Physica A 464 (2016) 221–230

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Community detection in complex networks using density-based clustering algorithm and manifold learning

Tao You [a], Hui-Min Cheng [a], Yi-Zi Ning [a], Ben-Chang Shia [b], Zhong-Yuan Zhang [a,*]

[a] *School of Statistics and Mathematics, Central University of Finance and Economics, Haidian District, Beijing 100081, China*

[b] *Big Data Research Center & School of Management, School of Health Care Administration, Taipei Medical University, Taiwan*

HIGHLIGHTS

- A density-based clustering framework is proposed for community structure detection.
- An improved partition density is proposed to evaluate the quality of the detected communities.
- The framework is insensitive to its parameters, and easy to implement.
- The comparisons performed on the synthetic benchmarks and the real-world networks show the effectiveness of the framework.

ARTICLE INFO

Article history:
Received 29 December 2015
Received in revised form 21 June 2016
Available online 5 August 2016

Keywords:
Complex network
Community detection
Density-based clustering
Manifold learning
IsoMap

ABSTRACT

Like clustering analysis, community detection aims at assigning nodes in a network into different communities. Fdp is a recently proposed density-based clustering algorithm which does not need the number of clusters as prior input and the result is insensitive to its parameter. However, Fdp cannot be directly applied to community detection due to its inability to recognize the community centers in the network. To solve the problem, a new community detection method (named IsoFdp) is proposed in this paper. First, we use IsoMap technique to map the network data into a low dimensional manifold which can reveal diverse pair-wised similarity. Then Fdp is applied to detect the communities in the network. An improved *partition density* function is proposed to select the proper number of communities automatically. We test our method on both synthetic and real-world networks, and the results demonstrate the effectiveness of our algorithm over the state-of-the-art methods.

1. Introduction

Network offers a fresh perspective to model the complex systems from various areas. Compared to the limits of reductionism, it is a simple yet powerful data-based mathematical tool to reveal the fundamental laws behind the whole system [1–3]. Community structure detection is an important research topic for understanding the topological structures of the networks. Intuitively speaking, a community can be considered as a set of nodes which are interconnected with higher probability than connected with the rest of the network [4].

There are lots of methods that have been proposed to detect community structures in complex networks, such as modularity-based algorithms [5,6], random walk-based algorithms [7,8], clustering-based algorithms [9–13] and matrix

[*] Corresponding author.
E-mail address: zhyuanzh@gmail.com (Z.-Y. Zhang).

http://dx.doi.org/10.1016/j.physa.2016.07.025

decomposition-based algorithms [14–17]. A more detailed analysis can be found in Ref. [18]. Community detection is similar to clustering analysis in many aspects, and the state-of-the-art clustering algorithms such as K-means and DBSCAN can be easily altered to detect communities in networks [19–21]. Compared to K-means, there is no need to give the number of clusters as prior input for DBSCAN. However, DBSCAN is sensitive to its parameters, and slightly different parameter settings may lead to very different results. Therefore it is a hard task to find the proper parameter settings, and it is not an ideal approach to use a single pair of global parameters to describe the whole dataset [22].

In order to address the sensitivity issue, a novel density-based clustering algorithm, which succeeds the advantages of DBSCAN, was proposed [23]. For convenience, the algorithm is denoted as Fdp, which is the title abbreviation of Ref. [23]. The only parameter needed for Fdp is d_c, furthermore Fdp is insensitive to d_c [23]. However, like DBSCAN [24], Fdp still suffers from the so-called curse of dimensionality, which can make the distance functions misleading, since the distances of the high dimensional data points can be more uniform or even identical [25]. Unfortunately, network data has similar distance characteristics of high dimensional datasets, and we will expound this fact in Section 2. So Fdp cannot be used directly to detect communities before the network nodes are mapped into a low dimensional latent space. Moreover, it is very difficult to distinguish the proper center nodes manually in the decision graph from the others, which is used as the cluster centers in Fdp, especially when the network structure is fuzzy or the number of centers is large. Hence it is necessary to find an approach to detect center nodes automatically.

Inspired by the concept of the hidden metric spaces [26,27], which can be considered as variations of hidden variables [28–30], we embed the network into a low dimensional manifold to preserve the key properties of the original network. While finalizing this paper, we have been made aware of similar ideas developed independently in Refs. [31,32]. In this paper we use IsoMap [33], the state-of-the-art manifold learning algorithm, to map the nodes of network into a low dimensional manifold which can reveal diverse pair-wised distances as well as preserve the key properties of the original network. We use an improved *partition density* function to evaluate the detected community structures and to choose the appropriate number of the reduced dimensions for IsoMap and of communities for Fdp.

The rest of the paper is organized as follows: Section 2 presents some related works; Section 3 is the details of our proposed algorithm; Section 4 gives the experimental results and Section 5 concludes.

2. Related work

DBSCAN classifies all points in the dataset into core set, border set and noise set based on two parameters *Eps* and *MinPts*. Specifically, *Eps* determines the radius of each point in dataset *D*. The points in this radius are neighbors, and the points that have more than *MinPts* neighbors are considered as cores. Finally the core points form the main body of clusters through directly density reachable chains, and the ends of the chains form the borders, and those which are not in the *Eps*-radius of any cores will be considered as noises [24]. From the above procedure we can see that DBSCAN determines the core points based on two parameters directly. Whether a point is a core highly depends on the specific value of *Eps* and *MinPts*, and as a consequence, a tiny change of parameter setting may yield a very different clustering result.

On the contrary, the Fdp algorithm does not distinguish the points into core or border directly according to the density. Instead it converts the cluster center selection problem into the outlier detection problem through the decision graph approach which is based on the idea that any cluster center possesses higher local density and at a relatively larger distance from other cluster centers. The idea can be realized by two delicately designed measures:

$$\rho_i = \sum_j \chi(d_{ij} - d_c), \tag{1}$$

and

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}), \tag{2}$$

where d_{ij} is the distance of data points i and j, d_c is a threshold such that $\chi(d_{ij} - d_c) = 1$ when $d_{ij} - d_c < 0$, and $\chi(d_{ij} - d_c) = 0$ otherwise, ρ_i is the local density of data point i, and δ_i is the smallest distance of point i to the points with higher ρ_i. Fig. 1 shows the two-dimensional plot of δ_i and ρ_i for each data point, which is the so-called decision graph [23].

The values of δ_i and ρ_i of the cluster centers are both significantly larger than other points such that the user can recognize the centers from the upper right corner of the decision graph manually. From above procedures we can see that Fdp takes the advantage of the relative value of two measures (i.e. the ordering information) to highlight the center points, hence it is not sensitive to the parameter d_c. Finally, the rest points in the data can be assigned into different clusters by the *higher nearest rule* in a single step. This means that after the detection of the cluster centers, the other points follow their nearest neighbor's cluster assignment, which has a higher density ρ_i [23].

However, there are also disadvantages of utilizing ordering information when Fdp is applied to detect communities in networks, since network data possesses similar distance characteristics of high dimensional datasets. For example, considering a network $G = (V, E)$, in which V and E are the set of nodes and edges respectively, with the adjacency matrix $A = [a_{ij}]_{n\times n}$, where $a_{ij} = 1$ if nodes i and j are connected, $a_{ij} = 0$ otherwise. If the adjacency matrix $A = [a_{ij}]_{n\times n}$ is considered as a dataset matrix, then each node of the network G can be considered as a point in a n-dimensional space, where n is often very large. In such a high dimensional space, the distance or similarity between nodes can usually be very