Comparing Fuzzy Algorithms on Overlapping Communities in Networks

Jian Liu

LMAM and School of Mathematical Sciences, Peking University,
Beijing 100871, P.R. China
dugujian@pku.edu.cn

Abstract. Uncovering the overlapping community structure exhibited by real networks is a crucial step toward an understanding of complex systems that goes beyond the local organization of their constituents. Here three fuzzy c-means methods, based on optimal prediction, diffusion distance and dissimilarity index, respectively, are test on two artificial networks, including the widely known ad hoc networks and a recently introduced LFR benchmarks with heterogeneous distributions of degree and community size. All of them have an excellent performance, with the additional advantage of low computational complexity, which enables one to analyze large systems. Moreover, successful applications to real world networks confirm the capability of the methods.

Keywords: Overlapping community structure, Fuzzy c-means, Optimal prediction, Diffusion distance, Dissimilarity index.

1 Introduction

The modern science of networks has brought significant advances to our understanding of complex systems [1,2]. One of the most relevant features of networks representing real systems is community structure, i.e. the organization of nodes in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such communities can be considered as fairly independent compartments of a network, playing a similar role like the tissues or the organs in the human body. Detecting communities is of great importance in sociology, biology and computer science, disciplines where systems are often represented as networks [3,4,5,6,7,8,9,10,11,12,13,14].

In a previous paper [11], an approach to partition the networks based on optimal prediction theory is derived. The basic idea is to associate the network with the random walker Markovian dynamics [15], then introduce a metric on the space of Markov chains (stochastic matrices), and optimally reduce the chain under this metric. The final minimization problem is solved by an analogy to the traditional fuzzy c-means algorithm in clustering analysis [16]. Another work [7] is also along the lines of random walker Markovian dynamics, then introduce the diffusion distance on the space of nodes and identify

R. Zhu et al. (Eds.): ICICA 2010, LNCS 6377, pp. 269–276, 2010.

the geometric centroid in the same framework. This proximity reflects the connectivity of nodes in a diffusion process. Under the same framework [6], a dissimilarity index for each pair of nodes is proposed, which one can measure the extent of proximity between nodes of a network and signify to what extent two nodes would like to be in the same community. They can motivate us to solve the partitioning problem also by fuzzy c-means algorithms [16] under these two measures.

We will compare the above three algorithms in fuzzy c-means formulation based on optimal prediction, diffusion distance and dissimilarity distance, respectively. From the numerical performance to the artificial networks: the ad hoc network and the LFR benchmark, we can see that the three methods identify the community structure during with a high degree of accuracy, while they also produce little different. Moreover, application to a real word social network, the karate club network, confirms the differences among them.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the three type of fuzzy c-means algorithms and the corresponding framework. In Section 3, we apply the algorithms to the representative examples mentioned before. Finally we make the conclusion in Section 4.

2 The Framework of Fuzzy c-Means Algorithms for Network Partition

2.1 The Fuzzy c-Means Based on Optimal Prediction

We will start with the probabilistic framework for network partition [12]. Let $G(S, E)$ be a network with n nodes and m edges, where S is the set of nodes, $E = \{e(x, y)\}_{x,y\in S}$ is the weight matrix and $e(x, y)$ is the weight for the edge connecting the nodes x and y. We can relate this network to a discrete-time Markov chain with stochastic matrix p whose entries are given by $p(x, y) = \frac{e(x,y)}{d(x)}$, where $d(x) = \sum_{z\in S} e(x, z)$ is the degree of the node x. This Markov chain has stationary distribution $\mu(x) = \frac{d(x)}{\sum_{z\in S} d(z)}$ and it satisfies the detailed balance condition with respect to μ. The basic idea in [12] is to introduce a metric for p with the form $\|p\|_\mu^2 = \sum_{x,y\in S} \frac{\mu(x)}{\mu(y)}|p(x, y)|^2$ and find the reduced Markov chain $\tilde{p}$ by minimizing the distance $\|\tilde{p} - p\|_\mu$. For a given partition of S as $S = \cup_{k=1}^N S_k$ with $S_k \cap S_l = \emptyset$ if $k \neq l$, let $\hat{p}_{kl}$ be the coarse grained transition probability from S_k to S_l on the state space $\mathbb{S} = \{S_1, \ldots, S_N\}$ which satisfies $\hat{p}_{kl} \geq 0$ and $\sum_{l=1}^N \hat{p}_{kl} = 1$. Let $\rho_k(x)$ be the probability of the node x belonging to the k-th community which needs the assumption that $\rho_k(x) \geq 0$ and $\sum_{k=1}^N \rho_k(x) = 1$. Naturally the matrix $\tilde{p}$ can be lifted to the space of stochastic matrices on the original state space S via

$$\tilde{p}(x, y) = \sum_{k,l=1}^{N} \rho_k(x)\hat{p}_{kl}\mu_l(y), \qquad x, y \in S, \tag{1}$$

where $\mu_k(x) = \frac{\rho_k(x)\mu(x)}{\hat{\mu}_k}$ and $\hat{\mu}_k = \sum_{z\in S} \rho_k(z)\mu(z)$. Given the number of the communities N, we optimally reduce the random walker dynamics by considering the following minimization problem

$$\min_{\rho,\hat{p}} J_{\mathrm{OP}} = \|p - \tilde{p}\|_\mu^2 = \sum_{x,y\in S} \mu(x)\mu(y) \left| \sum_{m,n=1}^{N} \rho_m(x)\rho_n(y)\frac{\hat{p}_{mn}}{\hat{\mu}_n} - \frac{p(x,y)}{\mu(y)} \right|^2 \quad (2)$$

subject to the constraints described before. To minimize (2), we define $\hat{p}^*_{kl} = \sum_{x,y\in S} \mu_k(x)p(x,y)\rho_l(y) = \frac{1}{\hat{\mu}_k}\sum_{x,y\in S} \mu(x)\rho_k(x)p(x,y)\rho_l(y)$. Then the Euler-Lagrange equations of (2) are derived as

$$\left(I_{\hat{\mu}}^{-1} \cdot \hat{\hat{\mu}}\right) \cdot \hat{p} \cdot \left(I_{\hat{\mu}}^{-1} \cdot \hat{\hat{\mu}}\right) = \hat{p}^*, \quad (3a)$$

$$\rho = I_{\hat{\mu}}\hat{p}^{-1}\hat{\hat{\mu}}^{-1}\rho p^T, \quad (3b)$$

where $\hat{\hat{\mu}} = \rho \cdot I_\mu \cdot \rho^T$. The diagonal matrices I_μ, $I_{\hat{\mu}}$ choose $\mu(x)$ and $\hat{\mu}_k$ as their diagonal entries, respectively. To ensure the nonnegativity and normalization for $\hat{p}$ and ρ, we add a projection step after each iteration and change (3) to

$$\hat{p} = \mathcal{P}\left(\hat{\hat{\mu}}^{-1} \cdot I_{\hat{\mu}} \cdot \hat{p}^* \cdot \hat{\hat{\mu}}^{-1} \cdot I_{\hat{\mu}}\right), \quad (4a)$$

$$\rho = \mathcal{P}\left(I_{\hat{\mu}}\hat{p}^{-1}\hat{\hat{\mu}}^{-1}\rho p^T\right). \quad (4b)$$

Here $\mathcal{P}$ is a projection operator which maps a real vector into a vector with nonnegative, normalized components.

2.2 The Fuzzy *c*-Means Based on Diffusion Distance

The main idea of [7] is to define a system of coordinates with an explicit metric that reflects the connectivity of nodes in a given network and the construction is also based on a Markov random walk. The transition matrix p has a set of left and right eigenvectors $\{\psi_i\}_{i=0}^{n-1}$, $\{\varphi_i\}_{i=0}^{n-1}$ and a set of eigenvalues $1 = \lambda_0 \geq |\lambda_1| \geq \cdots \geq |\lambda_{n-1}| \geq 0$. Let $q(t)$ be the largest index i such that $|\lambda_i|^t > \delta|\lambda_1|^t$ and if we introduce the diffusion map

$$\Psi_t : x \longmapsto \begin{pmatrix} \lambda_1\varphi_1^t(x) \\ \vdots \\ \lambda_{q(t)}\varphi_{q(t)}^t(x) \end{pmatrix}, \quad (5)$$

then the diffusion distance $D_t(x,y)$ between x and y is defined as the weighted L^2 distance and can be approximated to relative precision δ using the first $q(t)$ non-trivial eigenvectors and eigenvalues

$$D_t^2(x,y) = \sum_{z\in S} \frac{(p(x,z) - p(y,z))^2}{\mu(z)} \simeq \sum_{i=1}^{q(t)} \lambda_i^{2t}\left(\varphi_i(x) - \varphi_i(y)\right)^2 = \|\Psi_t(x) - \Psi_t(y)\|^2,$$

where the weight $\mu(z)^{-1}$ penalize discrepancies on domains of low density more than those of high density. The centroid $c(S_k)$ of community S_k is defined as $c(S_k) = \sum_{x \in S_k} \frac{\mu(x)}{\hat{\mu}(S_k)} \Psi_t(x)$, where $\hat{\mu}(S_k) = \sum_{x \in S_k} \mu(x)$. Similar as before, given the number of clusters N, we optimally reduce the random walker dynamics by

$$\min_{\rho, c} J_{\mathrm{DD}} = \sum_{k=1}^{N} \sum_{x \in S} \rho_k^2(x) \mu(x) \|\Psi_t(x) - c(S_k)\|^2 \tag{6}$$

subject to $\sum_{k=1}^{N} \rho_k(x) = 1$. To minimize (6), we define $\hat{\mu}_k = \sum_{z \in S} \rho_k^b(z) \mu(z)$. The Euler-Lagrange equations are given by

$$c = I_{\hat{\mu}}^{-1} \rho^b I_\mu \Psi_t, \tag{7a}$$

$$\rho = W I_{\mathbf{1} \cdot W}^{-1}, \tag{7b}$$

where $\rho^b = (\rho_k^b(x))_{k=1,\ldots,N, x \in S}$ and W is with entries $W_k(x) = \frac{1}{\|\Psi_t(x) - c(S_k)\|^{\frac{2}{b-1}}}$.

2.3 The Fuzzy c-Means Based on Dissimilarity Index

In [6], a dissimilarity index between pairs of nodes is defined and can measure the extent of proximity between nodes in graphs. Suppose the random walker is located at node x. The mean first passage time $t(x, y)$ is the average number of steps it takes before it reaches node y for the first time, which is given by

$$t(x, y) = p(x, y) + \sum_{j=1}^{+\infty} (j + 1) \cdot \sum_{z_1, \cdots, z_j \neq y} p(x, z_1) p(z_1, z_2) \cdots p(z_j, y). \tag{8}$$

It has been shown that $t(x, y)$ is the solution of the linear equation in [6]. The difference in the perspectives of nodes x and y about the network can be quantitatively measured. The dissimilarity index is defined by the following expression

$$\Lambda(x, y) = \frac{1}{n - 2} \left(\sum_{z \in S, z \neq x, y} \Big(t(x, z) - t(y, z) \Big)^2 \right)^{\frac{1}{2}}. \tag{9}$$

Then fuzzy c-means is considered to address the optimization issue

$$\min_{\rho, m} J_{\mathrm{DI}} = \sum_{k=1}^{N} \sum_{x \in S} \rho_k^2(x) \Lambda^2(m(S_k), x), \tag{10}$$

which guarantees convergence towards a local minimum [16]. The Euler-Lagrange equation for (10) with constraints $\sum_{k=1}^{N} \rho_k(x) = 1$ is given by the following

$$\rho_k(x) = \frac{1/\Lambda^2(m(S_k), x)}{\sum_{l=1}^{N} 1/\Lambda^2(m(S_l), x)}, \quad x \in S, \quad k = 1, \cdots, N, \tag{11a}$$

$$m(S_k) = \arg\min_{x \in S_k} \frac{1}{|S_k|} \sum_{y \in S_k, y \neq x} \Lambda(x, y), \quad k = 1, \cdots, N, \tag{11b}$$

where $|S_k|$ is the number of nodes in S_k and we set $x \in S_k$ if $k = \arg\max_l \rho_l(x)$.

3 Experimental Results

3.1 Ad Hoc Network with 128 Nodes

We apply our methods to the ad hoc network with 128 nodes. The ad hoc network is a typical benchmark problem considered in many papers [4,6,11,12,13,14]. Suppose we choose $n = 128$ nodes, split into 4 communities containing 32 nodes each. Assume pairs of nodes belonging to the same communities are linked with probability p_{in}, and pairs belonging to different communities with probability p_{out}. These values are chosen so that the average node degree, d, is fixed at $d = 16$. In other words p_{in} and p_{out} are related as $31p_{\text{in}} + 96p_{\text{out}} = 16$. Here we naturally choose the nodes group $S_1 = \{1 : 32\}, S_2 = \{33 : 64\}, S_3 = \{65 : 96\}, S_4 = \{97 : 128\}$. Testing an algorithm on any graph with built-in community structure also implies defining a quantitative criterion to estimate the goodness of the answer given by the algorithm as compared to the real answer that is expected. We change z_{out} from 0.5 to 8 and look into the normalized mutual information [8,9,10] produced by the three methods. From Figure 1, we can see that OP fuzzy c-means performs better than the two others, especially for the more diffusive cases when z_{out} is large.

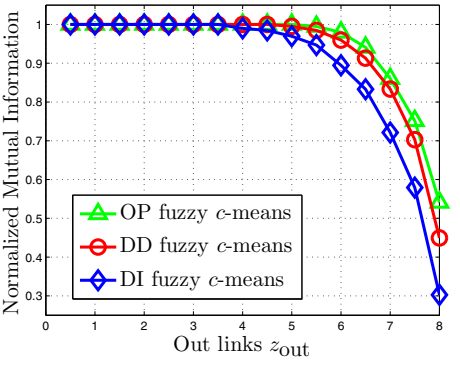


Fig. 1. Test of the three fuzzy c-means algorithms on the ad hoc network with 128 nodes with the normalized mutual information defined in [8]. Each point corresponds to an average over 20 graph realizations.

3.2 The LFR Benchmark

The LFR benchmark [9,10] is a special case of the planted partition model, in which groups are of different sizes and nodes have different degrees. The

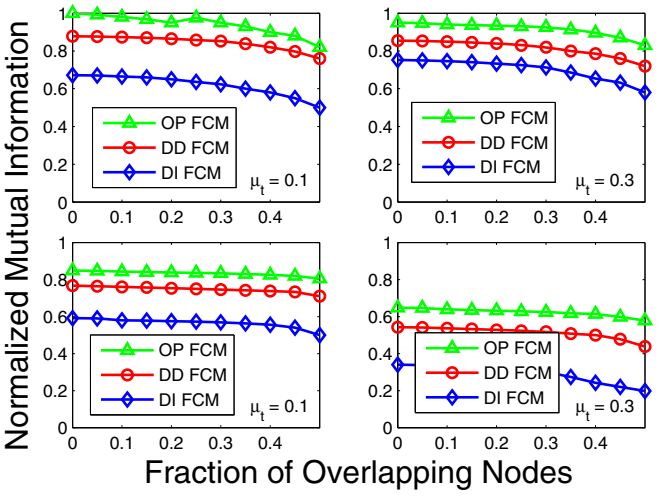
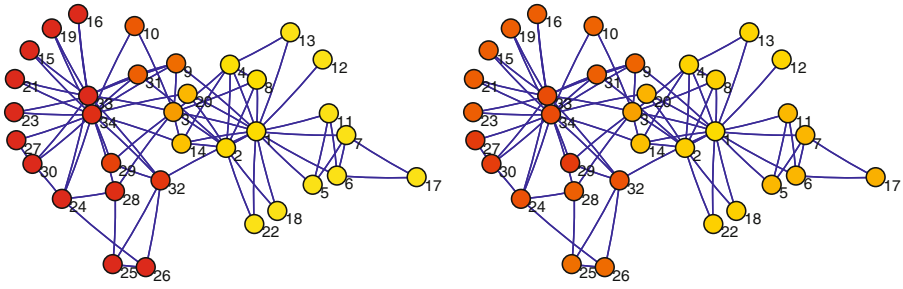


Fig. 2. Test of the three fuzzy c-means methods on the LFR benchmark for undirected and unweighted network with overlapping communities. The plot shows the variation of the normalized mutual information between the planted and the recovered partition, in its generalized form for overlapping communities [10]. The number of nodes $n = 1000$ and the average degree $\langle k \rangle = 20$, the other parameters are $\gamma = 2, \beta = 1$. For the upper two $s_{\min} = 10, s_{\max} = 50$ and for the lower two $s_{\min} = 20, s_{\max} = 100$. Each point corresponds to an average over 20 graph realizations.

node degrees are distributed according to a power law with exponent γ; the community sizes also obey a power law distribution, with exponent β. It is more practical to choose as independent parameter, the mixing parameter μ, which expresses the ratio between the external degree and the total degree of a node. In Figure 2,we show what happens if one operates the three fuzzy c-means methods on the benchmark, for $n = 1000$ and the average degree $\langle k \rangle = 20$. The other parameters are $\gamma = 2, \beta = 1$. We have chosen combinations of the extremes of the exponents' ranges in order to explore the widest spectrum of network structures. Each curve shows the variation of the normalized mutual information with the fraction of overlapping nodes. In general, we can infer that the fuzzy c-means type methods give good results.

3.3 The Karate Club Network

This network was constructed by Wayne Zachary after he observed social interactions between members of a karate club at an American university [17]. Soon after, a dispute arose between the clubs administrator and main teacher and the club split into two smaller clubs. It has been used in several papers to test the algorithms for finding community structure in networks [3,4,5,6,11,12,13,14]. The partitioning results are shown in Figure 3 and Table 1.



(a) (b)

Fig. 3. The fuzzy community structure for the karate club network, corresponding to two overlapping communities represent by the weighted color average described in [12]. (a)OP fuzzy c-means; (b)DD fuzzy c-means.

Table 1. The association probability of each node belonging to different communities for the karate club network. ρ_R or ρ_Y means the probability belonging to red or yellow colored community, respectively.

	Nodes	1	2	3	4	5	6	7	8	9	10	11	12
OP	ρ_R	0.0386	0.0782	0.4396	0	0	0	0	0.0037	0.6746	0.7640	0	0
	ρ_Y	0.9614	0.9218	0.5604	1.000	1.0000	1.0000	1.0000	0.9963	0.3254	0.2360	1.0000	1.0000
DD	ρ_R	0.1766	0.2510	0.3330	0.1942	0.3935	0.4356	0.4356	0.2095	0.5455	0.5318	0.3935	0.1962
	ρ_Y	0.8234	0.7490	0.6670	0.8058	0.6065	0.5644	0.5644	0.7905	0.4545	0.4682	0.6065	0.8038
	Nodes	13	14	15	16	17	18	19	20	21	22	23	24
OP	ρ_R	0	0.2271	1.0000	1.0000	0	0	1.0000	0.3030	1.0000	0	1.0000	1.0000
	ρ_Y	1.0000	0.7729	0	0	1.0000	1.0000	0	0.6970	0	1.0000	0	0
DD	ρ_R	0.1864	0.2426	0.6029	0.6029	0.4674	0.2227	0.6029	0.3191	0.6029	0.2227	0.6029	0.6054
	ρ_Y	0.8136	0.7574	0.3971	0.3971	0.5326	0.7773	0.3971	0.6809	0.3971	0.7773	0.3971	0.3946
	Nodes	25	26	27	28	29	30	31	32	33	34		
OP	ρ_R	1.0000	1.0000	1.0000	0.9651	0.8579	1.0000	0.7339	0.9103	1.0000	0.9631		
	ρ_Y	0	0	0	0.0349	0.1421	0	0.2661	0.0897	0	0.0369		
DD	ρ_R	0.5329	0.5472	0.7456	0.5569	0.6391	0.7353	0.5517	0.5734	0.7270	0.7287		
	ρ_Y	0.4671	0.4528	0.2544	0.4431	0.3609	0.2647	0.4483	0.4266	0.2730	0.2713		

4 Conclusions

In this paper, we test three fuzzy c-means methods, based on optimal prediction, diffusion distance and dissimilarity index, respectively, on two artificial networks, including the widely known ad hoc network with same community size and a recently introduced LFR benchmarks with heterogeneous distributions of degree and community size. All of them have an excellent performance, with the additional advantage of low computational complexity, which enables one to analyze large systems. They identify the community structure during iterations with a high degree of accuracy, with producing little different. Moreover, successful ap-

plications to real world networks confirm the capability among them and the differences and limits of them are revealed obviously.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under Grant 10871010 and the National Basic Research Program of China under Grant 2005CB321704.

References

1. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. 74(1), 47–97 (2002)
2. Newman, M., Barabási, A.L., Watts, D.J.: The structure and dynamics of networks. Princeton University Press, Princeton (2005)
3. Girvan, M., Newman, M.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99(12), 7821–7826 (2002)
4. Newman, M., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69(2), 026113 (2004)
5. Newman, M.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA 103(23), 8577–8582 (2006)
6. Zhou, H.: Distance, dissimilarity index, and network community structure. Phys. Rev. E 67(6), 061901 (2003)
7. Lafon, S., Lee, A.: Diffusion Maps and Coarse-Graining: A Unified Framework for Dimensionality Reduction, Graph Partitioning, and Data Set Parameterization. IEEE Trans. Pattern. Anal. Mach. Intel. 28, 1393–1403 (2006)
8. Danon, L., Diaz-Guilera, A., Duch, J., Arenas, A.: Comparing community structure identification. J. Stat. Mech. 9, P09008 (2005)
9. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E 78(4), 046110 (2008)
10. Lancichinetti, A., Fortunato, S.: Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys. Rev. E 80(1), 016118 (2009)
11. E.W., Li, T., Vanden-Eijnden, E.: Optimal partition and effective dynamics of complex networks. Proc. Natl. Acad. Sci. USA 105(23), 7907–7912 (2008)
12. Li, T., Liu, J., E.W.: Probabilistic Framework for Network Partition. Phys. Rev. E 80, 026106 (2009)
13. Liu, J.: Detecting the fuzzy clusters of complex networks. Patten Recognition 43, 1334–1345 (2010)
14. Liu, J., Liu, T.: Detecting community structure in complex networks using simulated annealing with k-means algorithms. Physica A 389, 2300–2309 (2010)
15. Lovasz, L.: Random walks on graphs: A survey. Combinatorics, Paul Erdos is Eighty 2, 1–46 (1993)
16. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York (2001)
17. Zachary, W.: An information flow model for conflict and fission in small groups. J. Anthrop. Res. 33(4), 452–473 (1977)