

Two Optimization Models for Clustering Problems

Nelson Maculan

Universidade Federal do Rio de Janeiro
Ilha do Fundão, Bloco H-223, Rio de Janeiro, Brasil
maculan@cos.ufrj.br

Marcos Negreiros

Universidade Estadual do Ceará
Av Silas Munguba, 1740, Campus do Itaperi, Fortaleza, Brasil
marcos.negreiros@uece.br

Renan Vicente Pinto

Universidade Federal Rural do Rio de Janeiro
Seropédica, Brasil
renanvp@ufrrj.br

ABSTRACT

This work presents two new formulations, one for the Unitary Weighted Weber-Problem (UWWP) and one for the Minimum Sum of Squares Clustering (MSSC). The UWWP is a clustering problem where the location of each cluster center minimizes the total Euclidean distance to assigned items. In MSSC problem, the squares of the Euclidean distances from cluster centers to assigned items is minimized. We apply new combinatorial approaches and obtain a convex quadratic model with second order cone constraints that can be efficiently solved by interior point algorithms.

KEYWORDS. Weber Problem, Clustering, Unconstrained Location.

Topics: Combinatorial Optimization, Mathematical Programming.

1. Introduction

The clustering problem is a well known partition problem where a number of items is to be divided into a number of hierarchical or non-hierarchical partitions that minimizes the total dissimilarity or maximize the total similarity of clusterings [6], [8], [7].

Dissimilarity and similarity functions can be designed to generate different clustering patterns to achieve specific purpose. In some sense these metrics can result a comprehensive duality between them. The measures of dissimilarity are mostly used because of their relation to human expectations on clustering by proximity [7]. In this paper, we study the measures of total Euclidean distance between items and their cluster centers and the total variance between clusters and their cluster centers both representing total dissimilarity measure for a non-hierarchical process.

In the literature, there were a wide number of papers about the min-sum-of-squares clustering (MSSC) problem. The same as the minimum total variance problem. For this problem, it is known that the cluster center is located in the average mean of the positions of the assigned items. This property facilitates the common use of heuristic methods based on k-Means, Exponential-Smooth, Bi-Level programming, Mathematical Programming based methods like Repeated B&B, Integer Programming, Branch-and-Cut SDP (semidefinite programming) and math-heuristics, [7], [2], [16], [17], [3], [1], [11], [12], [5].

In the context of the Euclidean distance as dissimilarity measure for the Weber Problem (WP), it was proposed in [9] a more general version of the problem including weights in the objective function for the distance between located clusters. The author applied his model in the context of communication network. In [14], it was proposed a linear model to solve WP but using Manhattan metrics as the dissimilarity function and [15] proposed a hyperbolic approach also to solve WP using Manhattan metrics. Much later, [4] considers the version where there are positive and negative weights between items and cluster centers. The authors propose a global optimization problem with both the objective function and constraint functions written as differences of convex functions. The authors use a math-heuristic to deal with the so called WAR Problem (Weber Problem with Attraction and Repulsion) and solved instances up to 1000 items. Although there are many works considering the WAR problem, as far as we know the case where the weights are positive unitary were not investigated.

To have a perception of the differences between the problems here stud-

ied, an example of an instance with 8 items and 2 clusters is shown in Figures 1 and 2, where the different solutions obtained for both Unitary Weighted Weber Problem (UWWP) and MSSC Problem are depicted. It can be noted the proximity of the locations between centers for this particular instance. But in fact there are still more differences in the assignment of items to clusters centers not revealed in this small example.

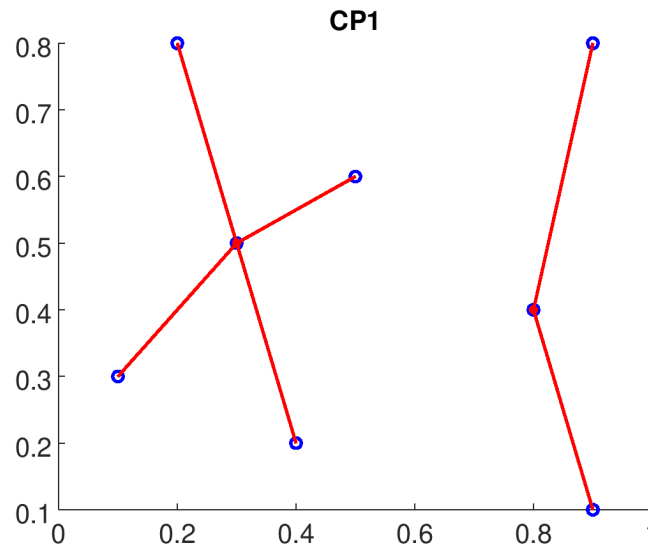


Figure 1: Instance result for the Unitary Weighted Weber Problem.

This paper is organized as follows. In Section 2 we develop and propose a new convex formulation for the UWWP. In Section 3 we develop and formulate a new convex and quadratic formulation to MSSC. We conclude the paper indicating some future work to reproduce the results of these new formulations for instances of the literature using regular QP or second-order conic programming (SOCP) state-of-the-art solvers.

2. Model for Unitary Weighed Weber Problem (UWWP)

Given a set of p points $a^i \in \mathbb{R}^n$, for $i \in P = \{1, 2, \dots, p\}$, we define the variables $x^j \in \mathbb{R}^n$, for $j \in K = \{1, 2, \dots, k\}$, to represent the location of the center of cluster C_j . We consider $p, k \in \mathbb{N}$ and $p \gg k$. Let the distance of a point

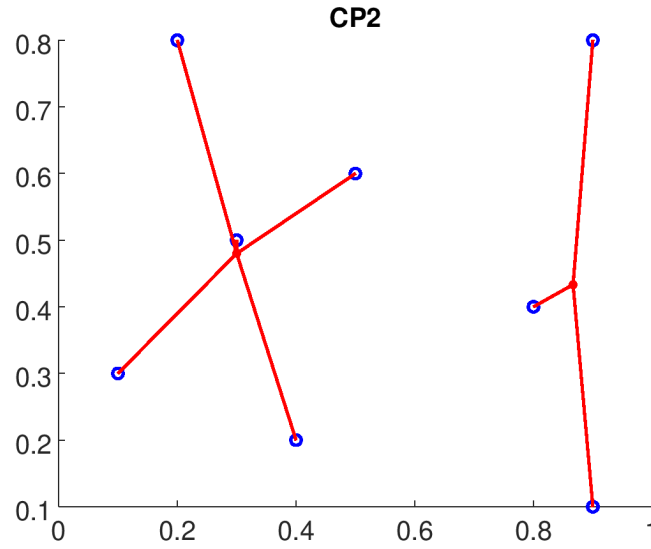


Figure 2: Instance result for the MSSC Problem.

a^i to the center of the cluster C_j be

$$d_{ij} = \|a^i - x^j\|_2 = \sqrt{\sum_{l=1}^n (a_l^i - x_l^j)^2}, \quad \forall i \in P, \forall j \in K.$$

The problem consists of finding the coordinates of x^j and which points a^i belong to each cluster in order to minimize the sum of intra-cluster distances. So we define the binary variables $y_{ij} \in \{0, 1\}$, such that $y_{ij} = 1$ if $a^i \in C_j$, and $y_{ij} = 0$, otherwise. The problem can be written as

$$\min \sum_{j \in K} \sum_{i \in P} d_{ij} y_{ij} \quad (1)$$

Each cluster must contain at least one point. We write

$$\sum_{i \in P} y_{ij} \geq 1, \quad \forall j \in K.$$

And, as each point must belong to one unique cluster, we write

$$\sum_{j \in K} y_{ij} = 1, \quad \forall i \in P.$$

As $y_{ij} \in \{0, 1\}$, we have $y_{ij}^2 = y_{ij}$, $\forall i \in P, \forall j \in K$. We can rewrite (1) as

$$\begin{aligned} \min \sum_{j \in K} \sum_{i \in P} d_{ij} y_{ij} &= \sum_{j \in K} \sum_{i \in P} \sqrt{y_{ij} \sum_{l=1}^n (a_l^i - x_l^j)^2} \\ &= \sum_{j \in K} \sum_{i \in P} \sqrt{\sum_{l=1}^n y_{ij}^2 (a_l^i - x_l^j)^2} \\ &= \sum_{j \in K} \sum_{i \in P} \sqrt{\sum_{l=1}^n m_{lij}^2} \end{aligned} \quad (2)$$

where we have introduced the variables $m_{lij} = y_{ij}(a_l^i - x_l^j)$, for all $i \in P$, for all $j \in K$, and for all $l = 1, \dots, n$. They will be well defined if we use the constraints

$$\begin{aligned} -(1 - y_{ij})M + (a_l^i - x_l^j) &\leq m_{lij} \leq (a_l^i - x_l^j) + (1 - y_{ij})M, \\ -My_{ij} &\leq m_{lij} \leq My_{ij}, \end{aligned}$$

where we used the constant value

$$M = \max_{1 \leq s < q \leq p} \|a^s - a^q\|_2.$$

For all $i \in P$ and for all $j \in K$, we define the variables

$$z_{ij} = \sqrt{\sum_{l=1}^n m_{lij}^2} \geq 0.$$

and substitute into the objective function (2). As the problem will minimize z_{ij} , we can use the constraints

$$z_{ij}^2 \geq \sum_{l=1}^n m_{lij}^2, \quad i \in P, j \in K.$$

Our model can be written as follows:

$$(CP1) : \text{Minimize } \sum_{j=1}^k \sum_{i=1}^p z_{ij}, \text{ subject to :} \quad (3)$$

$$z_{ij}^2 \geq \sum_{l=1}^n m_{lij}^2, \quad i \in P, j \in K, \quad (4)$$

$$-(1 - y_{ij})M + (a_l^i - x_l^j) \leq m_{lij} \leq (a_l^i - x_l^j) + (1 - y_{ij})M, \quad (5)$$

$$-My_{ij} \leq m_{lij} \leq My_{ij}, \quad (6)$$

$$z_{ij} \geq 0, \quad i \in P, j \in K, \quad (7)$$

$$\sum_{j=1}^k y_{ij} = 1, \quad i \in P, \quad (8)$$

$$\sum_{i=1}^p y_{ij} \geq 1, \quad j \in K. \quad (9)$$

$$y_{ij} \in \{0, 1\}, \quad i \in P, j \in K. \quad (10)$$

$$x^j \in R^n, \quad j \in K \quad (11)$$

where (5) and (6) are defined for all $i \in P$, for all $j \in K$ and for all $l = 1, \dots, n$. Constraints (4) and (7) define second order cone constraints [18].

If x_*^j , for a given $j \in K$, is an optimal solution of (CP1), then x_*^j will be a Weber center for C_j [13]. The continuous relaxation of (CP1) is a smooth and convex optimization problem. This model can be solved by XPRESS software that recognizes constraints (4) and (7) as second order cones and uses an interior point algorithm proposed in [18].

One important consideration is that if we preprocess the data, redefining each given point as $a^i := \frac{1}{M} a^i$, for all $i \in P$, we will have

$$|a_l^i - a_l^j| \leq 1, \quad \forall i, j \in P, i \neq j,$$

and we can use $M = 1$ in constraints (5) and (6).

3. Model for the MSSC Problem

The only difference between the problems UWWP and MSSC is the objective function. In MSSC, instead of (1), we use

$$\min \sum_{j \in K} \sum_{i \in P} d_{ij}^2 y_{ij}, \quad (12)$$

where the distances are squared. Using the variables m_{lij} as defined in Section 2, we can rewrite (12) as

$$\min \sum_{j \in K} \sum_{i \in P} \sum_{l=1}^n m_{lij}^2$$

Our model can be written as follows:

$$(CP2) : \text{Minimize } \sum_{j=1}^k \sum_{i=1}^p \sum_{l=1}^n m_{lij}^2, \quad (13)$$

subject to: (5), (6), (8)-(11).

If x_*^j , for a given $j \in K$, is an optimal solution of (CP2), then x_*^j will be the geometrical center of C_j .

4. Conclusion

In this paper we presented two new formulations, one for UWWP and one for the MSSC problem. For UWWP, we developed and formalized a quadratic integer problem with second order cone constraints, called (CP1), that can be solved by actual state-of-the-art quadratic optimization solvers like XPRESS-FICO.

Also, we developed and formalized, for the MSSC problem, a quadratic problem with linear constraints that can be solved by XPRESS-FICO, Gurobi, CPLEX, Mosek, Knitro and LINGO, all available on the platform NEOS Argonne Server [10].

As they are, although better than the previous ones used for the WP and MSSC, the formulations can only hold small instances up to 20 items.

The formulations can be processed much better by using triangle inequality cuts as proposed by [4] and [11], and also forbidden items to cluster assignment elimination cuts. Other approaches may be evaluated to hold large instances with small/average number of clusters by using the results of [17] to incorporate the solution process.

References

- [1] P. Aloise, D.; Hansen. A banch-and-cut sdp-based algorithm for minimum sum-of-squares clustering. *Pesquisa Operacional*, 29:503–516, 2009.
- [2] A.M. Bagirov. Modified global k-means algorithm for minimum sum-of-squares clustering problems. *Pattern Recognition*, 41:3192–3199, 2009.
- [3] M.J. Brusco. An enhanced branch-and-bound algorithm for a partitioning problem. *British Journal of Mathematical and Statistical Psychology*, 56:83–92, 2003.
- [4] P. C. Chen, P. Hansen, B. Jaumard, and H. Tuy. Weber’s problem with attraction and repulsion. *Journal of Regional Science*, 32:467–486, 1992.
- [5] O. Du Merle, P. Hansen, B. Jaumard, and N. Mladenovic. An interior point algorithm for minimum sum-of-squares clustering. 21:1485–1505, 2000.
- [6] P. Hansen and B. Jaumard. Cluster analysis and mathematical programming. *Mathematical Programming*, page 191–215, 1997.
- [7] A.K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31:651–666, 2010.
- [8] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM Press*, 31:264–323, 1999.
- [9] W. Miehle. Link length minimization in networks. *Operations Research*, 6, 1958.
- [10] Neos. Neos (network-enabled optimization system). <http://www.neos-server.org/neos/solvers/index.html>, 2020.
- [11] J. Peng and Y. Xia. A cutting algorithm for the minimum sum-of-squared error clustering. Proceedings of the SIAM International Data Mining Conference, 2005.
- [12] J. Peng and Y. Xia. Approximating k-means-type clustering via semidefinite programming. 18:186–205, 2007.
- [13] A. Weber. *Über den Standort der Industrien - English translation: The Theory of the Location of Industries*. Chicago, Chicago University Press, 1909.

- [14] G.O. Wesolowsky and R. F. Love. The optimal location of new facilities using rectangular distances. *Operations Research*, 1971.
- [15] G.O. Wesolowsky and R. F. Love. A non-linear method for solving a generalized rectangular distance weber problem. *Management Science*, 1972.
- [16] A.E. Xavier. The hyperbolic smoothing clustering method. *Pattern Recognition*, 43:731–737, 2010.
- [17] A.E. Xavier and V.L. Xavier. Solving the min-sum-of-squares clustering problem by hyperbolic smoothing and partition into boundary and gravitational regions. *Pattern Recognition*, 44:70–77, 2011.
- [18] G. Xue and Y. Ye. An efficient algorithm for minimizing a sum of euclidean norms with applications. *SIAM Journal on Optimization*, 4(7):1017–1036, 1997.