Decremental clustering for the exact solution of some large-scale p-dispersion problems

Claudio Contardo

Department of management and technology, ESG UQAM, GERAD and CIRRELT
e-mail address: claudio.contardo@gerad.ca

March 21, 2019

Abstract

Given n points, a symmetric dissimilarity matrix D of dimensions $n \times n$ and an integer $p \geq 2$, the p-dispersion problem (pDP) consists in selecting exactly p out of the n points in such a way that the minimum dissimilarity between any pair of selected points is maximum. This problem is $\mathcal{NP}$-hard when p is an input of the problem. We propose a decremental clustering method to reduce the problem to the solution of a series of smaller pDPs until reaching proven optimality. The proposed method can handle problems orders of magnitude larger than the limits of the state-of-the-art solver for the pDP for small values of p.

1 Introduction

In the p-dispersion problem (pDP) we are given a set of n points, a symmetric dissmilarity matrix $D = \{D(i,j) : 1 \leq i,j \leq n\}$ and an integer $p \geq 2$. The objective is to select p points from the set of n so as to maximize the minimum pairwise dissimilarity within the selected points. The pDP, as noticed by Erkut (1990), is $\mathcal{NP}$-hard when p makes part of the input parameters (otherwise it can be solved in $O(n^p)$ time by exhaustive enumeration). We denote this problem, for given input parameters D and p (n is implicitly given in the dimensions of D), as pDP(D, p).

This problem arises in a number of practical contexts. In location analysis, a pDP becomes useful to decide the placement of installations whose proximity may be hazardous —this is the case of power plants, oil storage tanks or ammunition—, or in the location of retail stores to prevent cannibalization (Kuby 1987). In multiobjective optimization, in the presence of multiple solutions for a given optimization problem, one may solve a pDP to select a subset of those solutions as complementary as possible with respect to the values for each of the objectives (Saboonchi et al. 2014). The same authors also mention applications in portfolio optimization to spread the investment risk when choosing multiple products.

The state-of-the-art solver for the pDP (Sayah and Irnich 2017) relies on the solution of an integer program containing $O(n^2)$ variables and constraints. The model remains tractable for medium-sized problems but memory/time limits may prevent the solution of problems containing more than a few hundred nodes. The problem size and the large amount of symmetries impact the model's performance. Our article contributes at narrowing this gap by allowing the solution of potentially much larger problems (in terms of the number of nodes n), under the assumption that parameter p remains low (typically ≤ 10 when going large-scale). To this end, we introduce a decremental clustering scheme that in a dynamic fashion forms clusters of points and constructs instances of the pDP that are smaller in size and with much better numerical properties (most notably a much smaller amount of symmetries). These smaller instances are shown to provide upper bounds of the original problem and are much more tractable than the original pDP. The proposed iterative mechanism can scale and solve problems containing up to 100,000 nodes to proven optimality within reasonable time limits, this is orders of magnitude larger than the limits of previous methods.

The remainder of this article is organized as follows. In Section 2 we present a review of the relevant literature for the pDP. In Section 3 we present the decremental clustering framework. In Section 4 we present the results of our computational campaign to assess the effectiveness of our method. Finally, Section 5 concludes the paper.

2 Literature review

Applications of the pDP can be found in multiple fields including location analysis, multiobjective optimization of portfolio optimization. Kuby (1987) mention the importance of locating facilites are far as possible from

each other when they represent a potential hazard for the surrounding communities. The same authors also mention applications in store location. If two stores of the same chain are located too close, cannibalism may prevent them from selling at full potential. Saboonchi et al. (2014) also discuss an interesting application of the pDP in multiobjective optimization. If the Pareto frontier of a problem contains multiple solutions, one shall solve a pDP to find p such solutions with distinct features. The same authors also describe an application in portfolio optimization to —given a set of potential investment opportunities— choose a subset that reduces the closeness in terms of features between the different investment options so as to reduce the risk associated with the portfolio.

The pDP is tightly related to facility location problems (FLP, Laporte et al. (2015)). In its simplest version, a FLP corresponds to the problem of, given a set of potential facility locations and a set of customers, select a subset of potential facility locations and allocate the customers to those facilities, at minimum total cost. Facility location problems and applications have been widely studied in the scientific literature, and several comprehensive surveys have been recently published that take into account several of the latest advances in the field (Laporte et al. 2015, Melo et al. 2009). The pDP differs from a typical FLP model in the importance of the notion of customer. While they are of key importance for the right choice of the facilities in the FLP, in the pDP they are irrelevant. Only the facility locations are of importance, and their choice must reflect the objective function to be optimized: to maximize the minimum distance between any two chosen facilities.

The pDP is also related to clustering problems, and more specifically to the maximin split clustering problem (MMSCP). In the MMSCP, we are given a set N of observations, a dissmilarity matrix D and a target number of clusters p. One has to group the observations into p groups such that the minimum dissimilarity between any two observations belonging to different groups is maximized. The MMSCP, unlike the pDP, is polynomially solvable (Delattre and Hansen 1980).

Regarding the methodological contributions to the solution of the pDP, a handful of articles have dealt with the problem of solving the pDP to proven optimality. Pisinger (2006) introduces a quadratic formulation for the pDP which is then partially solved by a series of relaxations including semidefinite programming, and linearization-reformulation. The bounds are embedded within a branch-and-bound framework and the author reports the solution of problems containing a few hundred nodes. Kuby (1987) introduces a mixed-integer linear formulation of the problem with a series of Big-M coefficients. The model can be seen as a linearization of that of Pisinger (2006) even though it was introduced almost 20 years earlier. The model is more compact than that of Pisinger (2006) but provides much weaker upper bounds. Sayah and Irnich (2017) introduces a novel pure binary compact formulation of the problem that the authors solve by branch-and-cut. Clique-like inequalities are used to strengthen the model. Problems with up to 1,000 nodes are solved to proven optimality as reported by the authors. The same authors also mention that linear and binary search methods may be used with the different formulations to speed up the solution process. Such techniques have already been studied by Chandrasekaran and Daughety (1981), Pisinger (2006) for the pDP. For this to be beneficial, the models need to exploit the availability of lower and upper bounds to fathom non-promising branches of the implicit enumeration tree.

The decremental clustering method introduced in this article is tightly related to other decremental relaxation mechanisms recently introduced in the literature for the solution of other MiniMax (or equivalently MaxiMin) combinatorial optimization problems to proven optimality. In the vertex p-center problem (VPCP), for the same input parameters n, D and p, one has to select p points and to allocate the remaining points to its closest center in such a way that the maximum dissimilarity between a node and its assigned center is minimized. Chen and Chen (2009) and Contardo et al. (2019) propose decremental relaxation mechanisms to ignore some node allocation constraints, which are only added as needed. The relaxed problem can thus be modeled as a smaller VPCP. Contardo et al. (2019) report the solution of problems containing up to 1M observations. The minimax diameter clustering problem (MMDCP) is another problem for which the decremental relaxation mechanism has proven useful. Aloise and Contardo (2018) introduced a sampling mechanism to solve smaller MMDCPs in a dynamic fashion, allowing the solution to proven optimality of problems containing up to 600k observations.

Using clustering mechanisms for finding feasible solutions for hard comninatorial optimizatino problems is not something totally new in the operations research community. Embedding a clustering scheme within a heuristic solver has been common practice for many years and for multiple classes of problems. In vehicle routing and scheduling, the so-called cluster-first-route-second (Solomon 1987, Bräysy and Gendreau 2005) and route-first-cluster-second (Beasley 1983, Prins et al. 2014) paradigms are both based on combining routing and clustering techniques so as to reduce the computational burden associated with the routing or scheduling substructures. Our technique differs from those mentioned in this paragraph in the fundamental property that our mechanism is capable of providing solutions with proven optimality.

3 Decremental clustering

In this section we describe the decremental clustering method for the pDP. This section is subdivided in 5 subsections. In the first subsection, we provide the theoretical foundations and a high-level description of the method. The next four sections describe the different procedures of the method.

3.1 High-level description and theoretical foundations

Let us introduce some notation and vocabulary first. A *partition* of the n nodes, and denoted by $\mathcal{C}$, is a family $\{C_i : i = 1 \dots m\}$ such that (i) $C_i \cap C_j = \emptyset$ for every $1 \leq i < j \leq m$ and (ii) $\bigcup\{C_i : i = 1 \dots m\} = \{1 \dots n\}$. A partition $\mathcal{C}$ is said to be *sufficiently refined* if, for every set $C_i \in \mathcal{C}$, $D(C_i) := \max\{D(u, v) : u, v \in C_i, u < v\} < z^*$, where z^* is the optimal value of problem $\texttt{pDP}(\texttt{D}, \texttt{p})$. For practical purposes, it is sufficient to test the refinement of a partition with respect to a lower bound $l \leq z^*$. The correctness of the decremental clustering method is supported on the following result.

Lemma 1. *Let $\mathcal{C}$ be a sufficiently refined partition of the nodes of size m. Let $D^{\mathcal{C}}$ be a $m \times m$ dissimilarity matrix where $D^{\mathcal{C}}(i, j) = \max\{D(u, v) : u \in C_i, v \in C_j\}$. The optimal value ζ^* of the problem $\texttt{pDP}(\texttt{D}^{\mathcal{C}}, \texttt{p})$ provides an upper bound for problem $\texttt{pDP}(\texttt{D}, \texttt{p})$.*

Proof. Let $S = \{s_1 \dots s_p\}$ be an optimal solution of problem $\texttt{pDP}(\texttt{D}, \texttt{p})$, of value z^*. Because the partition $\mathcal{C}$ is sufficiently refined, it follows that no two nodes in S can be found in the same cluster $C \in \mathcal{C}$. For every $s \in S$, let $k(s)$ denote the cluster index in $\mathcal{C}$ where node s lies. By construction of $D^{\mathcal{C}}$, we have that $D(s, t) \leq D^{\mathcal{C}}(k(s), k(t))$ for every two nodes $s, t \in S, s < t$ and therefore $z^* \leq \zeta^*$. □

Our method works as follows. First, a lower bound $l \leq z^*$ is computed using a simple heuristic. Using this value, we build an initial sufficiently refined partition containing, say, m clusters. The resulting pDP is then solved yielding an upper bound u. If at least one of the clusters in the optimal solution to this problem is composed of two or more nodes, then one such cluster must be split into two, and the algorithm repeats using $m + 1$ clusters. Because the clusters can only reduce their size, it follows that the resulting partition is also sufficiently refined. The pseudo-code provided in Algorithm 1 formalizes the main steps of our algorithm.

Algorithm 1 Decremental clustering for $\texttt{pDP}(\texttt{D}, \texttt{p})$

Require: D, p
Ensure: Set $X = \{x_1 \dots x_p\}$ of optimal locations
  $L \leftarrow \texttt{heuristicPDP}(\texttt{D}, \texttt{p})$, $U \leftarrow \max\{D(i, j) : 1 \leq i < j \leq n\}$
  $\mathcal{C}, D^{\mathcal{C}} \leftarrow \texttt{initialClustering}(\texttt{D}, \texttt{p}, \texttt{L})$
  $S \leftarrow \emptyset, X \leftarrow \emptyset$
  repeat
    $\mathcal{C}, D^{\mathcal{C}} \leftarrow \texttt{splitAndAdd}(\texttt{S}, \texttt{X}, \mathcal{C}, \texttt{D}^{\mathcal{C}})$
    $U, X \leftarrow \texttt{solvePDP}(\texttt{D}^{\mathcal{C}}, \texttt{p})$ ▷ $U \leftarrow$ optimal value, $X \leftarrow$ optimal solution
    $S \leftarrow \{x \in X : D^{\mathcal{C}}(x, x) > 0\}$
  until $S = \emptyset$
  return X

The following proposition formalizes the exactness of the decremental clustering procedure.

Proposition 1. *The decremental clustering method ends in at most n iterations and produces an optimal solution to problem $\texttt{pDP}(\texttt{D}, \texttt{p})$.*

Proof. Let $X = \{x_1 \dots x_p\}$ be the optimal solution of problem $\texttt{pDP}(\texttt{D}^{\mathcal{C}}, \texttt{p})$. If the clusters corresponding to the solution X are all singletons, then this is also a feasible solution to problem $\texttt{pDP}(\texttt{D}, \texttt{p})$ and therefore produces a lower bound that matches with the upper bound provided by problem $\texttt{pDP}(\texttt{D}^{\mathcal{C}}, \texttt{p})$. Otherwise, the method identifies at least one cluster that contains two or more nodes and splits it into two separate groups. This can be done at most n times when the clusters in $\mathcal{C}$ become all singletons. □

3.2 Procedure $\texttt{heuristicPDP}(\texttt{D}, \texttt{p})$

In this section we describe a simple procedure to compute a non-trivial lower bound L of problem $\texttt{pDP}(\texttt{D}, \texttt{p})$. This procedure is far from producing a near-optimal solution to the problem, but is sufficient to feed the procedure $\texttt{initialClustering}(\texttt{D}, \texttt{p}, \texttt{L})$ to be described later in Section 3.3. We execute a k-means algorithm using the dissimlarity matrix D to construct p clusters. For each of the p centers in the cluster, we find the

node in each cluster that is closest to its center. Let us call this set of points $X = \{x_1 \dots x_p\}$. We compute $d \leftarrow \min\{D(x_i, x_j) : 1 \leq i < j \leq p\}$. This procedure is performed not once but multiple times for as long as the value d keeps increasing. Indeed, we stop after 10 iterations without being able to improve this value. The highest possible such value d is returned as lower bound L.

3.3 Procedure `initialClustering`$(\mathtt{D}, \mathtt{p}, \mathtt{L})$

In this section we describe a two-step procedure used to build an initial sufficiently refined clustering of the n points, using the lower bound L as stopping point. In the first step, a p-clustering of the nodes is found using a k-means algorithm. This clustering may not be sufficiently refined and thus the second step is executed. This second step is iterative and goes as follows. At any given iteration —say when the number of clusters has reached a value of m—, we check if the sizes of each cluster are strictly lower than L. If yes, the current clustering $\mathcal{C}$ and dissimlarity matrix $D^{\mathcal{C}}$ are returned. Otherwise, we compute $i^* \leftarrow \arg\max\{D^{\mathcal{C}}(i,i) : i = 1 \dots m\}$ and execute a k-means algorithm to further divide cluster C_{i^*} into two clusters. The dissimilarity matrix $D^{\mathcal{C}}$ is then extended to dimensions $(m+1) \times (m+1)$. At this point, only the new rows and columns need to be recomputed to alleviate the computational effort.

3.4 Procedure `splitAndAdd`$(\mathtt{S}, \mathtt{X}, \mathcal{C}, \mathtt{D}^{\mathcal{C}})$

In this section we describe a procedure that, given a clustering $\mathcal{C}$, a dissmilarity matrix $D^{\mathcal{C}}$, a family S of cluster indices with $D^{\mathcal{C}}(i,i) > 0$ for every $i \in S$ and a set of optimal cluster locations (with $S \subseteq X$), selects one cluster from those indexed in S and splits it into two separate clusters. The extended clustering and dissimilarity matrix are returned. We first compute $(s^*, x^*) \leftarrow \arg\min\{D^{\mathcal{C}}(s,x), s \in S, x \in X\}$, which is the pair of indices in $S \times X$ with minimum dissimilarity. This computation excludes on purpose the pairs with both indices in $X \setminus S$ as both associated nodes are —by construction of set S— of zero dissimilarity. If $x^* \in S$, then for the following, the index with highest value of $D^{\mathcal{C}}(u,u)$ is kept, with $u \in \{s^*, x^*\}$. For the retained index, we execute a k-means algorithm similar to the one described in the previous section, to split the associated cluster into two separate clusters. We update and return the clustering $\mathcal{C}$ and the dissimilarity matrix $D^{\mathcal{C}}$ accordingly.

3.5 Procedure `solvePDP`$(\mathtt{D}^{\mathcal{C}}, \mathtt{p})$

In this section we introduce a heuristic and an exact solver for problem `pDP`$(\mathtt{D}^{\mathcal{C}}, \mathtt{p})$. Without loss of generality and to alleviate the reading, we will drop the superindex $\mathcal{C}$ from the dissimilarity matrix. Therefore, we will simply denote D to refer to it. It goes without saying that we always execute the heuristic solver before any attempt at executing the exact one.

3.5.1 Exact solver

Our exact solver uses the pure-integer formulation introduced by Sayah and Irnich (2017) and solves it by branch-and-cut embedded within a double binary search method. This formulation uses m binary variables —one per row/column of the matrix D— to represent the location decisions, and Δ binary variables z, where Δ is the number of different values appearing in the matrix D. We refer to Sayah and Irnich (2017) for details of the model and the associated valid inequalities.

Within the decremental clustering scheme, we exploit the existence of a monotonically decreasing upper bound U and exploit this further within a double binary search scheme, as follows. Let us denote by `exactPDP`$(\mathtt{D}, \mathtt{p}, \mathtt{L}, \mathtt{U})$ the solver of problem `pDP`$(\mathtt{D}, \mathtt{p})$ when feeded with the additional lower and upper bounds L and U. These bounds can be exploited in two aspects. First, to reduce the number of binary variables z. Second, to derive cutting planes to strengthen the model. The details of these two accelerating features can be found in full extent in Sayah and Irnich (2017). Our double binary search method starts with making $l = u \leftarrow U$. It iterates by executing `exactPDP`$(\mathtt{D}, \mathtt{p}, \mathtt{l}, \mathtt{u})$ at every iteration. If no feasible solution exists, the quantities are updated according to the formulas $l \leftarrow l - 2^t, u \leftarrow l - 1$, where t is the iteration number. When the problem becomes feasible, we abort the optimization as soon as one feasible solution is identified and its objective value is used to update the lower bound. At this point, the final quantities l, u are used to feed another binary search method with the aim of closing the gap between l and u. For as long as $u > l$, we make $r \leftarrow \lceil (l+u)/2 \rceil$ and execute `exactPDP`$(\mathtt{D}, \mathtt{p}, \mathtt{r}, \mathtt{u})$. If the problem is feasible, we make $l \leftarrow r$, otherwise we make $u \leftarrow r - 1$ and repeat.

3.5.2 Heuristic solver

We have observed that, in a large number of iterations, the optimal value of problem `pDP`$(\mathtt{D}, \mathtt{p})$ does not decrease from one iteration to the next. This type of degeneracy is often observed in decremental relaxation schemes

(Aloise and Contardo 2018, Contardo et al. 2019). Therefore, before resorting to executing the exact solver described in the previous section, our heuristic scheme checks if it is possible to select p points out of the $p+1$ points identified from the previous iteration —which includes $p-1$ optimal clusters that remain untouched, plus the one that has been split into two— as described in Section 3.4. If the value of this solution equals the upper bound U from the last iteration, the associated solution is then optimal and there is no need to execute the exact solver.

4 Computational experience

In this section we provide computational evidence of the effectiveness of the proposed method. Our method has been coded in Julia v1.1 using the JuMP interface v18.5 with Gurobi v8.0 as multipurpose optimization solver. It runs on an Intel Xeon E5-2637 v2 @ 3.50 GHz with 128 GB of RAM. Although this machine is capable of executing code in parallel, for reproducibility purposes we limit the number of threads to one. We consider instances from the TSPLIB containing between 1,621 and 104,815 points in the euclidean plane. The dissimilarity between two points is computed according to the TSPLIB standard and considers only integral distances.

For each instance in the dataset, we consider four values of p, namely $p \in \{5, 10, 15, 20\}$. In addition to the algorithm described in this paper, we have also implemented a variation of Sayah and Irnich (2017)'s algorithm embedded within the same binary search method described in Section 3.5. Using the notation described in our paper, this method resorts to executing procedure `solvePDP(D, p)` at once. We have executed both algorithms and given them a maximum CPU time of 86,400 seconds (1 day). Our implementation of Sayah and Irnich's method could not handle problems containing 3,000 nodes or more (it rapidly ran out of memory), so the comparison between both methods is restricted to the smaller ones.

In Tables **??**-**??** we report a comparison between our method and our implementation of Sayah and Irnich's method, restricted to problems containing strictly less than 3,000 nodes. We report, for each method, the final upper bounds (under column labeled `UB`) and the elapsed CPU times in seconds (under column labeled `CPU`). We mark with an (*) the instances that could be solved to proven optimality. As the results show, our method is more robust and is capable of solving to proven optimality all the problems in this restricted testbed, something that our implementation of Sayah and Irnich's method did not. For the problems solved by both methods, ours is always substantially faster.

In Table **??** we report the results obtained by our method for the problems containing 3,000 or more nodes. We report, for each value of p, the final lower bound (under column labeled `LB`), the final upper bound (under column labeled `UB`), the CPU time in seconds (under column labeled `CPU`), and the final number of clusters at the final iteration (under column labeled `C`). Once again, we mark with an (*) the problems that could be solved to proven optimality.

5 Concluding remarks

We have introduced a decremental clustering method for the solution of the p-dispersion problem (`pDP`). Our method works by iteratively clustering nodes that are close to each other and solving a restricted `pDP` to compute a non-increasing upper bound. In practice, for small values of p, we are capable of proving optimality within a few iterations. The method is capable of handling and solving `pDP`s containing up to 100,000 nodes within a day. This is orders of magnitude larger than the limits of previous methods. As an avenue of further research, we believe that the algorithm could be adapted to solve variants of the `pDP` or other problems with a potential to benefit from clustering techniques. While clustering techniques have been widely studied in the scientific literature for the design of powerful heuristics, their use within exact methods is new and their full potential is yet to be understood in more depth.

Acknowledgments

The author thanks the Natural Sciences and Engineering Research Council of Canada (NSERC) under Discovery Grant 435824-2013.

References

D. Aloise and C. Contardo. A sampling-based exact algorithm for the solution of the minimax diameter clustering problem. *Journal of Global Optimization*, pages 1–18, 2018.

J. Beasley. Route first-luster second methods for vehicle routing. *Omega*, 11(4):403–408, 1983.

O. Bräysy and M. Gendreau. Vehicle routing problem with time windows, part i: Route construction and local search algorithms. *Transportation Science*, 39(1):104–118, 2005.

R. Chandrasekaran and A. Daughety. Location on tree networks: P-centre and n-dispersion problems. *Mathematics of Operations Research*, 6(1):50–57, 1981.

D. Chen and R. Chen. New relaxation-based algorithms for the optimal solution of the continuous and discrete p-center problems. *Computers & Operations Research*, 36(5):1646–1655, 2009.

C. Contardo, M. Iori, and R. Kramer. A scalable exact algorithm for the vertex p-center problem. *Computers & Operations Research*, 103:211–220, 2019.

M. Delattre and P. Hansen. Bicriterion cluster analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(4):277–291, 1980.

E. Erkut. The discrete p-dispersion problem. *European Journal of Operational Research*, 46(1):48–60, 1990.

M. J. Kuby. Programming models for facility dispersion: The p-dispersion and maxisum dispersion problems. *Geographical Analysis*, 19(4):315–329, 1987.

G. Laporte, S. Nickel, and F. S. da Gama. *Location science*, volume 528. Springer, 2015.

M. T. Melo, S. Nickel, and F. Saldanha-Da-Gama. Facility location and supply chain management–a review. *European Journal of Operational Research*, 196(2):401–412, 2009.

D. Pisinger. Upper bounds and exact algorithms for p-dispersion problems. *Computers & Operations Research*, 33(5): 1380–1398, 2006.

C. Prins, P. Lacomme, and C. Prodhon. Order-first split-second methods for vehicle routing problems: A review. *Transportation Research Part C: Emerging Technologies*, 40:179–200, 2014.

B. Saboonchi, P. Hansen, and S. Perron. Maxminmin p-dispersion problem: A variable neighborhood search approach. *Computers & Operations Research*, 52:251–259, 2014.

D. Sayah and S. Irnich. A new compact formulation for the discrete p-dispersion problem. *European Journal of Operational Research*, 256(1):62–67, 2017.

M. M. Solomon. Algorithms for the vehicle routing and scheduling problems with time window constraints. *Operations Research*, 35(2):254–265, 1987.