A Study of Automatic Clustering Based on Evolutionary Many-objective Optimization

Shuwei Zhu
School of Electronics and Information Engineering, Tongji University, Shanghai, China
zswjiang@163.com

Lihong Xu
School of Electronics and Information Engineering, Tongji University, Shanghai, China
xulhk@163.com

Leilei Cao
School of Electronics and Information Engineering, Tongji University, Shanghai, China
mcaoleilei@sina.com

ABSTRACT

Automatic clustering problems, which need to detect the appropriate clustering solution without a pre-defined number of clusters, still remain challenging in unsupervised learning. In many related works, cluster validity indices (CVIs) play an important role to evaluate the goodness of partitioning of data sets. However, there is no CVI that is likely to ensure reliable results for different structures of data. In this paper, we present a study of evolutionary many-objective optimization (EMaO) based automatic clustering, in contrast to the weighted sum validity function defined in literature, several validity functions (more than 3) are considered to be optimized simultaneously here. Since the research of EMaO is still in its fancy, we take four state-of-the-art EMaO algorithms into consideration as the underlying optimization tool. To be more applicable and efficient for clustering problems, the encoding scheme and genetic operators are redesigned. Experiments show that, for the purpose of this study, it is promising to address automatic clustering problems based on a suitable EMaO approach.

CCS CONCEPTS

• **Computing methodologies → Optimization algorithms**; [1]

KEYWORDS

Automatic clustering; Many-objective optimization; Cluster validity indices; Cluster number; Evolutionary algorithm

1 INTRODUCTION

The problem of determining the best estimation of K is known as the automatic clustering problem, which is still an open issue. There are various research works that attempt to tackle this issue, among which the approaches using nature-inspired metaheuristics (either single-objective or multi-objective versions) have gained a lot of attention. The usual way is to set the value of K from the interval $[K_{\min}, K_{\max}]$, and then a cluster validity index (CVI) is adopted to evaluate the goodness of each clustering solution, such that CVIs can be used as objective functions to be optimized.

Thanks to the potential of multi-objective metaheuristics, multi-objective clustering (MOC) algorithms often show superiority when dealing with datasets with different structures. Commonly, they perform search by optimizing several clustering objective functions (usually two CVIs). It is known that the MOCK algorithm[1] is the most representative one in the field of MOC to address automatic clustering problem. Recently, an improved version of MOCK is proposed[2], named as Δ-MOCK, which can decrease the computational overhand as well as reduce the search space by a large margin. Up to now, there is no single CVI that shows superiority over others to deal with all cases. In order to take advantage of more CVIs to detect the cluster number as well as finding a good partition result, we present a study of evolutionary many-objective optimization based automatic clustering (EMaOC) using five objectives.

2 THE PROPOSED AUTOMATIC MANY-OBJECTIVE CLUSTERING

There are, basically, several major components in the framework of EMaOC, such as the encoding scheme, evolutionary operators, objective functions optimized, and determination of the final clustering result from non-dominated solutions[3]. In this study, we develop the automatic many-objective clustering based on the well-performed algorithm Δ-MOCK[2]. Hence, the locus-based encoding scheme, the uniform crossover and the neighborhood-biased mutation of Δ-MOCK are adopted here, whereas, the underlying optimization tool and the objectives optimized should be specified to suit the many-objective optimization property.

Note that, the application of different EMaO algorithms in the proposed automatic clustering framework is studied. Here, four popular EMaO algorithms, namely SPEA-II-SDE[4], NSGA-III[5], MOEA/DD[6], and RVEA[7] are considered. The proposed framework is outlined in Algorithm 1. For the sake of convenience, we denote the studied clustering approaches as Δ-MaOCK1(SPEA-II-SDE), Δ-MaOCK2 (NSGA-III), Δ-MaOCK3 (MOEA/DD), and Δ-MaOCK4 (RVEA), respectively.

Algorithm 1 EMaO based automatic clustering framework

1: Input parameters
2: Pre-computation for the loaded data set
  1) Compute nearest neighbors
  2) Compute the MST
3: Initialize the population
4: Execute the EMaO approach
  1) Mating selection (except for NSGA-III)
  2) Reproduction operation to generate offspring solutions
  3) Evaluate each offspring solution by five objectives
  4) Environmental selection
5: Output the non-dominated solutions

To obtain the initial population, a key parameter $k_{\max}$ should be set during the initialization of Δ-MOCK based on the true cluster number k^*, which may not be obtained in most real-world cases. In this study, we set the value of $k_{\max}$ to be $\sqrt{n}$ as

GECCO '18 Companion, July 15-19, 2018, Kyoto, Japan

https://doi.org/10.1145/3205651.3205759

recommended by many related works in literature. As the same as Δ-MOCK, the intra-cluster variance (VAR) and connectivity (Cnn) are employed as two objectives to keep a trade-off trend when k increases from 2 to k_{max}. They evaluate basically different but equally desirable desirable qualities of a clustering solution. Besides, another three famous CVIs, namely CH, DB, and DU_{53}, are taken into desirable qualities of a clustering solution. Besides, another three famous CVIs, namely CH, DB, and DU_{53}, are considered to be optimized, since they have shown a promising performance for automatic clustering[8]. In view of the length of this paper, the mathematic function of these CVIs can be referred to Ref. [8]. Note that, the time complexities of them are both $O(nkd)$ that is linearly scaled with respect to the number of points n and dimensions d, as well as k.

3 THE EXPERIMENT

3.1 Parameter settings and datasets

Experiments are conducted on the synthetic data sets, which have been considered in the analysis of Δ-MOCK. Table 1 presents the characteristics of these datasets. For all algorithms, we set the population size as N=50, the total generations T_{max}=50. For the evaluation of all clustering solutions, the external clustering validity criterion ARI (in [0,1]) is employed, which reaches the value of one as a perfect matching with the true clustering labels.

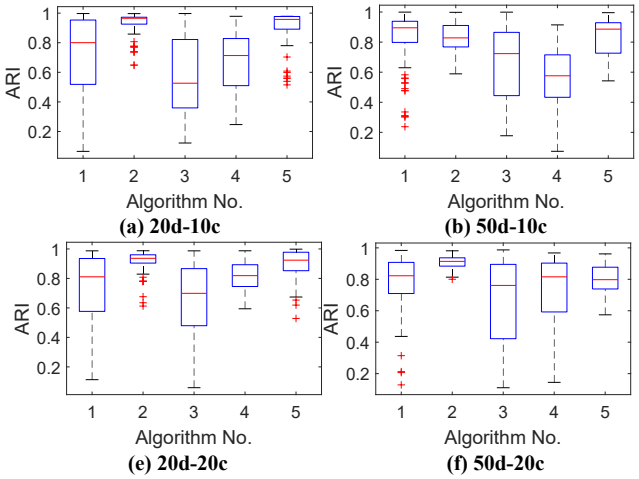
Table 1 The characteristics of datasets

Dataset	Number of clusters	Number of attributes	Number of points
20d-10c	10	20	3282
50d-10c	10	50	3184
20d-20c	20	20	5944
50d-20c	20	50	6247

3.2 Experiments and analysis

In this section, each optimization approach is unrepeatably conducted five times in the step 4 of Algorithm 1. Then, the produced non-dominated solutions of all five executions by each algorithm are collected. Fig. 1 shows the box plots of ARIs which is computed by the solutions of different algorithms on the four data sets. It is evident that Δ-MOCK and Δ-MaOCK2 (especially the latter) can produce wide ranges of ARIs in most cases. For Δ-MOCK, given $k_{max} = \sqrt{n}$ that ranges from ~55 (for 10c-type data sets) to ~75 (for 20c-type data sets), NSGA-II can produce a widely distributed values of k by optimizing VAR and Cnn, leading to some small ARIs if k is very small (as low as 2 or 3) or very high. For Δ-MOCK2, the NSGA-III algorithm focuses more on the diversity of solutions by replacing the environmental selection strategy of NSGA-II. Note that, NSGA-III can perform well on benchmark test suits since the defined problems are in basis of well distributed Pareto optimal solutions. However, this may not be true for other optimization problems with irregular shape of PF, especially real-world cases.

Overall, the Δ-MaOCK1 method which proposed based on SPEA-II-SDE performs better than the others. This may be owing to the fact that it is the extension of the algorithm SPEA-II with shift-based density estimation, which can produce solutions of better proximity without considerable sacrifice of diversity here.



*Note: Index 1~5 in x-axis, respectively, denotes clustering algorithms Δ-MOCK, and Δ-MaOCK1~Δ-MaOCK4.

Figure 1: Box-plots for the ARI comparisons.

4 CONCLUSIONS

The work done here focuses at dealing with automatic clustering using EMaO, for which four state-of-the-art EMaO methods are considered. Experiments show that, the SPEA-II-SDE based EMaOC (named as Δ-MaOCK1) technique shows a significant superiority over the others in terms of ARIs from a statistical viewpoint. Hence, it is considered to be a promising direction to address automatic clustering problems based on a suitable EMaO approach. Future research will be concentrated on the decision making step to determine the final clustering result.

ACKNOWLEDGMENTS

This work is based in part upon the National Natural Science Foundation of China under grant 61573258, and the U.S. National Science Foundation's BEACON Center for the Study of Evolution in Action, funded under Cooperative Agreement DBI-0939454.

REFERENCES

[1] Julia Handl and Joshua Knowles. An Evolutionary Approach to Multiobjective Clustering. 2007. *IEEE Transactions on Evolutionary Computation* 11, 1: 56-76.

[2] Mario Garza-Fabre, Julia Handl, and Joshua Knowles. An Improved and More Scalable Evolutionary Approach to Multiobjective Clustering. 2017. *IEEE Transactions on Evolutionary Computation*. DOI: https://doi.org/10.1109/TEVC.2017.2726341.

[3] Shuwei Zhu, Lihong Xu. Many-objective fuzzy centroids clustering algorithm for categorical data. 2018. *Expert Systems with Applications*, 96: 230-48.

[4] Miqing Li, Shengxiang Yang, and Xiaohui Liu. Shift-based density estimation for Pareto-based algorithms in many-objective optimization. 2014. *IEEE Transactions on Evolutionary Computation* 18, 3: 348-65.

[5] Kalyanmoy Deb and Himanshu Jain. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: solving problems with box constraints. 2014. *IEEE Transactions on Evolutionary Computation* 18, 4: 577-601.

[6] Ke Li, Kalyanmoy Deb, Qingfu Zhang, and Sam Kwong. An evolutionary many-objective optimization algorithm based on dominance and decomposition. 2015. *IEEE Transactions on Evolutionary Computation*, 19, 5: 694-716.

[7] Ran Cheng, Yaochu Jin, Markus Olhofer and Bernhard Sendhoff. A Reference Vector Guided Evolutionary Algorithm for Many-Objective Optimization. 2016. *IEEE Transactions on Evolutionary Computation* 20, 5: 773-91.

[8] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, and et al. An extensive comparative study of cluster validity indices. 2013. *Pattern Recognition* 46, 1: 243-56.