Contents lists available at ScienceDirect

Neurocomputing

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

Multi-objective community detection method by integrating users' behavior attributes

Peng Wu [a,b], Li Pan [a,b,*]

[a] Department of Electronic Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai, China
[b] National Engineering Laboratory for Information Content Analysis Technology, Shanghai Jiao Tong University, Shanghai, China

ARTICLE INFO

Article history:
Received 2 May 2015
Received in revised form
9 November 2015
Accepted 9 November 2015

Keywords:
Community detection
Structure clustering
Attribute categorization
Multi-objective optimization
Behavior analysis

ABSTRACT

Social networks usually have abundant attributes associated with users to describe their features. Behavior attribute is one of the most important types of attribute which can better reflect users' intrinsic interests. In practice, many network applications prefer communities that not only are densely intra-connected, but also have homogeneous attribute value on specific behavior attributes. Structure clustering and attribute categorization are two types of method which can take full advantage of structure information and attribute information to partition the network, respectively. In this paper, we propose a novel community detection method by realizing structure clustering technology and attribute categorization technology simultaneously. Specifically, structure clustering is realized by optimizing modularity which captures densely intra-connected nature of communities. As for attribute categorization, a new metric named as homogeneity is defined to achieve the goal that nodes within each community have homogeneous attribute value, while in different communities have diverse attribute values. A multi-objective optimization evolutionary mechanism is adopted to optimize modularity and homogeneity simultaneously. Extensive experiments on several real-world networks demonstrate that our method can get a set of community structures corresponding to different trade-offs between structure clustering and attribute categorization.

1. Introduction

Grouping similar objects together and keeping dissimilar objects apart [1,2] are important for data analysis. Community detection is one of the classic techniques to achieve this goal. Conventional community detection methods are mainly based on structure clustering, such as modularity optimization [3] and label propagation [4]. They take network topology structure as input and define a community as a group of nodes that are densely intra-connected while sparsely linked with the rest of the network. Besides connection information, real-world networks usually have abundant attributes associated with nodes to describe their properties, such as demographic data, behavior attributes and preference information of users in social networks. Note that the attribute information considered here is different from content information considered in other papers [5]. Attribute information is concise and has categorization ability, while content information is tanglesome and cannot categorize objects directly. Categorization methods [6,7] can take full advantage of attribute information to partition the networks intuitively. They take the nodes' attribute information as input and group objects with the same attribute values into the same groups. Each of structure clustering and attribute categorization makes use of only one type of information and ignore the other one. As a result, each tightly connected community detected by structure clustering methods may have a rather random distribution of attributes values, for which it is difficult to interpret the communities reasonably. In contrast, groups detected by attribute categorization methods may have very loose intra-community structures which fail to capture the frequent interaction nature within groups. One of the underlying reasons of above drawbacks is that although the formation of connections is partly influenced by the common attributes of objects, they do not collaborate all the time [8,9]. Therefore, both the topology structure and the node attributes should be taken into consideration for community detection. In fact, combing multiple types of information to solve problems has been demonstrated to be effective in many fields [10,11].

One of the most important attribute information in social networks is users' behavior attributes, such as retweeting, replying, clicking hyperlinks, clicking Like button, following famous homepages, etc. Behavior attributes can better reflect users' intrinsic interests, habits and characters. For example, in order to

[*] Corresponding author at: Department of Electronic Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai, China.
E-mail address: panli@sjtu.edu.cn (L. Pan).

http://dx.doi.org/10.1016/j.neucom.2015.11.128
0925-2312/© 2016 Elsevier B.V. All rights reserved.

promote products among a group of users interested in basketball, behavior attribute, e.g., following famous basketball star, can be adopted to guide the community detection. As for facilitating the important information diffusion in social networks, behavior attribute that denotes whether a user tends to retweet information or not can be used to detect active communities to launch the information. Furthermore, detecting communities with certain explicit behavior attributes is the basis of group behavior analysis and control in social networks. Above examples also indicate that community structures are usually homogeneous in certain attribute subsets rather than whole attribute set. Thus the dominant attribute subset used to detect communities should be selected explicitly and concisely based on the purpose of specific applications. In this paper, we study the problem about detecting community structures by combining structure and attribute information, especially behavior attributes information.

The community detection methods combining network structure and node attribute are designed mainly from two perspectives, i.e. pattern mining [12–15] and network partition [9,16–20]. The former allows nodes to belong to no community, in which case the community assignments of some key nodes may be ignored. What's more, pattern mining methods lack flexibility as they force every node in each community to have same or similar attribute values. This may lead to rather small or disconnected communities. Thus we are more interested in network partition methods which require each node to belong to at least one community. Such methods are further categorized into unified-model methods [9,16,17,19] and separate-model methods [18,20] based on different strategies handling two types of information. Unified-model methods treat the topology structure and node attributes in the same way by a unified model, such as a distance metric [9,21] or a Bayesian probabilistic model [17]. They ignore the fact that treating two types of information in a unified model will inevitably result in loss of information. Thus the unified-model methods fail to take full advantage of the partition ability of both types of information. What's more, they cannot adjust the relative importance of structure and attribute flexibly. In fact, the community structures with different relative importance of structure and attribute are usually different. Separate-model methods first model the topology structure and node attributes separately and then try to combine them to decide the final community structure. How to appropriately define such two models and how to flexibly combine and adjust them are crucial for obtaining meaningful community structures.

In this paper, we design a separate-model method which realizes structure clustering and attribute categorization simultaneously. The optimization of widely used modularity [22] which captures the densely connected feature of communities is adopted for structure clustering. In order to integrate attribute categorization into structure clustering flexibly, we define a new objective function named as homogeneity to evaluate the quality of attribute categorization. Then the attribute categorization can be realized by the optimization of homogeneity. The combination of structure clustering and attribute categorization can be realized by optimizing the modularity and the homogeneity simultaneously. Since the relative importance of structure and attribute is usually unknown in advance, we adopt a multi-objective optimization mechanism to obtain multiple community structures which correspond to different tradeoffs between structure information and attribute information. Community structures with respect to different tradeoffs are suitable for different applications. For example, for information diffusion control in social networks, solutions with larger modularity value may be selected because structure is more important for information diffusion. While for Internet marketing, community structures with larger homogeneity value are preferred because products tend to be promoted in some attribute-specific groups. Our method is named as Multi-objective Optimization Community Detection algorithms for networks with Attribute information which is termed as MOCDA for short. Extensive experiments on several real-world networks are implemented to demonstrate the good performance and illustrate the potential applications of the MOCDA.

The remainder of this paper is organized as follows. Section 2 presents related works. Section 3 describes the model formulations on both structure clustering and attributes categorization. Section 4 presents the multi-objective evolutionary algorithm. The experimental results are described in Section 5. Finally, Section 6 gives the conclusions.

2. Related works

Pattern mining and network partition are two kinds of community detection methods combining structure and attribute information. The pattern mining methods try to mine some node sets satisfying certain requirements. Moser et al. [12] study the problem of mining cohesive pattern which is a dense and connected subgraph that has homogeneous values in a large enough feature subspace. Pool et al. [13] try to find a diverse set of cohesive communities with concise descriptions. Similarly, Galbrun et al. [14] aim to find *k* communities so that the total edge density over all communities is maximized and each community is succinctly described by a set of labels. GAMer [15] is a synthesis of subspace clustering and dense subgraph mining. It tries to find sets of nodes that are densely connected within the associated graph and as well show high similarity regarding their attributes. The above pattern mining methods allow nodes to belong to no community and force nodes in each community to have same or similar values in some attributes. This may lead to rather small or disconnected communities. Thus we are more interested in the methods from network partition perspective.

The network partition methods are mainly categorized into unified-model methods and separated-model methods. The unified-model ones combine the topology information and attribute information of networks in a unified model. For example, SA-Cluster proposed by Zhou et al. [9] and its extended versions Inc-Cluster [16] define a unified neighborhood random walk distance on an augmented graph which combines both topology and attribute information. The K-Medoids method is adopted to cluster the network based on unified distance measure. BAGC method proposed by Xu et al. [17] adopts a Bayesian model to capture both structural and attribute information of a network. The community detection problem is transformed into a probabilistic inference problem and can be solved by an efficient variational algorithm. CODICIL [19] fuses the link strength with content similarity by creating content edges. However, structure and attribute are usually two completely different types of information. The unified-model methods will inevitably result in loss of information by treating them in a unified model.

On the other hand, the separated-model methods adopt different models to capture the topology and attribute information of networks and then try to combine them to decide the final community structures. The CESNA method proposed by Yang et al. [20] statistically models the links of the network and the node attributes by different probabilistic likelihood models respectively and combines them together with a hyperparameter which controls the scaling between two likelihoods. The hyperparameter needs to be set in advance. The AGCA proposed by Cruz et al. [18] adopts modularity and entropy to model the topology information and attribute information, respectively. However, the entropy defined by them does not capture the categorization nature of the attribute information. They define entropy of a group based on a