Information Sciences 435 (2018) 263–281

Contents lists available at ScienceDirect

Information Sciences

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

Nonnegative matrix factorization with mixed hypergraph regularization for community detection

Wenhui Wu [a], Sam Kwong [a,b,*], Yu Zhou [c], Yuheng Jia [a], Wei Gao [a]

[a] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
[b] City University of Hong Kong Shenzhen Research Institute, Shenzhen 51800, China
[c] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

ARTICLE INFO

Article history:
Received 26 July 2017
Revised 3 January 2018
Accepted 7 January 2018
Available online 8 January 2018

Keywords:
Community detection
Nonnegative matrix factorization
Hypergraph regularization

ABSTRACT

Community structure is the most significant attribute of networks, which is often identified to help discover the underlying organization of networks. Currently, nonnegative matrix factorization (NMF) based community detection method makes use of the related topology information and assumes that networks are able to be projected onto a latent low-dimensional space, in which the nodes can be efficiently clustered. In this paper, we propose a novel framework named mixed hypergraph regularized nonnegative matrix factorization (MHGNMF), which takes higher-order information among the nodes into consideration to enhance the clustering performance. The hypergraph regularization term forces the nodes within the identical hyperedge to be projected onto the same latent subspace, so that a more discriminative representation is achieved. In the proposed framework, we generate a set of hyperedges by mixing two kinds of neighbors for each centroid, which makes full use of topological connection information and structural similarity information. By testing on two artificial benchmarks and eight real-world networks, the proposed framework demonstrates better detection results than the other state-of-the-art methods.

1. Introduction

Networks can be widely seen in real world, which are one of the most complex data, e.g., transportation networks, social networks and biological networks [10,16,36]. Networks can be modeled as graphs which are composed of nodes and edges. For instance, the individuals or organizations can be represented as nodes in social networks, while the interactions are depicted by the edges. The analyses on the network structures are beneficial for the exploration and understanding of networks.

In networks, one of the most remarkable characteristics is the community structure, such as the overlapping community structure [25] and the hierarchical community structure [19]. A community is a group of nodes which are tightly connected with each other but rarely connected with the outside nodes. Community detection is a significant task for many applications, such as web page clustering [9], musical rhythmic pattern extraction [6] and protein function prediction [30]. Besides, some assessments of a partitioned community structure are researched [17,18,23].

* Corresponding author at: Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong.
E-mail addresses: wenhuiwu3-c@my.cityu.edu.hk (W. Wu), cssamk@cityu.edu.hk (S. Kwong), yu.zhou@my.cityu.edu.hk (Y. Zhou), yhjia3-c@my.cityu.edu.hk (Y. Jia), weigao5-c@my.cityu.edu.hk (W. Gao).

https://doi.org/10.1016/j.ins.2018.01.008
0020-0255/

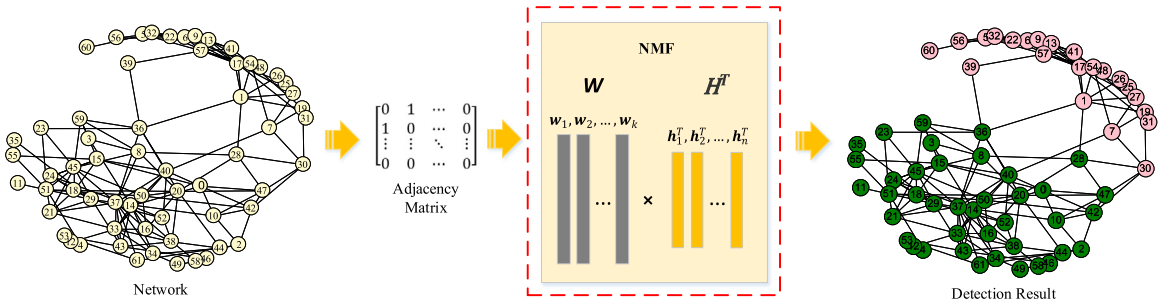

Fig. 1. Flowchart of the nonnegative matrix factorization based community detection algorithm. The topology information in the network can be encoded as an adjacency matrix. By decomposing the adjacency matrix into two nonnegative matrices, a low-dimensional representation can be achieved. Finally, nodes can be classified into different communities based on the new representation.

In the past few years, numerous community detection methods have been proposed. Among them, modularity-based algorithms are particularly popular, e.g., fast greedy modularity optimization [5], extremal optimization [7] and spectral optimization [24,44]. Another category is divisive algorithm, e.g., GN algorithm [10], which removes the intracommunity edges to obtain clusters. Many dynamic algorithms also proposed with the merits of low computational complexity, e.g., label propagation (LP) algorithm [27], Potts model. Besides, some overlapping community detection algorithms were proposed [17], such as seed expansion approach [37] and fuzzy clustering [32]. However, most of these methods only utilize the topology information. When it comes to complex and vague structure with many intracommunity links, the performance of these methods will be greatly affected. In order to address this problem, some research works try to make use of prior information to improve the detection performance in a semi-supervised manner. Cheng et al. [4] presented a semi-supervised method based on active learning to generate high-quality must-link and cannot-link constraints. Eaton et al. [8] proposed a semi-supervised algorithm, which is based on the spin-glass model. Liu et al. [20] combined the idea of graph regularization with pairwise constraint and proposed a semi-supervised nonnegative matrix factorization model. However, the real-world scenarios usually lack the desired prior information and it would consume much manpower to access the prior information. In addition, the performance of semi-supervised algorithms heavily depend on the quality of prior information, which might not be available in practice. Hence, given that semi-supervised methods are case-dependent and the prior information is scarce, the unsupervised methods are more desirable.

Without the prior information and human intervention, the unsupervised algorithms are self-organizing. More recently, one of the unsupervised algorithms, nonnegative matrix factorization (NMF) [26] based community detection, has attracted huge attention. Different from the previously described methods, NMF based method is a fuzzy community detection method where each node has a soft membership to each community. Inspired by the generative process of networks, NMF based method supposes that one network can be divided into a number of low-dimensional subspaces. The coefficient vectors in the new space are the soft membership vectors, which decide relationships between all pairs of nodes and communities. Fig. 1 shows the flowchart of a typical NMF based community detection method. Recently, theoretical advances on NMF enable some variants of NMF to be applied to community detection, such as symmetric NMF (SNMF) [34] and nonnegative matrix tri-factorization [43]. However, these methods only aim at minimizing the error between the factorized matrices and the original matrix, which ignore the underlying information in the network. Nodes within the same community may have totally different low-dimensional representations as shown in Fig. 2(b). Nodes 28 and 30 are two connected nodes, which belong to the same community. However, for nodes 28 and 30, the representations obtained from NMF based algorithm are totally different, which results in the misclassification of node 30. In order to solve the above problem, He et al. [13] proposed an extension of SNMF named graph regularized SNMF (GSNMF) method, where graph regularizer could smooth the variation in new space between two nodes with large similarity [3,12,45]. However, one community is a cluster of nodes, and the relationships among nodes in one community should not be simply pairwise, while GSNMF only considers the pairwise relationship of nodes. Essentially, modeling the higher-order relationship among nodes will significantly improve the performance of NMF [42].

Inspired from the theory of hypergraph learning [46], namely, a centroid within a hyperedge connected with more than two nodes by the hyperedge, we encode the higher-order information into NMF by hypergraph, and propose a framework named mixed hypergraph regularized NMF (MHGNMF) in this paper. The proposed MHGNMF contains two community detection methods, which are based on two kinds of cost functions for NMF, respectively. It is worth noting that the proposed framework can be extended to many existing variants of NMF. Our key insight of MHGNMF is that the new representation of nodes, which are within the same hyperedge, should be highly correlated. Since hypergraph construction is crucial to the proposed framework, we design a mixed hypergraph in this paper. Specifically, two sets of hyperedges are firstly generated by two approaches, i.e., topological connection and structural similarity, and then two hyperedges are mixed together for each centroid. Consequently, MHGNMF, which takes higher-order relationship into consideration and simultaneously takes advantages of topological connection and structural similarity, is able to learn a more discriminative representation. As shown in Fig. 2(c), contrast to the misclassification by NMF, the representations of nodes 28 and 30 obtained by MHGNMF