Communications of the Korean Statistical Society
2012, Vol. 19, No. 2, 213–217

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.2.213

Fuzzy k-Means Local Centers of the Social Networks

Wonseok Woo[a], Myung-Hoe Huh[1,a]

[a]Department of Statistics, Korea University

Abstract

Fuzzy k-means clustering is an attractive alternative to the ordinary k-means clustering in analyzing multivariate data. Fuzzy versions yield more natural output by allowing overlapped k groups. In this study, we modify a fuzzy k-means clustering algorithm to be used for undirected social networks, apply the algorithm to both real and simulated cases, and report the results.

Keywords: Fuzzy k-means clustering, social network analysis, local centers, communities.

1. Background and Aim of the Study

Fuzzy k-means clustering, which is developed by Dunn (1973) and improved by Bezdek (1981), can be stated as follows. For p-variate observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, determine k prototypes $\mathbf{c}_1, \ldots, \mathbf{c}_k$ in the same space by minimizing

$$\sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^{\phi} \| \mathbf{x}_i - \mathbf{c}_j \|^2$$

with respect to $\mathbf{c}_1, \ldots, \mathbf{c}_k$ and m_{ij}'s such that

$$m_{ij} \geq 0, \quad \text{for all } i, j; \qquad \sum_{j=1}^{k} m_{ij} = 1, \quad \text{for each } i = 1, \ldots, n.$$

There were no studies for the best choice of $\phi\,(> 1)$, which is normally set to 2. Larger ϕ results in more overlapped clusters.

We consider the social network with n nodes linked by undirected edges. Nodes of the network are denoted by serial numbers $1, \ldots, n$. The geodesic distances between node i and i' are denoted by $d(i, i')$.

For the social networks, Huh (2011) proposed the k-means algorithm to identify k nodes of the highest local centrality. In this study, we develop a fuzzy k-means algorithm to locate k local centers of the network. Then we apply the algorithm to both real and simulated cases and report the comparative results with those of the k-means algorithm.

[1] Corresponding author: Professor, Department of Statistics, Korea University, Anam-Dong, Sungbuk-Gu, Seoul 136-701, Korea. E-mail: stat420@korea.ac.kr

2. Proposed Algorithm

We propose the following fuzzy algorithm for the social networks to locate local centers of the highest local centrality.

0) Apply the max-min procedure, as stated in Huh (2011), to initialize k central nodes $a_1, \ldots, a_k$.

1) For each node $i\,(= 1, \ldots, n)$, its memberships m_{ij} to the group $j\,(= 1, \ldots, k)$ are derived as follows: If $i = a_j$ for some $j\,(= 1, \ldots, k)$, m_{ij} is set to 1 and $m_{ij'}$ is set to 0 for $j' \neq j$. Otherwise,

$$m_{ij} = \frac{[1/d(i, a_j)]^{\frac{1}{(\phi-1)}}}{\sum_{j'=1}^{k} [1/d(i, a_{j'})]^{\frac{1}{(\phi-1)}}}.$$

2) For each $j\,(= 1, \ldots, k)$, the central node a_j is re-obtained by minimizing

$$\sum_{i=1}^{n} m_{ij}^{\phi} d(i, a)$$

over $a = 1, \ldots, n$.

3) Repeat Step 1 and Step 2 until the decrease of the objective function

$$\sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^{\phi} d(i, a_j)$$

becomes negligible. In this study, the fuzziness parameter ϕ is set to 2.

As a numerical example, consider Zachary's (1978) Karate club network that consists of 34 members, eventually split into two groups due to an inner conflict. When the fuzzy k-means algorithm with $k = 2$ is applied, Node 1 (the instructor) and Node 34 (the administrator) are correctly identified as key members, or the local centers, in the dispute. The fuzzy k-means algorithm leaves Nodes 9, 14, 20, 25, 26, and 32 as neutral in the dispute, indicated by the membership profile $(0.5, 0.5)$. The splits of all the other nodes with uneven membership profiles are correctly identified. See Figure 1. On the other hand, the k-means algorithm of Huh (2011) incorrectly classifies Nodes 9, 25, 26, and 32.

3. Numerical Studies

Faux Magnolia High Network

This case comes from the survey of friendship relations among 1,461 students of a high school in the southern United States (Goodreau *et al.*, 2008; Hunter *et al.*, 2008). The fuzzy k-means algorithm, applied to the largest component of the network with $k = 10$, identifies Nodes 17, 30, 94, 102, 119, 141, 182, 242, 315, 392 as local centers, while the k-means algorithm lists Nodes 7, 30, 102, 156, 169, 182, 200, 239, 240 and 392 as local centers. See Figure 2, where filled circles represent local center nodes.

The average distance between local centers was 12.6, smaller than the corresponding average distance 13.8 between local centers by the k-means algorithm of Huh (2011). Possibly, such phenomenon is general in the networks composed of several communities. We explore a class of Monte Carlo networks in the next subsection.

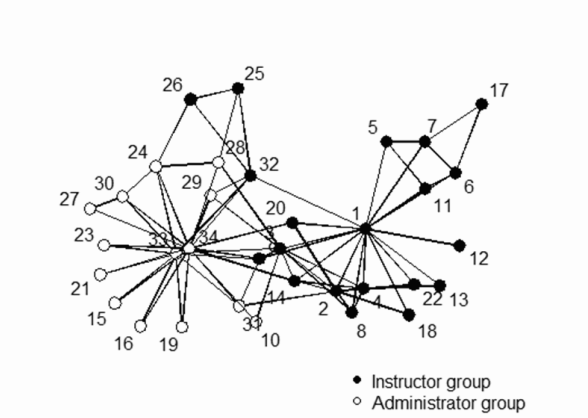

Figure 1: *Karate network*

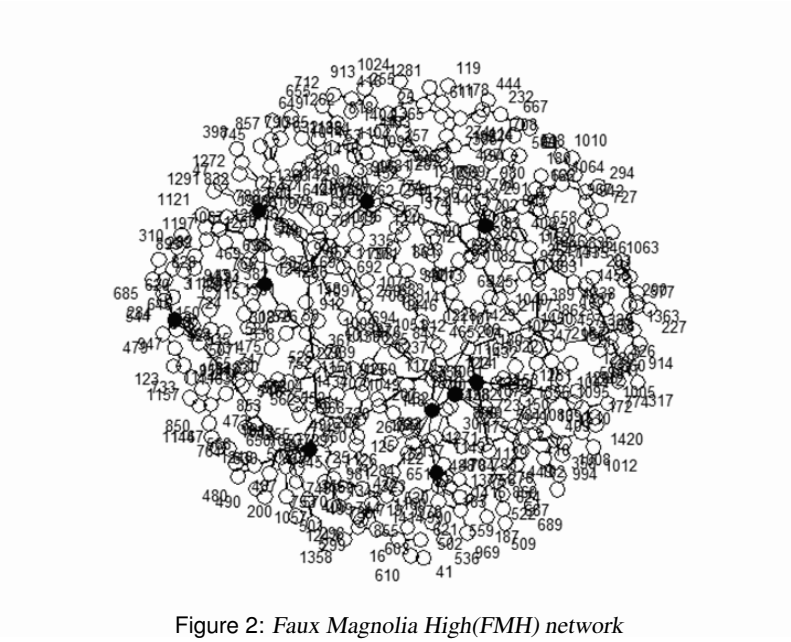

Figure 2: *Faux Magnolia High(FMH) network*

Monte Carlo Networks with Several Communities

Huh and Lee (2011) proposed a procedure for generating random networks with k communities as follows: First, generate k random communities of the same size completely split. Second, for each link of the network, one end node is switched to another with probability p. Thus, p is the parameter that influences the strength of connectedness. With 1,000 randomly generated networks with $n = 100$, $k = 5$, $p = 0.1, 0.2$ (Figure 3 and Figure 4) and $p = 0.4$, the following results are obtained.

For $p = 0.1$, the k-means algorithm produced the mean 4.17 and the s.d. 0.59 for the average distance between local centers, while the fuzzy k-means algorithm produced the mean 4.10 and the s.d. 0.60. For $p = 0.2$, the k-means algorithm produced the mean 3.58 and the s.d. 0.50, while the fuzzy k-means algorithm produced the mean 3.40 and the s.d. 0.51. For $p = 0.4$, the k-means algorithm produced the mean 3.05 and the s.d. 0.53, while the fuzzy k-means algorithm produced

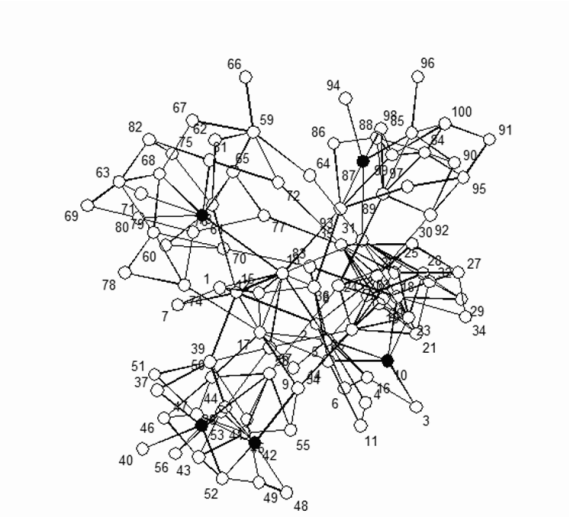

Figure 3: *A simulated network with five communities,* $p = 0.1$

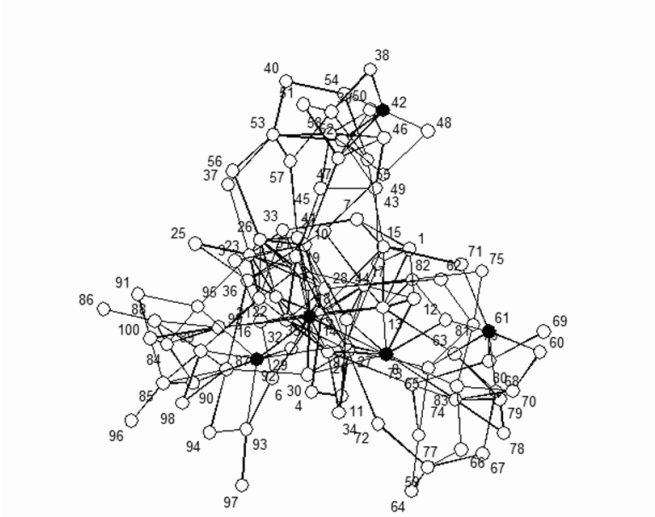

Figure 4: *A simulated network with five communities,* $p = 0.2$

the mean 2.65 and the s.d. 0.49. Thus, it is evident that the average distance between local centers by the fuzzy k-means algorithm is smaller than the corresponding average distance by the k-means algorithm; however, the difference is not large in the absolute sense.

4. Concluding Remarks

From real and simulated cases of the social networks, we observed that the difference between the fuzzy k-means algorithm and the k-means algorithm is more or less small. However, the fuzzy k-means algorithm allows overlapped subgrouping of the nodes in a natural way. Also, the fuzzy k-means algorithm produces the membership profiles for each node of the network, so that the analyst may identify the nodes located at the borderline of two or more communities.

References

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *Journal of Cybernetics*, **3**, 32–57.

Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T. and Morris, M. (2008). A `statnet` tutorial, *Journal of Statistical Software*, **24**, 1–26.

Huh, M. H. (2011). Local centers of the social network, *Communications of the Korean Statistical Society*, **18**, 213–217.

Huh, M. H. and Lee, Y. (2011). Random generation of the social network with several communities, *Communications of the Korean Statistical Society*, **18**, 595–601.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M. and Morris, M. (2008). `ergm`: A package to fit, simulate and diagnose exponential-family models for networks, *Journal of Statistical Software*, **24**, 1–29.

Zachary, W. W. (1978). An information flow model for conflict and fission in small groups, *Journal of Anthropological Research*, **33**, 452–473.

Received November 8, 2011; Revised November 28, 2011; Accepted January 5, 2012