ARTICLE IN PRESS

Physica A ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Physica A

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

A clustering algorithm for determining community structure in complex networks

Hong Jin, Wei Yu, ShiJun Li *

State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072, China
Computer School, Wuhan University, Wuhan 430072, China

HIGHLIGHTS

- By spectral analysis DENCLUE is able to deal with community detection problem.
- It reduces the number of parameters and Sheather–Jones plug-in can select its value.
- It is able to find community structure with arbitrary size and shape.
- It has good scalability in deal with high dimensional data.

ARTICLE INFO

Article history:
Received 19 January 2017
Received in revised form 26 September 2017
Available online xxxx

Keywords:
Community detection
Density based clustering
Spectral analysis
Parameter estimation

ABSTRACT

Clustering algorithms are attractive for the task of community detection in complex networks. DENCLUE is a representative density based clustering algorithm which has a firm mathematical basis and good clustering properties allowing for arbitrarily shaped clusters in high dimensional datasets. However, this method cannot be directly applied to community discovering due to its inability to deal with network data. Moreover, it requires a careful selection of the density parameter and the noise threshold. To solve these issues, a new community detection method is proposed in this paper. First, we use a spectral analysis technique to map the network data into a low dimensional Euclidean Space which can preserve node structural characteristics. Then, DENCLUE is applied to detect the communities in the network. A mathematical method named Sheather–Jones plug-in is chosen to select the density parameter which can describe the intrinsic clustering structure accurately. Moreover, every node on the network is meaningful so there were no noise nodes as a result the noise threshold can be ignored. We test our algorithm on both benchmark and real-life networks, and the results demonstrate the effectiveness of our algorithm over other popularity density based clustering algorithms adopted to community detection.

1. Introduction

Networks are extensively used to represent complex systems in social science, engineering, biology and so on. A common feature of these networks is community structure [1], forming group structures in networks based on work, friendship, family, and other types of relations [2]. In the World Wide Web communities may correspond to groups of pages dealing with the same or related topics. While in metabolic networks they may be related to functional modules such as cycles and

* Corresponding author at: Computer School, Wuhan University, Wuhan 430072, China.
E-mail address: shjli@whu.edu.cn (S. Li).

https://doi.org/10.1016/j.physa.2017.11.029
0378-4371/

pathways. Communities are also called clusters or modules with high concentrations of edges within special groups and low concentrations between these groups [3]. The nodes in one community probably share common properties or play similar roles within the group. In view of this, community detection can lead to important advances in complex systems analysis.

For this reason, a large variety of community detection algorithms have been proposed such as modularity-based algorithms, random walk-based algorithms, clustering based algorithms and matrix decomposition-based algorithms [4–9]. A more detailed analysis can be found in [10]. To some extent the process of community detection is similar to clustering analysis.

Because communities in networks often have an arbitrary size and shape, finding communities in complex networks is a challenging task [11]. Under these circumstances, new clustering methods have recently been introduced to solve this highly complex problem. Density-based clustering methods look for clusters of arbitrary size and shape and have hence been widely used in community detection. Recently, Huang et al. proposed a novel density-based network clustering method called graph-skeleton-based clustering (gSkeletonClu) [12]. This algorithm can find communities as well as hubs and outliers. However, the main difficulty for the gSkeletonClu algorithm is the relatively complicated parameter selection process. Though it provides a convenient way to automatically set the parameter value, excessive evaluation of some indexes is required to finally determine the appropriate parameter setting.

In order to address the sensitivity issue, a generalized density-based method was proposed, denoted as GDENCLUE for convenience, with the particular goal of community detection in the complex networks. Because all nodes are meaningful comparing with the original density based clustering algorithm, the parameter noise threshold can be eliminated, so that only the parameter the influence factor is needed for GDENCLUE. As a kernel based clustering approach, it introduces the influence function to characterize the density distribution of the dataset. The approximation of the density distribution function is obtained by summing up the influence functions. The process to study local minima of the density distribution function can be viewed as a tool to find the clusters from datasets. Moreover, the selection of the kernel scale parameter determines the number of clusters. In this algorithm named influence factor is determined by a Sheather–Jones plug-in.

This paper focuses on one approach of density based clustering in computer science. Combined with the spectral analysis, the complex network data is first transformed before applying a further clustering procedure [13]. It is able to produce excellent results as various approximate optimization techniques. However, it does not require an impractical large computational effort like other algorithms when optimizing the modularity exhaustively. When converted to Euclidean Space there was no noise present in the data, alleviating the dependency on parameters for density based clustering.

The rest of the paper is organized as follows. Section 2 provides a general introduction to spectral clustering analysis. While in Section 3 the nonparametric estimation approach sheather-Jones plug-in method is described. Then the clustering algorithm proposed for determining community structure is detailed in Section 4. Section 5 provides several experimental evaluations and discussions about our approach. Finally, the conclusions are drawn in Section 6 along with some issues for further work.

2. Common spectral clustering analysis

Given a set of data points $x_1, \ldots, x_n$ and some measurement of similarity $s_{ij} \geq 0$ between all pairs of data points x_i and x_j, common spectral clustering represents the data in form of the similarity graph $G = (V, E)$ [14]. In this graph each vertex v_i represents a data point x_i. If the similarity s_{ij} between the corresponding data points x_i and x_j is positive or larger than a certain threshold, then there exists an edge between these two vertices [15]. The most commonly used similarity measurement is Gaussian function as formula (1) shows:

$$s_{ij} = \exp\left(\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \tag{1}$$

In the following graph G is assumed to be weighted, each edge between two vertices v_i and v_j carries a non-negative weight $w_{ij} \geq 0$. Let $W = \left(w_{ij}\right)_{i,j=1,\ldots,n}$ be the weighted adjacency matrix of the graph. Then the problem of clustering can now be formulated using the weighted similarity graph. That is to find a partition of the graph such that the edges between different groups have very low weights and the edges within a group have high weights.

For a given complex network $G = (V, E)$ itself can be seen as a kind of similarity graph, its adjacency matrix A can be regarded as the weight matrix W of the graph. In which element A_{ij} is equal to 1 if node i has directly connected to node j and 0 otherwise. It can be represented as formula (2) shows:

$$A_{ij} = \begin{cases} 1 & \text{if } i, j \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The main tools for spectral clustering are graph Laplacian matrices that derived from W [16]. There is no unique convention how the different matrices denoted, we choose the normalized symmetric Laplacian as formula (3) shows:

$$L_{sym} = I - D^{-1/2} A D^{-1/2} \tag{3}$$

Please cite this article in press as: H. Jin, et al., A clustering algorithm for determining community structure in complex networks, Physica A (2017), https://doi.org/10.1016/j.physa.2017.11.029.