Knowledge-Based Systems 89 (2015) 487–496

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Label propagation based evolutionary clustering for detecting overlapping and non-overlapping communities in dynamic networks

Ke Liu [a,b], Jianbin Huang [a,*], Heli Sun [c], Mengjie Wan [a], Yutao Qi [b], He Li [a]

[a] School of Software, Xidian University, Xian 710071, China
[b] School of Computer Science and Technology, Xidian University, Xian 710071, China
[c] Department of Computer Science and Technology, Xian Jiaotong University, Xian 710049, China

ARTICLE INFO

Article history:
Received 20 January 2015
Received in revised form 2 August 2015
Accepted 4 August 2015
Available online 28 August 2015

Keywords:
Dynamic network
Overlapping community
Non-overlapping community
Label propagation

ABSTRACT

Since real-world networks evolve over time, detecting communities in dynamic networks is a challenging research problem with wide applications. In this paper, we first improve our previous method and propose a more stable algorithm which is label-propagation-based for the discovery of communities in complex networks. Then, we present a novel evolutionary clustering approach DLPAE for dynamic networks based on the stable algorithm. According to DLPAE, community labels of nodes are determined by their neighbors, and a confidence (i.e., the importance of its neighbor to the node) is attached to each neighbor. During clustering, the confidences of nodes are calculated in terms of the structures of the current network and the network at last timestamp. We compute confidences' variance of each node and update nodes' labels in a descending order according to the values. In our setting, each node can keep one or more labels with belonging coefficients no less than a threshold, which renders DLPAE suitable for detecting overlapping and non-overlapping communities in dynamic networks. Experimental results on both real and synthetic datasets show the ability of DLPAE to detect overlapping and non-overlapping communities in dynamic networks, and demonstrate its higher accuracy compared to other related methods.

1. Introduction

In recent years, the research of complex networks has attracted more and more attention owing to their great potential in capturing natural and social phenomena. Since complex networks usually change over time, dynamic networks are formed. The discovery of communities in dynamic networks has become a critical task.

Evolutionary clustering is an effective method for detecting communities in dynamic networks. Chakrabarti et al. [1] first addressed this issue and proposed an evolutionary clustering framework. The framework assumes that the cluster structure of a dynamic network changes little in a very short time, and therefore each community in the dynamic network should smooth out over time. For smoothing, the framework trades-off two distinct criteria, the snapshot quality and the history cost, at each timestamp. The snapshot quality simply reflects how well the clustering result captures the current network, and the history cost determines how much the current clustering result has deviated from the previous clustering result. Obviously, higher snapshot quality and lower history cost are expected in order to perform well. Inspired by this framework, several evolutionary clustering methods have been proposed [2–4].

However, all of the evolutionary clustering methods were designed for detecting disjoint communities in dynamic networks. In real social networks, communities are overlapped sometimes. Some other methods [5,6] can be used for the discovery of overlapping communities in dynamic networks, but both are incremental.

In this paper, we first improve the Dominant Label Propagation Algorithm DLPA [7] and propose a more stable algorithm DLPA+. Then, we present an evolutionary clustering approach DLPAE for dynamic networks based on DLPA+. According to DLPAE, community labels of nodes are voted by their neighbors, and a confidence is attached to each neighbor. During evolutionary clustering, confidences are computed between nodes and their neighbors. Each confidence here consists of two parts: the confidence of nodes in the current network and the confidence of the same nodes in the network at last timestamp. A user-defined parameter α is a trade-off between the two. After that, we compute confidences' variance of each node and update nodes' labels in a descending order according to the variance. In the setting of DLPAE, each node can possess one or more community labels with belonging coefficients no less than a threshold, and this property endows DLPAE with the ability to detect non-overlapping and overlapping

* Corresponding author.

http://dx.doi.org/10.1016/j.knosys.2015.08.015

communities in dynamic networks. By iteratively updating nodes' labels, each node in the dynamic network keeps one or more community labels in the end. For the discovery of non-overlapping communities, we choose the labels with the greatest belonging coefficients as the community labels of nodes, and for the discovery of overlapping communities, we preserve all the labels.

We summarize the main contributions of this paper as follows: (1) We improve the dominant label propagation algorithm DLPA and make it more stable. (2) We propose a novel clustering approach DLPAE for dynamic networks. (3) DLPAE has the ability to detect overlapping and non-overlapping communities in dynamic networks.

The rest of the paper is organized as follows: in Section 2 we review the related work. Section 3 presents the preliminaries. We show the improved dominant label propagation algorithm DLPA+ in Section 4 and our evolutionary clustering approach DLPAE is described in Section 5. The experimental results and analysis are given in Section 6. Finally, Section 7 concludes the study.

2. Related work

Community discovery in complex networks is a challenging research issue in recent years [8–11]. Label propagation [12–14] has been shown as a very efficient approach in this field owing to its simplicity and near-linear time complexity. However, all of these algorithms can only handle disjoint communities. Recently, two improved label propagation algorithms COPRA [15] and SLPA [16] were proposed to reveal overlapping communities in complex networks. We also proposed a dominant label propagation algorithm DLPA [7] to detect overlapping and non-overlapping communities in networks simultaneously.

There exists a growing body of literature on detecting communities in dynamic networks. Chakrabarti et al. [1] first introduced the concept of evolutionary clustering and proposed an evolutionary clustering framework. Chi et al. [2] inherited similar concept and proposed an evolutionary spectral clustering algorithm. Lin et al. [3] presented the FacetNet framework to analyze communities and their evolutions in a unified process in terms of non-negative matrix factorization. Kim and Han [4] designed a particle-and-density based evolutionary clustering method KH to detect a variable number of communities of arbitrary forming and dissolving in dynamic networks. Additionally, there also exist several evolutionary clustering methods for detecting overlapping communities in dynamic networks. Nguyen et al. [5] designed a two-phase framework called AFOCS for detecting overlapping communities as well as tracing their evolution in dynamic mobile networks. Cazabet et al. [6] proposed iLCD, an evolutionary clustering method that efficiently discovers overlapping communities via adding edges and then merging similar edges in dynamic networks. Some other works [17–21] also analyzed communities and their evolutions in dynamic networks.

3. Preliminaries

We have proposed a dominant label propagation algorithm DLPA [7] based on the traditional label propagation algorithm LPA [12]. It can be used for the discovery of overlapping and non-overlapping communities in networks. In the following, we introduce the related definitions which can be found in the literature [7].

3.1. Basic concepts

Similar to LPA, according to DLPA, nodes' community labels are determined by their neighbors. However, different neighbors may have different importance which we refer to as confidence. Given a node u, the confidence $\delta_u(v)$ of u w.r.t. its neighbor node v is defined as follows:

$$\delta_u(v) = \frac{sim(u,v)}{\sum_{w\in Nb(u)} sim(u,w)}, \tag{1}$$

where $Nb(u)$ denotes the neighbors of node u, and $sim(u,v)$ denotes the similarity between node u and node v which can be defined by the Jaccard similarity function:

$$sim(x,z) = \frac{|\Gamma(x)\cap\Gamma(z)|}{|\Gamma(x)\cup\Gamma(z)|}, \tag{2}$$

where $\Gamma(x) = Nb(x)\cup\{x\}$.

When updating nodes' labels, the belonging coefficient of a node u to a community c which also is the update rule for node u is defined as follows:

$$b(c,u) = \frac{\sum_{y\in Nb(u)} \theta(c,y)\cdot b(c,y)\cdot\delta_u(y)}{\sum_{y\in Nb(u)} b(Dl_y,y)\cdot\delta_u(y)}, \tag{3}$$

where $\theta(c,y) = \begin{cases} 1 & Dl_y = c, \\ 0 & others. \end{cases}$ Dl_y represents the dominant label of node y which is defined as follows:

$$Dl_y = \arg\max_c(b(c,y)). \tag{4}$$

Obviously, the dominant label of a node represents the community that the node most likely belongs to.

After the updating process, we introduce the inflation operator γ_{in} and the deletion operator ψ_τ for every node. The inflation operator γ_{in} is defined as follows:

$$\gamma_{in} b(c,u) = \frac{b(c,u)^{in}}{\sum_{j\in C} b(j,u)^{in}}, \tag{5}$$

where in is a user-defined inflation parameter and C is the label set attached to node u. After having been applied to the inflation operator, the belonging coefficient rises to the inth power. So, larger belonging coefficients get increases in values while smaller ones diminish their values.

The deletion operator ψ_τ is defined as follows:

$$\psi_\tau(u) = \{c|b(c,u) > \tau\}, \tag{6}$$

where τ is the cut threshold which is default set to $1/|Nb(u)|$. The operator is applied to delete the useless labels of each node. That is, DLPA preserves the labels whose belonging coefficients are greater than $1/|Nb(u)|$. If there is no label whose belonging coefficient is greater than $1/|Nb(u)|$, DLPA reserves only one label with the greatest belonging coefficient, and delete the others.

Like LPA [12], DLPA updates nodes' labels iteratively in a random order. The process does not terminate until a maximum iteration number T is reached or labels of nodes become stable. Consequently, each node keeps one or more labels. For detecting non-overlapping communities, DLPA chooses the dominant label of each node as its community label, and for overlapping communities, all the remaining labels are selected. Therefore, DLPA has the ability to detect non-overlapping and overlapping communities in networks.

3.2. Problem statement

We define a dynamic network $\mathcal{G} = \{G_1, G_2, \ldots G_t, \ldots\}$. At each timestamp $t, G_t(V_t, E_t)$ (V_t is the node set and E_t is the edge set) represents the network at that timestamp. Our goal is that, at each timestamp t, detecting the overlapping and non-overlapping community structures $CR_t^o = \{c_1^o, c_2^o, \ldots, c_m^o\}$ and $CR_t^n = \{c_1^n, c_2^n, \ldots, c_q^n\}$ of G_t, and for each community $c \in CR_t$, it tries to smooth out over time.