Information Sciences 382–383 (2017) 179–200

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Less is more: Solving the Max-Mean diversity problem with variable neighborhood search

Jack Brimberg[a], Nenad Mladenović[b,c], Raca Todosijević[b,c], Dragan Urošević[b,*]

[a] Royal Military College of Canada, Kingston, Ontario, Canada
[b] Mathematical Institute, Belgrade, Serbia
[c] LAMIH, Université de Valenciennes et du Hainaut Cambrésis, Valenciennes, France

ARTICLE INFO

Article history:
Received 3 February 2016
Revised 6 December 2016
Accepted 13 December 2016
Available online 14 December 2016

Keywords:
Variable neighborhood search
Diversity/dispersion problem
Maximum-mean

ABSTRACT

Within the broad class of diversity/dispersion problems we find an important variant known as the Max–Mean Diversity Problem, which requires finding a subset of a given set of elements in order to maximize the quotient of the sum of all edges belonging to that subset and the cardinality of the subset. In this paper we develop a new application of general variable neighborhood search for solving this problem. Extensive computational results show that our new heuristic significantly outperforms the current state-of-the-art heuristic. Moreover, the best known solutions have been improved on 58 out of 60 large test instances from the literature. In other words, despite the simplicity of our method, which is a desirable property for any heuristic, we achieve significantly better results than a more complex heuristic that represents the state-of-the-art. Thus, simplicity can lead to more efficient and effective methods: when heuristics are used, less can be more.

1. Introduction

Given a set N of $n = |N|$ elements labeled by integer numbers from 1 to n, and the distance $d_{ij} (= d_{ji})$ between any two elements i and j, a dispersion or diversity problem (DP) consists generally of finding a subset $S \subset N$ so that an objective function based on the distances between elements in S is maximized or minimized. According to the objective function, two main classes of dispersion problems are distinguished: those that use efficiency-based objective functions and those that use equity-based objective functions. An efficiency-based objective function reflects the dispersion quantity for the entire selection S, while an equity-based objective function guarantees equitable dispersion among the selected elements. Among widely studied problems that use efficiency-based objective functions are: the Maximum Diversity Problem (MDP) (see e.g., [7]), whose goal is to find a subset S so that the sum of the distances between the selected elements is maximized, and the Max-Min Diversity Problem (MMDP) (see e.g., [5]), where the goal is to find a subset S so that the minimum distance between the selected elements is maximized. On the other hand, the problems that consider equity-based measures introduced by Prokopyev et al. [17] are: the Maximum Mean Dispersion Problem (Max-Mean DP), Minimum Differential Dispersion Problem (Min-Diff DP), and Maximum Minsum Dispersion Problem (Max-Minsum DP). The first of these problems requires finding a subset S, so that the average distance between the selected elements is maximized, while the second one

* Corresponding author.
E-mail addresses: Jack.Brimberg@rmc.ca (J. Brimberg), nenad@mi.sanu.ac.rs (N. Mladenović), racatodosijevic@gmail.com (R. Todosijević), draganu@mi.sanu.ac.rs (D. Urošević).

http://dx.doi.org/10.1016/j.ins.2016.12.021
0020-0255/

concerns finding a subset S so that the difference between the maximum sum and the minimum sum of the distances from a node in S to the other selected elements in S is minimized. Finally, the Max-Minsum DP consists of finding a subset S so that the minimum sum of distances to the other selected elements is maximized.

Some applications of diversity problems arise in the context of facility location [5,6,11,18,19], maximally diverse/similar group selection (e.g., biological diversity, admissions policy formulation, committee formation, curriculum design, market planning, etc.) [1,2,7,8,12,20], densest subgraph identification [10] and network flow problems [3].

In this paper, we study the Maximum Mean Dispersion Problem (Max-Mean DP). As proved in Prokopyev et al. [17], the Max-Mean DP is strongly NP-hard if the distances (diversity measures) take both positive and negative values. A basic example where distances may have negative values is in the case of identifying a diverse group of people. As noted in Lee et al. [13], the cosine similarity is a popular measure of similarity. It is computed for two given individuals, x, y, as

$$d_{xy} = \frac{\sum_{k=1}^{p} x_k y_k}{\sqrt{\sum_{k=1}^{p} x_k^2}\sqrt{\sum_{k=1}^{p} y_k^2}} \tag{1}$$

where x_i and y_i represent the i-th attribute values of individuals x and y, respectively, which can be positive or negative. The imposed measure d_{xy} reflects the affinity between the individuals and obviously may take both positive and negative values.

Because of the NP-hardness of the Max-Mean DP there is a need for a heuristic method that can provide high-quality solutions in a reasonable amount of time. Some steps in building such heuristics have been already made. In 2013, Marti and Sandoya [14] proposed a hybrid approach which combines GRASP and Path Relinking, and uses a Variable Neighborhood Descent (VND) procedure as a local search. Recently, Carrasco et al. [4] have proposed a dynamic tabu search algorithm that examines three different neighborhood structures.

In this paper, we propose a general variable neighborhood search (GVNS) for the Max–Mean Diversity Problem that uses Sequential Variable Neighborhood Descent (SeqVND) as the local search. Within this SeqVND, three neighborhood structures are examined in an efficient way. The merit of the proposed approach is demonstrated on benchmark instances from the literature, where we obtain significantly better results than the state-of-the-art heuristic with tabu search [4]. For almost all examined large problem instances, we are able to find significantly better solutions than those previously reported, and for all small problem instances, we reach the best known results from the literature. Additionally, we propose a new set of large scale test instances for the Max-Mean DP.

We bring in the term “less is more” [16] to emphasize that the newly proposed method involves just a subset of ingredients used by the current state-of-the-art heuristic for solving the Max–Mean DP [4]. Indeed, three Tabu search based heuristics proposed in [4] use *Add, Drop* and *Swap* neighborhood structures as we do in our VNS based heuristic. However, in [4] they are just a part of several more complex routines such as short and long term memory routines, nested local searches, etc. Hence, our contribution consists in finding the minimum number of ingredients that makes a VNS based heuristic as simple and user friendly as possible, while at the same time achieving, despite its relative simplicity, significantly better results than the state-of-the-art. Therefore, we can conclude that adding many ideas in the search does not necessarily lead to better computational results; on the contrary, sometimes “less can yield more”.

The rest of the paper is organized as follows. The mathematical formulation of the problem is presented in the next section. The new heuristic is described in detail in Section 3, including a description of the data structure used for an efficient implementation of the SeqVND local search, an illustrative example to depict the work of the proposed GVNS, and a way to parallelize GVNS. In Section 4 we first perform a comparison of several SeqVND procedures, and then present a summary of the computational results obtained by our GVNS on a wide range of test problems. In addition, Section 4 includes comparison of Sequential and Parallel GVNS. Finally, Section 5 concludes the paper and gives some possible future research directions.

2. Mathematical formulation

Unlike the other dispersion problems, where the cardinality of the set S is known in advance, the cardinality of S is variable in the Max-Mean DP. Thus, the Max-Mean DP consists of finding a subset $S \subseteq N$, $|S| > 1$, with maximal value $f(S)$, where $f(S)$ represents the mean dispersion of the set $S \subseteq N$, calculated as

$$f(S) = \frac{\sum_{i \in S} \sum_{j \in S} d_{ij}}{2m}, \tag{2}$$

where $m = |S| > 1$ is unknown. Formally, the Max-Mean DP may be stated as:

$$\max \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} x_i x_j}{\sum_{i=1}^{n} x_i}, \tag{3}$$

subject to

$$\sum_{i=1}^{n} x_i \geq 2 \tag{4}$$