Minimum Spanning Tree-based Clustering of Large Pareto Archives

Andrzej Jaszkiewicz
Poznan University of Technology, Faculty of Computing
andrzej.jaszkiewicz@put.poznan.pl

ABSTRACT

In this paper we propose the use of Minimum Spanning Tree-based clustering to recursively cluster large sets of potentially Pareto-optimal solutions. We present preliminary results for the multiobjective traveling salesperson problem. The clustering is based on the distances between solutions defined in the decision space but it generates clusters that correspond to clearly defined regions in the objective space.

CCS CONCEPTS

• **Computing methodologies → Discrete space search**; • **Applied computing → Multi-criterion optimization and decision-making**;

KEYWORDS

Multiobjective optimization, clustering

ACM Reference Format:
Andrzej Jaszkiewicz. 2019. Minimum Spanning Tree-based Clustering of Large Pareto Archives. In *Genetic and Evolutionary Computation Conference Companion (GECCO '19 Companion), July 13–17, 2019, Prague, Czech Republic.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3319619.3326883

1 INTRODUCTION

Recent advances in multiobjective optimization with evolutionary and metaheuristic algorithms made it possible to efficiently generate, manage, and evaluate even large sets of potentially Pareto-optimal solutions (Pareto archives). Examples of such recent advances are:

- ND-Tree data structure and algorithms for efficient update and management of large Pareto archives [7].
- (Many objective) Pareto local search method for efficient generation of large and high quality Pareto archives [5, 9, 10].
- Efficient methods for calculation of the hypervolume indicator [1, 4, 11, 12, 14].

These recent advances allow avoiding the traditional approach of using Pareto archives with a bounded size [2]. Note that the use of a bounded archive always reduces the quality of the archive since some potentially Pareto-optimal solutions are discarded, while each of the discarded solutions could be the one that would be selected by the decision maker (DM) if the full archive was known. The use of bounded archives is especially disadvantageous in the case of many-objective optimization, because in a many dimensional space it is difficult to well represent a large set with a smaller sample of solutions.

A large Pareto archive composed of thousands or even millions of solutions may constitute, however, a challenge for the DM who needs to select the single best solution for implementation in practice. A possible solution to this problem is clustering of similar Pareto-optimal solutions. The DM may start by seeing a smaller number of solutions representative to the main clusters and then go inside one or more interesting clusters. This process may be recursively repeated at different levels until the DM is satisfied with a solution. Of course, the DM may also retract the search to a higher level. In this way an interactive analysis of the Pareto archive is obtained [6]. The clustering may be performed in the objective space, however, in practice the DMs are often interested also in other aspects of solutions than the values of objectives. Thus, a solution representative to a cluster should be representative also in the decision space, and the cluster should contain solutions close in both the objective and the decision spaces.

In this paper we propose to use Minimum Spanning Tree-based (MST-based) clustering [3, 8, 13]. Namely, we use the well-known Kruskal's algorithm to build a spanning tree based on the distances of all solutions in the decision space. Then we iteratively remove the longest edges obtaining lower level spanning trees and corresponding clusters. The advantages of this approach are:

- Efficient, polynomial-time algorithms for finding MST like the Kruskal's algorithm may be used.
- A hierarchy of clusters is naturally obtained.
- MST-based clustering is capable of finding clusters of irregular shape, while some other methods tend to prefer some specific shapes of clusters.

2 PRELIMINARY RESULTS

We perform a proof of concept using a three objective traveling salesperson (TSP) problem with 100 nodes. As a distance measure we use the number of different edges between two solutions. We have generated more than 6 000 potentially Pareto-optimal solutions using many-objective Pareto local search [5] initialized with seed solutions generated with Lin-Kernighan heuristic. We start by testing the correlation between the distances in the objective and decisions spaces. As the distance in the objective space we use the Euclidean distance. The results are presented in Figure 1. As one can see there is a very strong correlation between the distances in the two spaces. This result is obviously not surprising

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3326883

since it is a typical motivation for applying Pareto local search to multiobjective TSP [9, 10].

Figure 2 shows a 2D projection of the set of potentially Pareto-optimal solutions in the objective space and their split into 30 clusters obtained through MST-based clustering. One can see that the clusters obtained in the decision space correspond to well defined regions in the objective space of quite irregular shapes. One of the clusters (the topmost) is marked and the triangle on the right shows the edges (a coincidence matrix) that apear in at least 80% of solutions from this cluster. In this case, there are 62 such edges (18 of them are common to all solutions from this cluster). In other words, these solutions are very similar in the decision space. In a practical problem such frequent edges could be a meaningful information for the DM. They constitute frequent subpaths that could be presented on a digital map and the DM may be informed that they correspond to some some specific (approximate) objective values. Figure 3 shows a more deailed split into a higher number of clusters and frequent edges for the rightmost cluster. In this case, there are 78 (out of 100) edges that appear in at least 80% solutions and 49 of them are common to all solutions from this cluster.

There are of course a number of open issues in this preliminary work:

- How to evaluate the quality of a split of potentially Pareto-optimal solutions?
- How does MST-based clustering compare to other approaches to clustering in either the objective or decision space?

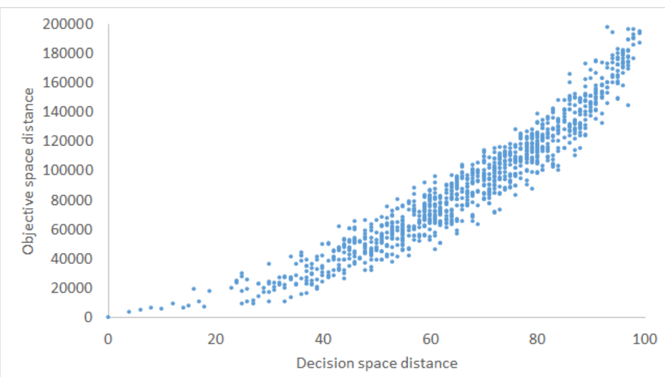

Figure 1: Objective space distance vs. decisions space distance

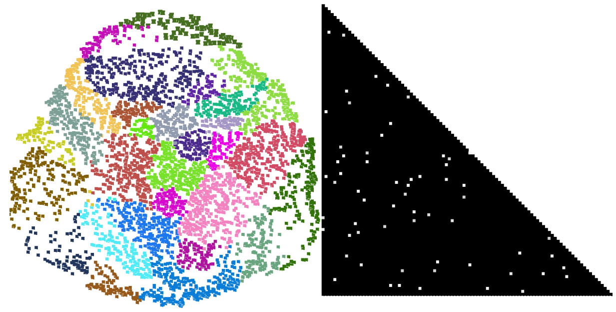

Figure 2: Visualisation of obtained clusters

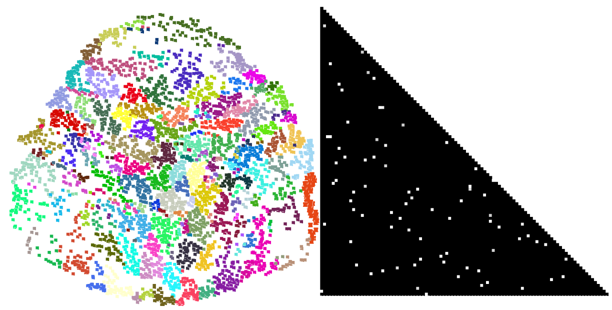

Figure 3: Visualisation of obtained clusters

REFERENCES

[1] Nicola Beume, Carlos M. Fonseca, Manuel Lopez-Ibanez, Luis Paquete, and Jan Vahrenhold. 2009. On the Complexity of Computing the Hypervolume Indicator. *IEEE Transactions on Evolutionary Computation* 13, 5 (OCT 2009), 1075–1082. https://doi.org/{10.1109/TEVC.2009.2015575}

[2] J. E. Fieldsend, R. M. Everson, and S. Singh. 2003. Using unconstrained elite archives for multiobjective optimization. *IEEE Transactions on Evolutionary Computation* 7, 3 (June 2003), 305–323. https://doi.org/10.1109/TEVC.2003.810733

[3] O. Grygorash, Y. Zhou, and Z. Jorgensen. 2006. Minimum Spanning Tree Based Clustering Algorithms. In *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*. 73–81. https://doi.org/10.1109/ICTAI.2006.83

[4] Andrzej Jaszkiewicz. 2018. Improved quick hypervolume algorithm. *Computers & Operations Research* 90 (2018), 72 – 83. https://doi.org/10.1016/j.cor.2017.09.016

[5] Andrzej Jaszkiewicz. 2018. Many-Objective Pareto Local Search. *European Journal of Operational Research* 271, 3 (2018), 1001 – 1013. https://doi.org/10.1016/j.ejor.2018.06.009

[6] Andrzej Jaszkiewicz and Jürgen Branke. 2008. *Interactive Multiobjective Evolutionary Algorithms*. Springer Berlin Heidelberg, Berlin, Heidelberg, 179–193. https://doi.org/10.1007/978-3-540-88908-3_7

[7] A. Jaszkiewicz and T. Lust. 2018. ND-Tree-Based Update: A Fast Algorithm for the Dynamic Nondominance Problem. *IEEE Transactions on Evolutionary Computation* 22, 5 (Oct 2018), 778–791. https://doi.org/10.1109/TEVC.2018.2799684

[8] R. Jothi, Sraban Kumar Mohanty, and Aparajita Ojha. 2018. Fast approximate minimum spanning tree based clustering algorithm. *Neurocomputing* 272 (2018), 542 – 557. https://doi.org/10.1016/j.neucom.2017.07.038

[9] T. Lust and A. Jaszkiewicz. 2010. Speed-up techniques for solving large-scale biobjective TSP. *Computers & Operations Research* 37, 3 (2010), 521 – 533. https://doi.org/10.1016/j.cor.2009.01.005

[10] Thibaut Lust and Jacques Teghem. 2010. Two-phase Pareto local search for the biobjective traveling salesman problem. *Journal of Heuristics* 16, 3 (01 Jun 2010), 475–510. https://doi.org/10.1007/s10732-009-9103-9

[11] Luis M. S. Russo and Alexandre P. Francisco. 2014. Quick Hypervolume. *IEEE Transactions on Evolutionary Computation* 18, 4 (AUG 2014), 481–502. https://doi.org/{10.1109/TEVC.2013.2281525}

[12] Luis M. S. Russo and Alexandre P. Francisco. 2016. Extending quick hypervolume. *Journal of Heuristics* 22, 3 (JUN 2016), 245–271. https://doi.org/{10.1007/s10732-016-9309-6}

[13] X. Wang, X. Wang, and D. Mitchell Wilkes. 2009. A Divide-and-Conquer Approach for Minimum Spanning Tree-Based Clustering. *IEEE Transactions on Knowledge and Data Engineering* 21, 7 (July 2009), 945–958. https://doi.org/10.1109/TKDE.2009.37

[14] Lyndon While and Lucas Bradstreet. 2012. Applying the WFG Algorithm to Calculate Incremental Hypervolumes. In *2012 IEEE Congress on Evolutionary Computation (CEC) (IEEE Congress on Evolutionary Computation)*. IEEE, IEEE, 345 E 47TH ST, NEW YORK, NY 10017 USA. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) / International Joint Conference on Neural Networks (IJCNN) / IEEE Congress on Evolutionary Computation (IEEE-CEC) / IEEE World Congress on Computational Intelligence (IEEE-WCCI), Brisbane, AUSTRALIA, JUN 10-15, 2012.