Les Cahiers du GERAD ISSN: 0711–2440

Sum-of-Squares Clustering on Networks

E. Carrizosa, N. Mladenović, R. Todosijević

G–2012–10

March 2012

Sum-of-Squares Clustering on Networks

Emilio Carrizosa

Faculdad de Matemáticas
Universidad de Sevilla
Sevilla, Spain
ecarrizosa@us.es

Nenad Mladenović

GERAD & School of Mathematics
Brunel University-West London
London, United Kingdom
nenad.mladenovic@brunel.ac.uk

Raca Todosijević

Faculty of Mathematics
University of Belgrade
Belgrade, Serbia
racatodosijevic@gmail.com

March 2012

Les Cahiers du GERAD
G–2012–10

Abstract

Finding p prototypes by minimizing the sum of the squared distances from a set of points to its closest prototype is a well-studied problem in clustering, data analysis and continuous location. In this note, this very same problem is addressed assuming, for the first time, that the space of possible prototype locations is a network. We develop some interesting properties of such clustering problem. We also show that optimal cluster prototypes are not necessary located at vertices of the network.

Key Words: Networks, clustering, location, p-median.

Acknowledgments: This research is partially supported by the bilateral Serbian-Spanish project AIB2010SE-00318, and projects MTM2009-14039 (Ministry of Science and Innovation, Spain), FQM-329 (Junta de Andalucia, Spain) and EU European Regional Development Funds. Last two authors are also partially supported by Project #172010, financed by Serbian Ministry of Sciences.

1 Introduction

Let $N(V, E)$ be a connected and undirected network with a node set V and an edge set E . Each edge is represented by its endpoints and its length. Let x be a point on an edge, then its location is determined by its distance from a prescribed endpoint of that edge. If an edge has endpoints (u, v) and length l, then any real number $x \in [0, l]$ denotes the location in that edge for which the length of sub-edge $[u, x]$ is x . For any two points x and y in N , let $d(x, y)$ denote the length of a shortest path connecting x and y. For any nonempty finite set P of points, let $d(x, P)$ denote the minimum distance from x to P , i.e

$$d(x, P) = \min\{d(x, p) : p \in P\} \tag{1}$$

Consider a set P of p locations for prototypes on a network N and suppose that the cardinality of the set V is equal to n. Let each vertex $x_i \in V$ has nonnegative weight w_i . The objective for our minimum sum-of-squares clustering on the network (MSSCN) is to minimize the following sum for all choices of p prototype locations:

$$\min_P f(P) = \sum_{i=1}^{n} w_i (d(x_i, P))^2 \tag{2}$$

The problem of finding continuous medians on the network is considered in [2]. It is shown that the continuous p-median is always located at vertices V , whenever the Euclidean distance is used, i.e., when the objective function is not squared Euclidean distance. See also [2, 4, 5] for related continuous network location problems, and [3] for complexity results. Network location problems on tree are considered in [7].

In this paper we show that if the square distances between any two nodes of the network are used, then the set of p points are not necessarily located at nodes. This result is illustrated on a simple example.

The paper is organized as follows. In the next section, we give theoretical properties of MSSCN problem. In Section 3, we give conclusions and possible directions for the future work.

2 Structural properties

We assume that the number p of prototypes to be located is strictly smaller than the number of nodes with positive weight. Otherwise, the problem is solved in a straightforward manner, since locating prototypes at all nodes with positive weight would yield an objective value of zero, which is optimal.

We also assume that prototypes can be located at nodes so as at points in the interior of edges. If the prototype locations were restricted to be nodes, then the problem can be formulated as a p-median problem: considering the set of nodes as both the set of consumers and the set of candidate sites for the facilities, and the distance between them as the squared shortest-path distance.

Allowing prototypes to be located in the interior of edges, as done in this paper, yields to a related yet different problem. It is intuitively obvious that the optimal locations for the prototypes are likely to be different if not only nodes but also interior points; so interior points are also allowed to be prototypes. However, it is unclear if both problems yield to different clusters because the clustering of nodes, in the two models, to their closest prototype can be the same or may be different. Some simple experiments were performed to show that the optimal allocations may be different if prototypes are chosen from the set of nodes or from the interior of the edges. We tested this on the following simple example: A rectilinear network, identified with the interval $(0, 1)$, with 10 nodes randomly distributed in $(0, 1)$ is built. Two prototypes $(p = 2)$ are to be chosen, either prototypes are only nodes (case 1) or any point in (0,1) (case 2). Since the dimension of the problem is small, the optimal solutions in both cases are found by complete enumeration. This example, as well as, the optimal allocations of both cases are given in Figures 1 and 2 below. Optimal locations for prototypes are denoted by +, and the points belonging to the same cluster are colored in the same color.

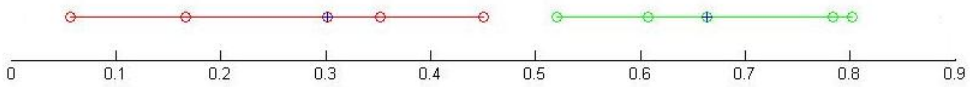

Figure 1: Optimal solution for case 1: one prototype is at node 3, and the other at node 8.

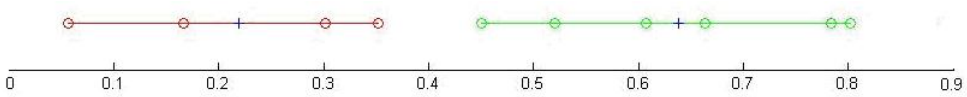

Figure 2: Optimal solution for case 2: one prototype is between nodes 2 and 3 the other between nodes 7 and 8.

We intended to find how frequent is the case depicted in Figures 1 and 2. To do this, we generated the experiment 1000 times, with 1000 different sets of 10 nodes randomly chosen in (0,1). Out of these 1000 instances, 160 instances yielded different clusters for cases 1 and 2. This shows that, although in many cases the clusters are the same, different clusters configurations can be obtained by allowing prototypes to be interior points in the edges.

Our experience shows, that the MSSCN problem (2) with prototypes on edges may yield (and does yield) the clusters different from those when the MSSCN is restricted to nodes. Some basic properties with evident algorithmic consequences of this new problem are now stated.

Property 2.1 *If solution P of problem (2) is optimal, then the shortest path from a point to its closest prototype cannot pass through the other prototypes.*

Proof. It is obvious since any optimal solution minimizes the sum of increasing functions of the distances (point, prototype). □

Property 2.2 *The interior of a given edge contains at most one optimal prototype of Problem (2).*

Proof. Suppose that the interior of a given edge e contains two optimal prototypes, denoted v_1 and v_2. By Proposition 2.1, the ending nodes of e are not prototypes. Each customer allocated to prototype v_1 must pass through the same end of the edge which contains those two prototypes. If we place prototype v_1 on such end node of e, we will obtain a better solution than the current solution, which cannot be optimal. Hence, we conclude that the interior of e must contain at most one optimal prototype. □

Property 2.3 *If an optimal prototype is located at a node, then the interior of all adjacent edges contain no optimal prototypes.*

Proof. Assume that the opposite holds, i.e., that an optimal prototype is located at a node v, and that the interior of an adjacent edge e with endpoints v and v^* contains an optimal prototype x. By Proposition 2.1, for each node allocated to prototype at x, its shortest path to x must pass through v^*, the other endpoint of e. Obviously, if we move prototype at x to v^*, we will obtain a solution better than the optimal one, which is a contradiction. □

An important case is the single-prototype problem, $p = 1$ .The idea, similar to the procedures described, e.g. in [4, 5], is that within each edge the objective function is piecewise convex and smooth; hence, once the different pieces are identified, one only needs to find the local (and thus global) minimum for each piece and store the best solution found. Moreover, since the objective function is piecewise quadratic, the critical points are obtained by solving (constrained) quadratic problems in one variable, thus the exact solution can be computed. This case, though simple, is critical if location-allocation algorithms are to be used: at each stage, one has the node set split into p clusters, and for each such cluster X , one needs to find the optimal

prototype by solving the 1-facility case considering X as a set of nodes. The procedure for the single-facility case and set X is as follows:

For each edge with both end nodes from X , suppose that the optimal solution of MSSCN problem belongs to this edge. Then we find that optimal solution and denote it with q.

We choose the best point (i.e., the point for which the value of the MSSCN is the smallest) of all points q found in step 1. Obviously, that point is the optimal solution of MSSCN problem in that cluster.

Suppose now that a cluster contains points $x_1, x_2, \ldots, x_n$. Assume further that the edge e belongs to that cluster and that we are looking for the solution of MSSCN, for $p = 1$, problem on such edge. Denote end nodes of edge e by u and $v(e = (u, v))$. For every point x_i we can determine the point y_i on edge e, if such exists, whose distance from point x_i to point y_i via node u is the same as the distance from point x_i to point y_i via node v . If we suppose that the distance from u to y_i is equal to z and that the length of edge e is equal to l, then we can obtain a formula for calculating the position of y_i:

$$z = (d(x_i, v) + l - d(x_i, u))/2 \tag{3}$$

The value of z can be calculated for any point x_i . Such value determines whether the shortest path from x_i to any point that belongs to edge e passes through node u or v :

- If the value of z is less than or equal to 0, then the length of the path from x_i to any point on edge e via point v is shorter than the length of the path from x_i to the same point via point u .
- If the value of z is bigger than or equal to l, then the length of the path from x_i to any point on the edge e via point u is shorter than the length of the path from x_i to the same point via v .
- If $0 < z < l$ then, if the distance between point t on edge e and point u is less than or equal to z , then the shortest path from x_i to t passes through point u , otherwise it passes through point v .

Points y_i split the edge e into at most n-1 parts. For each of such parts, we know whether a shortest path from each point in the cluster to any point on such a part passes through point u or point v because we know the value of z for each point in the cluster. So, we can easily detect the minimum of a quadratic function, which is our objective function for MSSCN problem, when the feasible set of locations for the prototypes is restricted to such part.

The minimum of the function on such part is a vertex of a parabola if feasible, otherwise it is the end point closer to the vertex of the parabola. The best point among such minima will be the optimal solution on that edge.

3 Conclusions

In this paper we have considered a continuous p-median problem on a network, taking the minimization of the weighted sum of squared distances between any two nodes as a criterion. We show that if we extend feasible set of prototypes from set of nodes to set of all points of the network, we can obtain different optimal clusters. We illustrate this property on a simple example.

Future research includes the extension of this result to other types of location and clustering problems on networks. We are now working on designing local search heuristics and metaheuristics, based on Variable Neighborhood Search approach of [6].

References

[1] Francis, R.L., Lowe, T.J., Rayco, M.B., and Tamir, A., Aggregation error for location models: survey and analysis,*Annals of Operations Research*, (2009) 171–208.

[2] Hansen, P., and Labbé, M., The continuous p-Median of a network, *Networks*, **19** (1989) 595–606.

[3] Hassin, R., and Tamir, A., Improved complexity bounds for location problems on the real line, *Operations Reaserch Letters*, **10** (1991) 395–402.

[4] Lopez-de-los-Mozos, M., and Mesa, J., The maximum absolute deviation measure in location problems on networks, *European Journal of Operational Reserach*, **135** (2001) 184–194.

[5] Lopez-de-los-Mozos, M., Mesa, J., and Puerto, J., A generalized model of equality measures in network location problems, *Computers and Operations research*, **35** (2008) 651–660.

[6] Mladenović, N., and Hansen, P., Variable neighborhood search, *Computers and Operations research*, **24** (1997) 1097–1100.

[7] Perez-Brito, D., Mladenović, N., and Moreno-Perez, J.A., A note on spanning trees for network location problems, *Yugoslav Journal of Operations Research*, **8** (1998) 129–135.