_1, ..., G_n\}$ corresponds to one of the possible partitions containing at least $k_{\min}$ different clusters and at most $k_{\max}$ different clusters.

Consider how, with the help of constraints, a requirement that $k$ should belong to interval $[k_{\min}, k_{\max}]$, is formalized. The condition on the upper boundary $k_{\max}$ of the interval for $k$ is taken into account in specifying a set of possible cluster indices $\{1, ..., k_{\max}\}$. The condition on the lower boundary $k_{\min}$ of the interval considered is taken into account by the following constraint: $AtLeast(G,k_{\min},1)$. According to this constraint, at least one of variables $G = \{G_1, ..., G_n\}$ takes the value $k_{\min}$ in the resulting complete assignment.

To minimize the partition diameter, a constraint to be generated for each pair of objects (for each element of matrix $[d_{ij}]$), is the following:

$$(d_{ij} > D) \rightarrow (G_i \neq G_j). \quad (1)$$

Here $d_{ij}$ is a constant denoting the distance between the objects $o_i$ and $o_j$. Variable $D$ is the partition diameter, which initially takes the values from interval $[d_{\min}, d_{\max}]$, where $d_{\min}$ and $d_{\max}$ are the minimum and the maximum elements of matrix $[d_{ij}]$.

In addition to the obligatory constraints above, some additional user constraints can be added to the model, according to [2]:

---

IOP Publishing
Journal of Physics: Conference Series 1801 (2021) 012041
doi:10.1088/1742-6596/1801/1/012041
1) \( AtLeast(G, G, \alpha) \) are constraints on the minimal size of any cluster in the partition: the number of points in any cluster of any partition is not less than the specified number \( \alpha \);

2) \( AtMost(G, c, \beta) \) are constraints on the maximal size of the cluster with index \( c \in \{1, \ldots, k_{\text{max}}\} \); the number of points in any cluster of any partition is not greater than the specified number \( \beta \).

3) \( G_i = G_j \) are must-link constraints, according to which the pair of objects \( o_i \) and \( o_j \) is placed into one cluster.

4) \( G_i \neq G_j \) are cannot-link constraints according to which the pair of objects \( o_i \) and \( o_j \) is not placed into one cluster.

In contrast to the studies in [2], this study suggests that a cluster and objects clustered are represented as multisets, and the distance \( (d_{ij}) \) between these is determined by metrics in the Petrovsky multisets space. It allows the class of the considered problems to be extended at the account of the problems which need a decision made by a group of experts, and, if necessary, to use the relevant data aggregation techniques [9].

In addition, a significant problem of the base model [2] is that of organizing effective processing of non-numerical constraints like (1). The problem is that a great number of similar constraints cannot be, in aggregate, effectively processed by the available constraint programming solvers.

Thus, an actual problem of the research is to develop the ways to speed up similar constraints processing (new non-numerical constraints propagation techniques). This field of research is the most fully represented in [10]. Another field of research which is actively being developed in the present study is aimed at generating the constraints of the type (1) only for some pairs of objects rather than for all of them, which allows the dimension of the problem to be substantially reduced. Let’s briefly clarify the main idea. Assume that, as a result of application of some methods, we reduced the initial interval of possible partition diameter values from \( D \in [d_{\text{min}}, d_{\text{max}}] \) to \( D \in [d_1, d_2] \), where \( d_1 \) and \( d_2 \) are new upper and lower boundaries of the interval. In this case, it is not necessary to generate constraints for those matrix elements of distances \( d_{ij} \), for which \( d_{ij} \in [d_{\text{min}}, d_1] \) is true, because constraint (1) is identically true. For those elements of matrix \( d_{ij} \), for which \( d_{ij} \in (d_2, d_{\text{max}}] \) is true, constraint (1) is reduced to become constraint \( (G_i \neq G_j) \), which is simpler to process and propagate. What is left is to generate constraints for the objects, for which \( d_{ij} \in [d_1, d_2] \). The number of the objects left is, as a rule, much smaller than the number of the initial ones. Consider below in detail the approach proposed.

The approach proposed:

Step 1. To estimate the range of values, into which the required optimal diameter of partition must be placed. To determine the initial partition, it is suggested to use the FPF method (Furthest Point First) presented in [2]. This approximate method allows the optimal partition diameter to be estimated. If we consider that the partition diameter obtained by the FPF-method is equal to \( d \), then the optimal partition diameter can be estimated as: \( D \in [(d/2), d] \). Based on the values obtained one can generate cannot-link constraints for the pairs of clusters for which \( d_y > d \). The new constraints generated are added to other constraints set by experts, if any.

Step 2. To specify the upper boundary of the interval \( D \in [(d/2), d] \) determined at Step 1. It is done by multisets hierarchical clustering presented in [9]. A significant modification of this procedure is in that the cannot-link constraints are analyzed in clustering [11]. Being applied, this method increases the efficiency of computational procedures, allowing some variants of union of clusters to be eliminated from consideration. Step 2 results in a new interval of possible values of variable \( D \).

Step 3. To generate constraints for a systematic search of the CSP-problem solutions. Steps 1 and 2 allow generating constraints not for all the pairs of the objects to be clustered, as it has been shown above. Constraints (1) are represented as table constraints, notably the \( D \)-type smart tables, being

### References

[9]

[10]

[11]
proposed by one of the authors [11]. These constraints are processed by highly effective non-numerical constraints satisfaction methods developed by one of the authors.

Step 4. To solve the Constrained Clustering problem generated at the previous step, by the heuristics for searching a variable and its values. The systematic search method is based on the following heuristics used in choosing a variable at the current stage of search: choice is made of the variable whose domain has the minimum number of values. The choice of the variable value is based on the following rule: as the variable represents one of the objects to be clustered and its value is the cluster’s number, the variable is assigned the number of the cluster that is closer to the object considered (calculation is made of the distances between the corresponding multisets).

3. Results and Discussion
Consider the example. Let the objects to be clustered are 14 cells into which one of the sites of the rock mass located within the Kukiswumchorr apatite-nepheline deposit, is divided. The aim of clustering is to reveal the zones with different seismic activity. Each seismic event assigned to a certain spatial cell is described by a certain set of factors that, according to the opinion of experts, influence the occurrence of seismic events. according to the opinion of the experts, induce seismic events.

The input data are presented in the table, in which the groups of seismic events are represented as a set of multisets. The representation of the objects, which are characterized by quantitative and/or qualitative attributes and being presented in several versions (copies) in a form of multisets allows qualitative attributes not to be transformed into quantitative ones when calculation procedures are being made.

The attributes (factors) used are as follows: F1 – fault 1; F2 – fault 2; Sb1 – stope boundaries; Sb2 – stope boundaries of the overlying level; W – workings; Ob – ore body; Hr – host rocks; Ob/Hr – ore body/host rocks; Hw – hanging wall of the ore deposit; Fw – foot wall of the ore body; N – the number of seismic events in a cell.

Table 1 is a part of the table with input data used in clustering. In each cell of the table there are two elements of a multiset. For instance, for the factor F1 in the first row, “2 0” means that the multiplicity of value “factor F1 is present” is equal to 2, and the multiplicity of value “factor F1 is absent” is equal to 0. Each row of the table is a multiset. This table can represent the opinion of a single expert, or the aggregate of opinions of several experts in the case of group decision-making.

| Cell | F1 | F2 | Sb1 | Sb2 | Ob | Hr | Ob/Hr | W | Hw | Fw | N |
|------|----|----|-----|-----|----|----|-------|---|----|----|---|
| 1    | 2 0| 2 0| 2 0 | 2 0 | 0 2| 2 0| 2 0    | 0 | 2 0| 2 0| 2 |
| 2    | 0 2| 2 0| 2 0 | 2 0 | 0 2| 2 0| 2 0    | 0 | 2 0| 2 0| 2 |
| 3    | 6 0| 6 0| 6 0 | 6 0 | 6 0| 6 0| 6 0    | 6 | 6 0| 6 0| 6 |
| …   |

Table 2 shows the calculation results for the distances between the initial clusters (one object – one cluster). In calculations, use was made of metrics.

\[ d_{i_1}(o_i, o_j) = \sum_{l=1}^{n} |k_{A_i}^{l}(x_i^{l}) - k_{A_j}^{l}(x_j^{l})|, \]  

(2)

where \( A_i \) and \( A_j \) are multisets corresponding to objects \( o_i \) and \( o_j \).

Having applied the FPF method, we obtain the following: the first cluster includes objects 5, 6; the second cluster – objects 7-10; the third cluster – objects 1-4, 11-14. The first, second and third cluster diameters are 832, 930, and 344, respectively. The partition diameter is 930. We get the following estimate of the interval of values: \( D \in [465, 930] \).
After that the seismic events are being clustered according to a modified method of hierarchical clustering. Take into account cannot-link constraints for $d_{ij} > 930$, which are obtained by the FPF-method (marked in Table 2 in dark colour).

Table 2. The dimension matrix.

|   | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1 | 0  | 4  | 40 | 84 | 1062| 1264| 446| 1088| 242| 442 | 104 | 328| 10  | 66 |
| 2 | 4  | 0  | 44 | 80 | 1066| 1262| 450| 1084| 246| 438 | 108 | 324| 10  | 62 |
| 3 | 40 | 44 | 0  | 52 | 1046| 1272| 438| 1104| 250| 454 | 112 | 344| 60  | 58 |
| 4 | 84 | 80 | 52 | 0  | 1050| 1260| 480| 1100| 270| 450 | 140 | 340| 100 | 72 |
| 5 | 1062| 1066|1046|1050| 0  | 832 | 980 | 1294| 1020|1156| 1030| 1006|1070|1098|
| 6 | 1264| 1262|1272|1260|832 | 0  | 1260| 936 | 1260|1083|1250| 1132|1280|1288|
| 7 | 446 | 450 | 438 | 480 |980 | 1260| 0  | 810 | 200 |104 | 410 | 386 | 450 | 478 |
| 8 | 1088| 1084|1104|1100|1294| 936 | 810 | 0  | 930 | 650|1080 |1016| 1080|1108|
| 9 | 242 | 246 | 250 | 270 |1020|1260| 200 | 930 | 0  | 280 | 210 | 270 | 250 | 278 |
| 10 |442 | 438 | 454 | 450 |1156|1088| 104 | 650 | 280 | 0  | 430 | 366 | 430 | 458 |
| 11 |104 | 108 | 112 | 140 |1030|1250| 410 |1080 | 210 | 430 | 0  | 240 | 100 | 114 |
| 12 |328 | 324 | 344 | 340 |1006|1132| 386 |1016 | 270 | 366 | 240 | 0  | 320 | 320 |
| 13 |10  | 10  | 60  | 100 |1070|1260| 450 |1080 | 250 | 430 | 100 | 320 | 0  | 70  |
| 14 |66  | 62  | 58  | 72  |1098|1288| 478 |1108 | 278 | 458 | 114 | 320 | 70  | 0  |

At each step of hierarchical clustering, all the constraints imposed on each cluster, are imposed on a new united cluster. In calculating the distances by equation (2), only legal combinations (non-coloured in Table 2) are analyzed, which allows the amount of calculations to be substantially reduced. Some steps of clustering are shown in Tables 2-4. The cells with united clusters are in boxes.

Table 3. Step 11 in clustering.

|     | 1-4,9,11-14 | 5   | 6   | 7,8,10 |
|-----|-------------|-----|-----|--------|
| 1-4,9,11-14 | 0  | 950 | 1134| 1718   |
| 5   | 950         | 0   | 832 | 1424   |
| 6   | 1134        | 832 | 0   | 1456   |
| 7,8,10 | 1718   | 1424| 1456| 0      |

Table 4. Final partition.

|     | 1-4,9,11-14 | 7,8,10 | 5,6  |
|-----|-------------|--------|------|
| 1-4,9,11-14 | 0  | 1718 | 1912 |
| 7,8,10  | 1718 | 0   | 1786 |
| 5,6    | 1912 | 1786| 0    |

The final partition: the first cluster – objects 1-4, 9, 11-14; the second cluster – objects 7, 8, 10; the third cluster – objects 5, 6. The clusters diameters are 344, 810 and 832, respectively. Thus, the upper boundary of clustering is updated: $D \in [465, 832]$.

It is clear now that it is not necessary to generate constraints for the elements of the distance matrix, for which $d_{ij} < 465$. The constraints ($G \neq G'$) are generated only for the elements of the distance matrix, for which $d_{ij} > 832$. There are only five elements left of the distance matrix, which satisfy the condition $d_{ij} \in [465, 832]$ (if we look at the top triangle of the distance matrix): $d_{4,7} = 480$, $d_{5,6} = 832$, $d_{7,8} = 810$, $d_{7,14} = 478$, $d_{8,10} = 650$. Hence, there will be only five constraints of the type (1) generated for the Constrained Clustering problem. It should be noted that without preparatory steps, such as the application of the FPF-method and hierarchical clustering, with the constraints taken
into account, the number of the generated constraints of the type (1) would be equal to \((14\cdot13)/2 = 91\).

4. Conclusion
The approach proposed is implemented within the Constraint Programming paradigm and includes the elements, each of which is of scientific novelty: 1) the method of multisets clustering to be used in estimation of the optimal value of the clustering criterion, taking into account the subject domain constraints; 2) the constraints generation method based on the estimation obtained and on the representation and processing the qualitative constraints in a kind of a new type of table constraints, notably, the D-type smart tables; 3) the solution search method to be used in solving the CSP, based on the original heuristics and the author’s algorithms of non-numerical constraints propagation. The methods proposed provide, in aggregate, a possibility to produce a global optimum for the Constrained Clustering problems of high space complexity.

5. Acknowledgments
The study was funded by the Russian Foundation for Basic Research (RFBR), according to the research project № 18-07-00615a, 20-07-00708 a, 17-29-07021 ofi_m.

References
[1] Wagsta K and Cardie C 2000 Clustering with instance-level constraints. Proc. of the 17th Int. Conf. on Machine Learning pp 1103-1110
[2] Dao T, Duong K and Vrain C 2017 Constrained Clustering by Constraint Programming. Artificial Intelligence Journal 244 70-94
[3] Pelleg D and Baras D 2007 K-Means with Large and Noisy Constraint Sets In Machine Learning: ECML vol 4701 of Lecture Notes in Computer Science (Springer Berlin Heidelberg) 674–682
[4] Davidson I and Ravi S S 2005 Agglomerative hierarchical clustering with constraints: Theoretical and empirical results Proc. of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases pp 59-70
[5] Wang X and Davidson I. 2010 Flexible constrained spectral clustering. Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp 563-572
[6] Russel S and Norvig P 2010 Artificial Intelligence: A Modern Approach (3rd ed. Prentice Hall) p 1132
[7] Rossi F, van Beek P and Walsh T 2008 Chapter 4 “Constraint Programming”. In: Foundations of Artificial Intelligence 3 pp 181-211
[8] Petrovsky A 2019 Metrics in multiset spaces Journal of Intelligent and Fuzzy Systems 36(4) 3073-3085
[9] Petrovsky A 2018 Hierarchical aggregation of object attributes in multiple criteria decision making Artificial intelligence Proc. of the 16th Russian conf. Communications and Information Science vol 934 (Springer Nature Switzerland AG) pp 125-137
[10] Zuenko A, Oleynik Y, Yakovlev S and Shemyakin A 2020 Matrix-Like Representation of Production Rules in AI Planning Problems Proc. of the Fourth Int. Scientific Conf. Intelligent Information Technologies for Industry (IITI’19) Advances in Intelligent Systems and Computing vol 1156 (Springer, Cham) pp 393-402
[11] Zuenko A A, Fridman O V and Zhuravleva O G 2019 Application of constrained clustering methods for the study of technogenic seismicity Bulletin of Voronezh State University Ser.: System analysis and information technologies 3 29-41