Pattern Recognition 81 (2018) 50–59

Contents lists available at ScienceDirect

Pattern Recognition

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

Community discovery in networks with deep sparse filtering

Yu Xie[a], Maoguo Gong[a,*], Shanfeng Wang[a], Bin Yu[b]

[a] *Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Xidian University, Xi'an, Shaanxi Province 710071, China*
[b] *School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi Province 710071, China*

ARTICLE INFO

Article history:
Received 8 April 2017
Revised 8 December 2017
Accepted 27 March 2018
Available online 27 March 2018

Keywords:
Community discovery
Sparse filtering
Deep learning
Network representation

ABSTRACT

In the past decade, network community discovery has attracted great attention from quite a few researchers, and community structure is one of the most significant properties in complex networks. This paper presents a novel method for network community discovery based on deep sparse filtering. The features of the network are extracted by sparse filtering, an unsupervised deep learning algorithm, from an efficient representation of the network. Consequently, extracted features are employed to partition the network. Experiment results on both synthetic and real-world network datasets indicate that the proposed algorithm especially based on S∅rensen–Dice's similarity matrix representation of the network is efficient and it outperforms several state-of-art algorithms in discovering community structure.

1. Introduction

With the rapid development of Internet, especially the era of Web 2.0 coming, unstructured data grow in petabytes per day. A large portion of them are graph data, produced from various online social networks, such as Facebook, Twitter, LinkedIn, WeChat and so on. In real world, all kinds of complex systems, such as both online and off-line social networks [1], computer networks [2], biochemical networks [3], protein-protein interaction (PPI) networks [4] and citation networks [5], can be represented as network graphs of highly abstract. These network graphs generally have an important feature, community structure [6], in addition to small world effect [7] and scale-free effect [8]. Community discovery is a technology of great significance in mining and analyzing the potentially valuable information and it has wide applications in the analysis of social networks, machine learning, biology, medical science and criminology, etc.

Community discovery aims at separating the whole network into several tightly connected parts which are also called communities. In other word, the intra connections in the same community are much denser than the inter connections with other communities. A large number of efforts have been devoted to developing community discovery algorithms. These efforts are closely related to graphical segmentation in computer science and hierarchical clustering in sociology. These conventional methods can be divided into three types: decompose algorithms, agglomerative algorithms and optimization algorithms. Based on graph partitioning problem in graph theory, decompose algorithms begin with their concerned network, try to find node pairs of low similarity and remove them step by step till divide the graph into small and relatively independent partitions. Although they are usually non-deterministic polynomial (NP) problems, a large number of algorithms can achieve some effects by estimating some index. For instance, Girvan et al. presented famous GN algorithm in [9], which holds a novel view that connections within the community are very tight, and connections between communities are sparse. Although decompose algorithms can be used in community discovery of large networks, it is hard to determine the pros and cons of divisive results because of their uncertainty. It is necessary for agglomerative algorithms to calculate similarities of all pairs of nodes, start with nodes of high similarity and add edges into these nodes. Girvan et al. presented FN algorithm in [10], based on the edge betweenness. Clauset et al. presented CNM algorithm [11], an improved version of FN algorithm by reducing its complexity. For agglomerative algorithms, the core nodes can be well divided, but easily wrong for the peripheral nodes. Optimization algorithms are based on the maximization of an objective, such as optimize to seek the largest modularity of the graph, including extremal optimization [12], simulated annealing [13], greedy algorithm [14], specifically. For example, Duch et al. presented EO algorithm [12], based on an extremal optimization of the value of modularity. Compared with the former two types of algorithms, optimization algorithms can effectively find a proper solution with high quality in a reasonable period of time. However, some of them are easily getting into local optimal solutions, sensitive to the optimized order or they have parameters to tune.

[*] Corresponding author.
E-mail address: gong@ieee.org (M. Gong).

https://doi.org/10.1016/j.patcog.2018.03.026
0031-3203/© 2018 Elsevier Ltd. All rights reserved.

More recently, DeepWalk [15], LINE [16], Node2vec [17] and Struc2vec [18] learn node representations to characterize the network structure based on random walks on graphs. Although deep feature learning models are popular by virtue of its strong nonlinear representation power [19], there is only a little work on deep learning for efficient unsupervised learning of network features. Tian et al. [20] dealt with a deep representation for graph clustering, in which stacked sparse autoencoders are adopted to reconstruct normalized similarity matrix of a network and then k-means algorithm is run on the embeddings to obtain clustering results. When applied to large networks, the normalized similarity matrix representation of a network is not conducive to become parallelized and it is difficult to calculate the normalized similarity, for which calculating and multiplying the inverse of the degree matrix is necessary. Yang et al. [21] investigated the strong power of stacked autoencoders to represent nonlinear features of modularity matrix of a network in community detection and extended it to a semi-supervised version. Almost both of them are challenging to implement for the reason that there are many super parameters to be fined tuned in autoencoders.

In reality, the size of a network is extremely large and the connection structure is also very sparse and complex. Even though an autoencoder learns models that can provide good approximations of the true distribution, it scales poorly to large sets of features. In contrast, sparse filtering (SF) [22], a simple deep feature learning algorithm presented by Ngiam et al. in 2011, can not only avoid explicitly modeling the data distribution and tuning many super parameters, but also scale gracefully to handle high-dimensional inputs. Inspired by this, we pay our attention to presenting a flexible and efficient model to discover communities in networks, based on deep sparse filtering algorithm and k-means algorithm. The main highlights of our proposed model are as follows:

(1) Based on sparse filtering and k-means clustering, a new algorithm for community discovery is proposed to cluster different nodes into different communities. In order to find a desirable network representation, four different network representations are introduced as the input of our algorithm.
(2) A new similarity constraint is proposed to make the proposed algorithm learn more efficient features for community discovery.
(3) The performance of the proposed algorithms and four different network representations are investigated by using two types of synthetic networks and seven real-world networks.

The remainder of this paper is organized as follows. Section 2 deals with the related backgrounds of the proposed method. The detailed descriptions of the proposed algorithm are given in Section 3. In Section 4, the performance of the proposed algorithm is validated on both synthetic networks and real-world networks, we also compare our algorithm and its constrained version with six state-of-the-art approaches. Finally, the concluding remarks are summarized in Section 5.

2. Related backgrounds

As has been mentioned roughly before, the proposed algorithm is highly related to community discovery, deep learning and sparse filtering. In this section, the related backgrounds of the proposed algorithm are given in detail.

2.1. Community and community discovery

To be convenient for analyzing a complex network, a network can be expressed by a graph that consists of nodes and edges, as shown in Fig. 1(a). Let graph $G = \{V, E\}$ represent a network where V representing the network objects is a set of nodes, and E representing the relations between the objects is a set of edges. The graph G has $\mid V \mid = m$ nodes and $\mid E \mid = n$ edges. The connections of the graph G can be represented as an adjacent matrix A, whose every element represents the relation between two nodes.

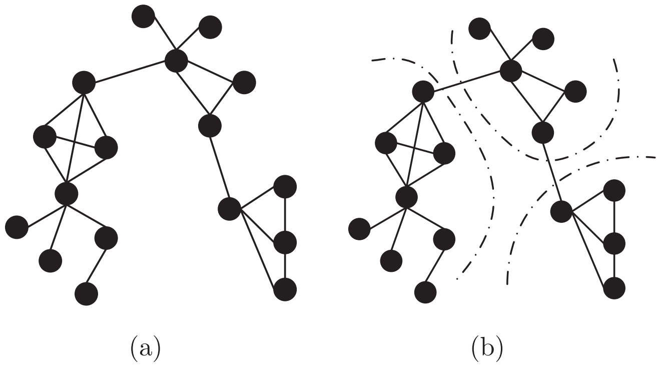

Fig. 1. Graphical illustration of (a) graph-modeled network and (b) network community discovery.

In [23], Radicchi et al. introduced a definition of community based on the degree of nodes. Suppose that a generic node i belongs to a community C, a subgraph of G, the degree of node i can be represented as $k_i(C)= k_i^{in}(C)+k_i^{out}(C)$, where $k_i^{in}(C)=\sum_{j\in C} A_{ij}$ is the number of edges connecting node i to other nodes belonging to C and $k_i^{out}(C)= \sum_{j\notin C} A_{ij}$ is clearly the number of connections toward nodes in the rest of the network. A community in a strong sense can be formulated as $k_i^{in}(C) > k_i^{out}(C), \forall i \in C$. A community in a weak sense can be formulated as $\sum_{i\in C} k_i^{in}(C) > \sum_{i\in C} k_i^{out}(C), \forall i \in C$. It means that, in a strong community, each node has more connections within the community than with the rest of the graph, and in a weak community, the sum of the degrees within the sub-graph is larger than the sum of degrees towards the rest of the network.

A complex network is usually made up of a great many of groups or clusters. So far, community discovery does not have a unified and specific definition in the literature. The long accepted definitions are based on linking density of nodes or connectivity of the graph respectively. A little similar to general clustering, both of them in a network can be regarded as finding its sub-graphs, in which internal degree, the total weight of the edges within the sub-graph, is larger than its external degree, the total weight of edges towards the rest of the graph, as shown in the toy model in Fig. 1(b).

2.2. Deep learning and sparse filtering

Deep learning is motivated from neuroscience findings [24–26], which reveals the principles of information processing in the mammalian brain. In neocortex, the raw information propagates through a hierarchical deep architecture instead of being explicitly preprocessed [25]. Deep learning just learns to represent the raw information with abstract and conceptualized features. Inspired from the mammalian brain, the breakthrough of deep learning is extracting complex and abstract features by a fast greedy training layer by layer instead of low-level features. As a boom of deep learning in both academia and industry, its several models have become more and more popular like deep belief networks (DBN) [27], followed by convolutional neural networks (CNN) [28], autoencoders (AE) [29], Sparse filtering [22] and so on. Deep learning achieved similar properties to that of neocortex and excellent performance in a wide variety of applications, such as image recognition [30–33], speech recognition [34,35], image enhancement [36], text understanding [37], image labeling [38], time-series problem [39,40] and image segmentations [41,42], etc. At the same time,