May 2nd, 11:00 AM - 1:00 PM

Clustering and Multifacility Location with Constraints via Distance Function Penalty Methods and DC Programming

Tuyen Tran
Portland State University

Samuel Reynolds
Portland State University

Thai An Nguyen
Duy Tan University

Mau Nam Nguyen
Portland State University

CLUSTERING AND MULTIFACILITY LOCATION WITH CONSTRAINTS VIA DISTANCE FUNCTION PENALTY METHODS AND DC PROGRAMMING

Tuyen Tran, Samuel Reynolds, Thai An Nguyen[a]

Faculty Advisor: Dr. Mau Nam Nguyen

Fariborz Maseeh Department of Mathematics and Statistics

PO Box 751, Portland, Oregon 97207

[a]Institute of Research and Development, Duy Tan University, Vietnam (thaian2784@gmail.com).

Research of this author was supported by the Vietnam National Foundation for Science and Technology Development under grant #101.01-2017.325.

Abstract

This is a continuation of our effort in using mathematical optimization involving DC programming in clustering and multifacility location. We study a penalty method based on distance functions and apply it particularly to a number of problems in clustering and multifacility location in which the centers to be found must lie in some given set constraints. We also provide numerical examples to test our method.

In the current time of "big data", clustering is a very important problem that helps classify data in many fields such as machine learning, pattern recognition, image analysis, data compression, and computer graphics. Given a finite number of data points with a measurement distance, a centroid-based clustering problem seeks a finite number of cluster centers with each data point assigned to the nearest cluster center in a way that a certain measurement distance is minimized.

The main focus here is on solving a number of clustering and multifacility location problems with constraints. We use a penalty method with squared Euclidean distance functions to convert constrained problems to unconstrained problems. Then appropriate DC decompositions and the DCA are used to minimize the penalized objective functions. In the case where the measurement distance is defined by the Euclidean norm instead of the squared Euclidean norm, we use Nesterov's smoothing techniques for *reducing the nonsmoothness* of the model and for providing a DC decomposition that is favorable for applying the DCA. Our method opens up the possibility of using distance function penalty methods for other problems of DC programming.

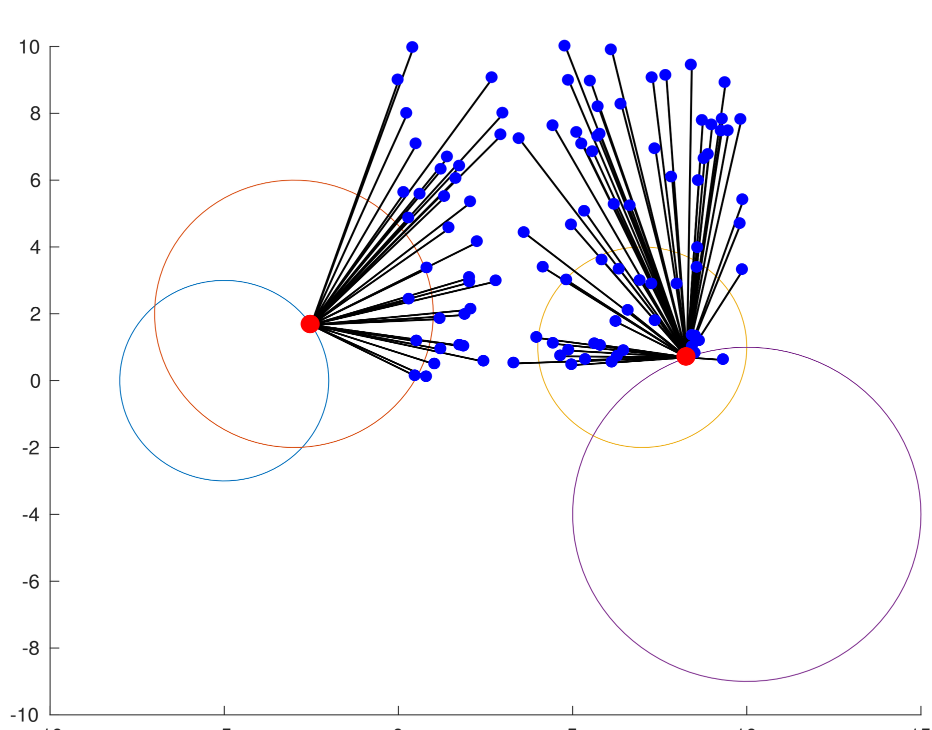

Figure: A clustered data set with $m = 100$ random nodes and $k = 2$

Clustering with Constraints

We study problems of *clustering with constraints* in which the measurement distance is defined by the squared Euclidean norm. We seek k centers $x^1, \ldots, x^k \in \mathbb{R}^d$ of m data nodes $a^1, \ldots, a^m \in \mathbb{R}^d$ and impose the restriction that each $x^\ell \in \cap_{i=1}^q \Omega_i^\ell$ for some nonempty closed convex set $\Omega_i^\ell \subset \mathbb{R}^d$ with $\ell = 1, \ldots, k$ and $i = 1, \ldots, q$. The problem we are concerned with is given by

$$\begin{aligned} &\min && \psi(x^1, \ldots, x^k) = \Sigma_{i=1}^m \min_{\ell=1,\ldots,k} \|x^\ell - a^i\|^2 \\ &\text{subject to} && x^\ell \in \cap_{j=1}^q \Omega_j^\ell \text{ for } \ell = 1, \ldots, k. \end{aligned}$$

This problem can be converted to an unconstrained minimization problem:

$$\begin{aligned} \min\ & f(x^1, \ldots, x^k) = \frac{1}{2} \Sigma_{i=1}^m \min_{\ell=1,\ldots,k} \|x^\ell - a^i\|^2 + \frac{\tau}{2} \Sigma_{\ell=1}^k \Sigma_{i=1}^q [d(x^\ell; \Omega_i^\ell)]^2, \\ & x^1, \ldots, x^k \in \mathbb{R}^d, \end{aligned}$$

where $\tau > 0$ is a penalty parameter.

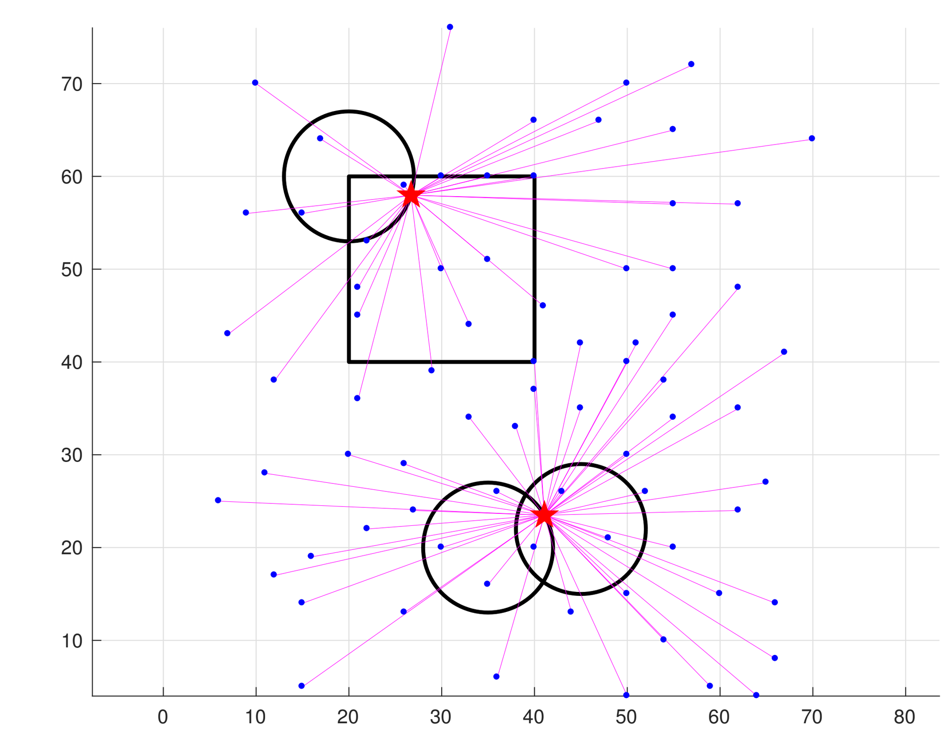

Figure: A 2-center constrained clustering problem for dataset EIL76 taken from the Traveling Salesman Problem Library [3].

Set Clustering with Constraints

In this section, we turn our attention to a model of *set clustering* with constraints, i.e., for given m subsets $\Lambda_1, \ldots, \Lambda_m \subset \mathbb{R}^d$, we seek k cluster centers $x^\ell \in \cap_{j=1}^q \Omega_j^\ell$ for $\ell = 1, \ldots, k$, where each Ω_j^ℓ is a subset of $\mathbb{R}^d$. The measurement distance is defined by the squared distance functions to the sets involved. The optimization modeling of the problem to be solved is given by

$$\begin{aligned} &\min && \psi(x^1, \ldots, x^k) = \Sigma_{i=1}^m \min_{\ell=1,\ldots,k} [d(x^\ell; \Lambda_i)]^2 \\ &\text{subject to} && x^\ell \in \cap_{j=1}^q \Omega_j^\ell \text{ for } \ell = 1, \ldots, k. \end{aligned}$$

Using the penalty method based on distance functions with a parameter $\tau > 0$, we consider the constrained set clustering model:

$$\begin{aligned} \min\ & f(x^1, \ldots, x^k) = \tfrac{1}{2} \Sigma_{i=1}^m \min_{\ell=1,\ldots,k} [d(x^\ell; \Lambda_i)]^2 + \tfrac{\tau}{2} \Sigma_{\ell=1}^k \Sigma_{j=1}^q [d(x^\ell; \Omega_j^\ell)]^2, \\ & x^1, \ldots, x^k \in \mathbb{R}^d. \end{aligned}$$

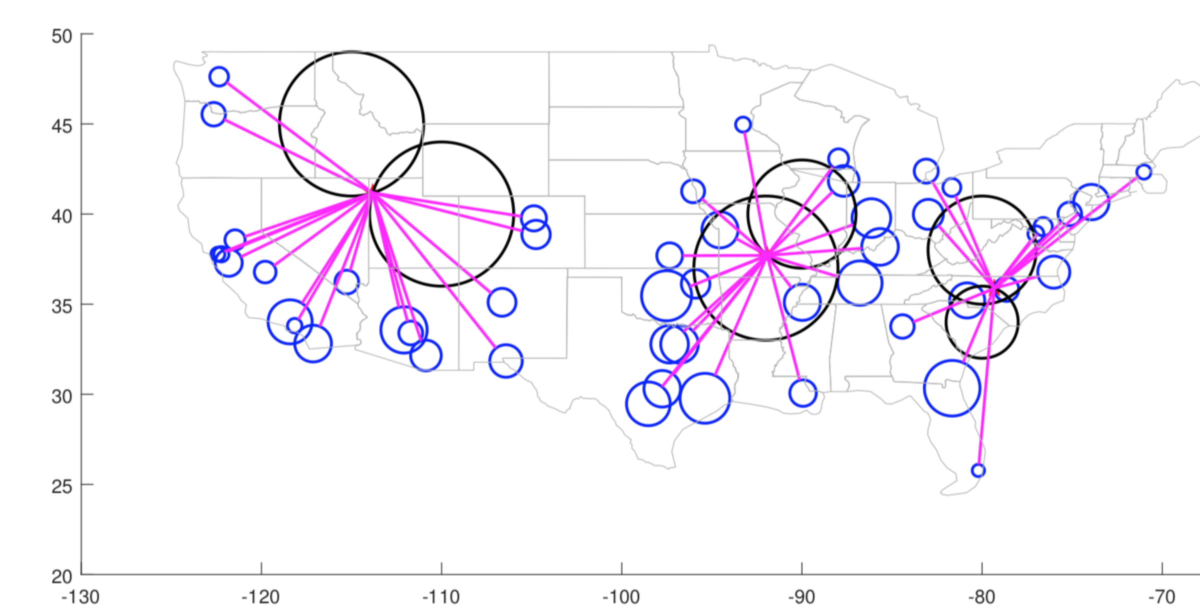

Figure: A 3-center set clustering problems with 50 most populous US cities. Each city is approximated by a ball proportional to its area.

Multifacility Location with Constraints

Given a set of m points (nodes) $a^1, a^2, \ldots, a^m$ in $\mathbb{R}^d$, our goal is find k centers x^ℓ for $\ell = 1, \ldots, k$, which must be in constraint sets $\cap_{i=1}^q \Omega_i^\ell$ for $l = 1, .., k$, such that the *transportation cost* to the nodes is minimized. The total cost is given by

$$\psi(\mathbf{X}) = \psi(x^1, \ldots, x^k) = \sum_{i=1}^m \min_{\ell=1,\ldots,k} \|x^\ell - a^i\|.$$

We apply Nesterov's smoothing techniques from [4] with smoothing parameter μ and penalty parameter τ to approximate the objective function f_τ by a new DC function which is favorable for applying the DCA.

$$\begin{aligned} f_{\tau,\mu}(x^1, \ldots, x^k) = & \Big(\frac{\mu}{2} \sum_{i=1}^m \sum_{\ell=1}^k \left\| \frac{x^\ell - a^i}{\mu} \right\|^2 + \frac{\tau q}{2} \sum_{\ell=1}^k \|x^\ell\|^2\Big) \\ & - \Big(\frac{\mu}{2} \sum_{i=1}^m \sum_{\ell=1}^k \left[d\left(\frac{x^\ell - a^i}{\mu}; \mathbb{B} \right) \right]^2 + \sum_{i=1}^m \max_{r=1,\ldots,k} \sum_{\ell=1,\ell\neq r}^k \|x^\ell - a^i\| + \frac{\tau}{2} \sum_{\ell=1}^k \sum_{i=1}^q \varphi_{\Omega_i^\ell}(x^\ell)\Big). \end{aligned}$$

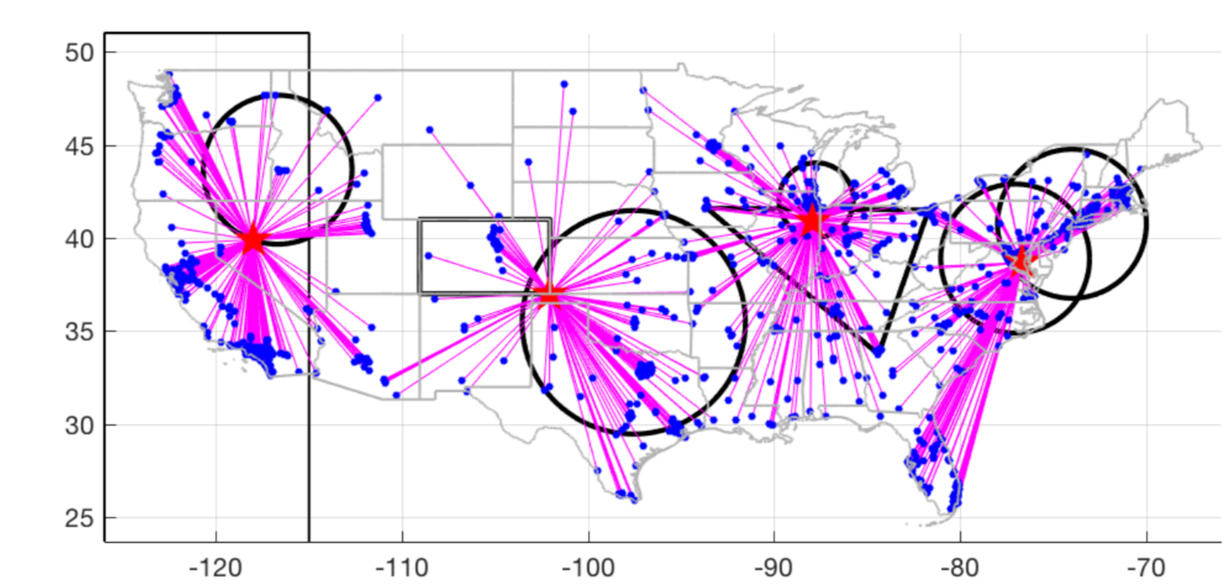

Figure: A 4-center constrained multifacility location problem with US cities dataset.

References

[1] L. T. H. An, M. T. Belghiti, P. D. Tao: A new efficient algorithm based on DC programming and DCA for clustering. J. Glob. Optim., **27**, 503–608 (2007)

[2] B. S. Mordukhovich and N. M. Nam, *An Easy Path to Convex Analysis and Applications*, Morgan & Claypool Publishers, San Rafael, CA, 2014.

[3] G. Reinelt, TSPLIB: A Traveling Salesman Problem Library, *ORSA Journal of Computing*. **3**, 376-384, 1991.

[4] N. M. Nam, R.B. Rector, D. Giles: Minimizing Differences of Convex Functions with Applications to Facility Location and Clustering, submitted.

[5] Y. Nesterov: Smooth minimization of non-smooth functions. Math. Program. **103**, 127–152 (2005)

[6] N. M. Nam, W. Geremew, S. Reynolds, T. Tran, The Nesterov Smoothing Technique and Minimizing Differences of Convex Functions for Hierarchical Clustering, submitted.