Computers & Operations Research 71 (2016) 16–22

Contents lists available at ScienceDirect

Computers & Operations Research

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

A hybrid three-phase approach for the Max-Mean Dispersion Problem

Federico Della Croce [a,b], Michele Garraffa [a,*], Fabio Salassa [a]

[a] D.A.I., Politecnico di Torino, Italy
[b] CNR, IEIIT, Torino, Italy

ARTICLE INFO

Available online 13 January 2016

Keywords:
Max-Mean Dispersion Problem
Fractional combinatorial optimization
Quadratic binary programming
Hybrid heuristics

ABSTRACT

This paper deals with the Max-Mean Dispersion Problem (*Max-Mean DP*) belonging to the general category of clustering problems which aim to find a subset of a set which maximizes a measure of dispersion/similarity between elements. A three-phase hybrid heuristic was developed, which combines a mixed integer non-linear solver, a local branching scheme and a path relinking procedure. Computational results performed on the literature instances show that the proposed procedure outperforms the state-of-the-art approaches.

1. Introduction

In recent years, several mathematical programming based heuristics have been developed in order to solve complex combinatorial optimization problems. These heuristics rely on a general purpose solver for the solution of subproblems which can be more tractable. To this end, linear programming (LP) and mixed integer linear programming (MILP) models have been exploited so far, since they can take advantage of the impressive effort in the improvements of the related solvers. Moreover, there has been a significant evolution in the performances on solving non-linear programming (NLP) models, in particular in specific case of the quadratic–quadratic integer programming (QP–QIP). In this paper, we consider the Max-Mean Dispersion Problem (*Max−Mean DP*), also called Equitable Dispersion Problem (EDP), and propose a hybrid heuristic approach based on the solution of a QIP formulation. *Max−Mean DP* belongs to a general category of clustering problems whose aim is to find a subset M of a set N which maximizes a measure of dispersion/similarity of the elements in M. More formally, suppose N is a set of elements with cardinality n, and D a matrix whose components $d_{i,j}$ (that may be positive, negative or null) indicate distance/proximity between items $i \in N$ and $j \in N$. We assume that the matrix D is symmetric, namely $d_{i,j} = d_{j,i} \forall i, j \in N$, where the values on the diagonal are equal to 0 ($d_{i,i} = 0, \forall i \in N$).

When the measure of dispersion/similarity of the elements in M is the sum of the $d_{i,j}$ s between elements $i, j \in M$ (that is $\sum_{i,j \in M} d_{i,j}$) and the cardinality of subset M is given a priori ($|M| = m$ with m predefined), then we have the Maximum Diversity Problem [5,7], which is known to be strongly NP-Hard. The Maximum Diversity Problem has been referred to with several names, which have been carefully collected in [10]. Recently, it has also been called *k−cluster* problem [9]. Hereafter, we will refer to it as *Max−Sum DP*. When the measure of dispersion/similarity of the elements in M is the minimum of the $d_{i,j}$ s between elements $i, j \in M$ (that is $\min_{i,j \in M} d_{i,j}$) and the cardinality of subset M is given a priori, we have the *Max−Min DP* [3,14], which is also known to be strongly NP-Hard. Finally, when the measure of dispersion/similarity of the elements in M is the average of the distances between elements $i, j \in M$ (that is $\sum_{i,j \in M} d_{i,j}/|M|$), but the cardinality of subset M is not given a priori, we have the *Max−Mean DP* which is the object of this work. It is of interest to exploit the relationship between *Max−Sum DP* and *Max−MeanDP*. Let $OPT_{m=i}(Max{-}Sum\ DP)$ denote the optimal solution value of *Max−Sum DP* for $m = i$. The optimal solution of *Max −Mean DP* can be simply computed by iteratively solving *Max−Sum DP* for $m = 2, \dots, n$ and subsequently taking the maximum $OPT_{m=i}(Max{-}Sum\ DP)/i\ \forall i \in \{2, \dots, n\}$.

This problem has a real importance in fields like architectural space planning and analysis of social networks (as claimed in [11]). Another real-world application is about web pages rank (see [8]). A different application is also shown in [15], where authors aim at selecting individuals with different abilities, in order to determine very productive non-homogeneous work teams. In such domains $d_{i,j}$ s can violate the triangular inequality $d_{i,j} \le d_{i,k} + d_{k,j}\ \forall i, k, j \in N$ and the non-negativity condition $d_{i,j} \ge 0\ \forall i, j \in N$. Finally, the considered problem can be of interest for communities mining (see [16]) where relationships between individuals and/or elements can be either positive or negative, such as like–dislike and trust–distrust. This can be useful for market surveys, pattern recognition and social network analysis and more generally when studies have to be conducted in order to derive specific (balanced) characteristic of subset of elements in a larger community. In this context, it is a common approach to tackle clustering problems with the aim of maximizing a measure of

* Corresponding author at: via Giacomo Puccini 34, 91100 Trapani. Italy.

http://dx.doi.org/10.1016/j.cor.2016.01.003
0305-0548/

the average diversity among individuals or elements where it is important also to take into account the cluster size.

To the authors' knowledge, the state of the art literature on *Max−Mean DP* is quite limited. In [13], *Max−Mean DP* is shown to be strongly NP-Hard whenever $d_{i,j}$ s can take both positive and negative values. Those authors have also presented a mixed integer non-linear programming (MINLP) formulation and an equivalent ILP formulation. In [11], a randomized GRASP with path relinking is proposed for *Max−Mean DP*. The presented computational experiments dealt with a set of real world instances from a social network application. The authors also noted that a specific ILP solver (CPLEX) iteratively applied to the ILP formulation of *Max−Sum DP* reached better performances than the same solver applied just once to the ILP formulation of *Max−Mean DP*.

In this paper, we propose a three-phase hybrid heuristic procedure whose first phase repeatedly solves a QIP formulation of *Max−Sum DP* in order to determine an initial solutions set. The following phases enhance the quality of the initial solutions set by means of a local branching scheme and a path relinking procedure. The proposed approach proved to be computationally superior to the approach proposed in [11]. A preliminary version of this work was presented at ISCO2014 [2]. Here, we provide an extended computational experiments section and significantly improved results due to an enhanced parameters configuration and to the addition of the path relinking procedure.

2. Mathematical formulations

Max−Mean DP has a straightforward non-linear and fractional formulation, as introduced in [13]. Define vector $\mathbf{x} \in \{0,1\}^n$ where, for each component x_i, we have $x_i = 1$ if and only if element i is included in the subset M, otherwise 0. The formulation below follows directly from the definition of *Max−Mean DP*:

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

subject to:

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

$$x_i \in \{0,1\} \quad \forall i \in N \tag{3}$$

Since D is symmetric, it can be written in the following vectorial form:

$$\min -\frac{\frac{1}{2}\mathbf{x}^T D \mathbf{x}}{\mathbf{u}^T \mathbf{x}} \tag{4}$$

subject to:

$$\mathbf{u}^T \mathbf{x} \geq 2 \tag{5}$$

$$\mathbf{x} \in \{0,1\}^n \tag{6}$$

where $\mathbf{u}^T = \overbrace{(1,1,\ldots,1)}^{n}$, and, for convenience, the problem is converted into a minimization problem where the sign of the objective is changed. The following proposition indicates that if the integrality constraints are relaxed, then the resulting mathematical program is not convex if no other assumptions are given on the matrix D.

Proposition 1. *The function* $f(\mathbf{x}) : \Gamma \to \mathbb{R}$:

$$f(\mathbf{x}) = -\frac{\frac{1}{2}\mathbf{x}^T D \mathbf{x}}{\mathbf{u}^T \mathbf{x}} \tag{7}$$

where $\Gamma = \{\mathbf{x} \in [0,1]^n : \mathbf{u}^T \mathbf{x} \geq 2\}$, *is convex if and only if* $D \preceq 0$.

Proof. Set $\mathbf{u}^T \mathbf{x} = y$. The above function can be written as:

$$f(\mathbf{x}, y) = -\frac{\frac{1}{2}\mathbf{x}^T D \mathbf{x}}{y}$$

over the domain $\Gamma' = \{\mathbf{x} \in [0,1]^n, y \in \mathbb{R} : y = \mathbf{u}^T \mathbf{x} \geq 2\}$

The Hessian $\nabla^2 f(\mathbf{x}, y)$ is equal to:

$$\nabla^2 f(\mathbf{x}, y) = -\frac{1}{y^3}\begin{bmatrix} Dy^2 & -D\mathbf{x}y \\ -(D\mathbf{x})^T y & \mathbf{x}^t D \mathbf{x} \end{bmatrix} = -\frac{1}{y^3} D \begin{bmatrix} y \\ -\mathbf{x} \end{bmatrix} \begin{bmatrix} y \\ -\mathbf{x} \end{bmatrix}^T$$

Function $f(\mathbf{x}, y)$ is defined on the convex set Γ' and is convex if and only if $\nabla^2 f(\mathbf{x}, y) \succcurlyeq 0$, thus if $D \preceq 0$. □

With real world application data such as those considered in [11], the condition $D \preceq 0$ typically does not hold, hence the problem given by the continuous relaxation of (4)–(6) is not convex in the general case. NLP solvers can clearly be applied to this formulation (here we used XPRESS-SLP by Fair-Isaac), as it is shown in Section 4, even though just local maxima can be guaranteed.

On the other hand, the following straightforward QIP model holds for *Max−Sum DP*:

Max−Sum DP **QIP Formulation 1.**

$$\max \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{i,j} x_i x_j \tag{8}$$

subject to:

$$\sum_{i=1}^{n} x_i = m \tag{9}$$

$$x_i \in \{0,1\} \quad \forall i \in N \tag{10}$$

Even if general convexity results do not hold also for the continuous relaxation of this QIP, we note that for *Max−Sum DP*, the formulation corresponds to a 0/1 quadratic knapsack problem (with equality constraint) that has been much more tackled in the literature (see, e.g. [12]) and can be efficiently tackled by means of QIP solvers, such as, for instance, CPLEX. In [11], computational experiments dealt only with ILP formulations of *Max−Mean DP* and *Max−Sum DP* showing that the iterative solution of *Max−Sum DP* was superior to the one-shot solution of *Max−Mean DP*. In our preliminary tests on *Max−Sum DP*, we determined that CPLEX 12.5 was more efficient when applied to QIP Formulation 1 rather than to its standard linearization indicated in [11]. In fact, the latter required a computational time which is higher by more than an order of magnitude on instances with $n = 35$. The purpose of this work is to embed the repeated solution of the QIP formulation of *Max−Sum DP* into a heuristic framework for *Max−Mean DP*.

3. A hybrid heuristic approach

In [11], it was shown that in their real world small instances the value of m associated with the optimal solution lies in an interval, such that there are high quality solutions with values of m inside this interval. Although these promising intervals may be disjoint in large instances, the basic conclusion is very