Expert Systems With Applications 91 (2018) 277–285

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An overlapping network community partition algorithm based on semi-supervised matrix factorization and random walk

Weimin Li[a,*], Jun Xie[a], Mingjun Xin[a], Jun Mo[a,b]

[a] School of Computer Engineering and Technology, Shanghai University, Shanghai, China
[b] Shanghai Key Laboratory of Computer Software Evaluating and Testing, Shanghai, China

ARTICLE INFO

Article history:
Received 16 June 2017
Revised 7 September 2017
Accepted 8 September 2017
Available online 9 September 2017

Keywords:
Matrix factorization
Random walk
Node convergence degree
Node influence

ABSTRACT

The discovery of community structure is the basis of understanding the topology structure and social function of the network. It is also an important factor for recommendation technology, information dissemination, event prediction, and more. In this paper, we consider the structure and characteristics of the social network and propose an algorithm based on semi-supervised matrix factorization and random walk. The proposed method first calculates the transition probability between nodes through the topology of the network. The random walk model is then used to obtain the final walk probability, and the feature matrix is constructed. At the same time, we combine a priori content information in the network to build a must-link matrix and a cannot-link matrix. We then merge them into the feature matrix of the random walk to form a new feature matrix. Finally, the expectation of the number of edges is defined according to the factorized membership matrix. Results demonstrate the effectiveness and better performance of our method.

1. Introduction

Many complex systems in the real world such as social networks can be analyzed using the network. Community structure exists in all real networks, so it is significant to identify the community structure, to understand the function and dynamic change of the network. At present, some of the commonly used community partition algorithms are based on modularity optimization (Duch & Arenas, 2005), graph partitioning (Kernighan & Lin, 1970), node label broadcast (Raghavan, Albert, & Kumara, 2007) and other algorithms. In many social networks, the community is naturally overlapping. For example, a person will also belong to the family network and friend network. To solve the problem of community overlapping, researchers proposed the algorithms such as clique filtering algorithm (Palla, Derényi, Farkas, & Vicsek, 2005), node local diffusion algorithm (Lancichinetti, Fortunato, & Kertész, 2009), and fuzzy clustering algorithm (Gregory, 2011). All communities cannot be determined with limited information. These undiscovered communities should have a critical role in overlapping community detection. Many researchers use nonnegative matrix factorization (NMF) for community partitioning and obtain the potential community structure information of the network by analyzing the low-dimension matrix after factorization of the matrix.

NMF is an unsupervised learning algorithm, which is usually used for data dimensionality reduction and feature extraction. For a nonnegative matrix X, it can be decomposed into the product of two matrices, denoted as $X \approx U \times V$, where U and V are nonnegative. NMF is widely used in various fields such as predicting the score in recommendation systems (Koren, Bell, & Volinsky, 2009). NMF was used to transform the adjacency matrix of the network into the low-dimension matrix for analysis (Zhang, Wang, & Zhang, 2007). Wang et al. proposed a symmetric nonnegative matrix factorization (SNMF) algorithm (Wang, Li, Wang, Zhu, & Ding, 2011) to decompose the feature matrix of the network into two symmetric low-dimension matrices. In other areas, researchers add priori knowledge of the network to the matrix factorization model, to guide the learning process. Zhang proposed an enhanced network community partition algorithm based on partial network information (Zhang, 2013). Ma et al. proposed a network community-partitioning algorithm combined with semi-supervised clustering and SNMF (Ma, Gao, Yong, & Fu, 2010). In most existing NMF algorithms, the feature matrix is represented by the adjacency matrix of the network; so the information of nodes without edges cannot be expressed, and the network information is limited. Although these methods obtained excellent results, most of them omitted that one node may participate in more than one community. In this paper, we propose an overlapping community partition algorithm based on random walk and semi-supervised matrix factorization. Firstly, we can calculate the transition probability between nodes by using the topology of the network. Then, the random

* Corresponding author.
E-mail addresses: wmli@shu.edu.cn, 108wml@gmail.com (W. Li), xie_jun@i.shu.edu.cn (J. Xie), xinmj@shu.edu.cn (M. Xin).

http://dx.doi.org/10.1016/j.eswa.2017.09.007
0957-4174/

walk model obtains the final walk probability between nodes. The feature matrix is constructed based on the random walk and proposed node convergence. At the same time, a must-link matrix and a cannot-link matrix are formed according to a priori knowledge of the network. A new feature matrix is formed by fusing them into the feature matrix of the random walk. Finally, the expectation of the number of edges is defined according to the factorized membership matrix. Results demonstrate the effectiveness and better performance of our method.

The main contributions of this paper are as follows:

(1) The convergence of the node is defined to integrate the structure and attribute of the nodes.
(2) We model the random walk based on the convergence of the node and construct the feature matrix based on the random walk and node convergence.
(3) A community partition algorithm based on random walk and semi-supervised matrix factorization is presented to obtain better community structure than traditional matrix decomposition models.
(4) We report the results of several experiments on original data obtained from Polblogs and Digital Bibliography & Library Project (DBLP).

This article is organized as follows: Section 2 introduces the related work of overlap network community. Section 3 presents the algorithm of this paper. Section 4 describes the experiment and the conclusion follows.

2. Related works

With the continuous development of the network, the network community partition algorithm has become a popular research topic. In recent years, many network community partition algorithms have been proposed, such as module-based optimization algorithms (Newman & Girvan, 2004; Newman, 2004), graph partition algorithms (White & Smyth, 2005; Zhou, Cheng, & Yu, 2009), label propagation (Gregory, 2010), and hierarchical clustering (Nepusz, Petróczi, Négyessy, & Bazsó, 2008; Yang, Di, Liu, & Liu, 2013). However, these algorithms are non-overlapping network community partition algorithms, that is, nodes in the network belong to only one community. In the real world, the community in the network will overlap with each other, that is, a node may belong to multiple communities at the same time. Several researches studied overlapping community partition algorithms to solve this problem. These algorithms are divided into two categories: node-based and edge-based overlapping network community partition algorithms.

The node-based overlapping network community partition algorithms divide nodes into different communities based on the structure and attribute information of the nodes. Palla et al. proposed the clique percolation method (CPM) to mine the overlapping structure in the network (Palla et al., 2005). Lancichinetti et al. proposed an algorithm based on node local diffusion and fitness function optimization, which can discover overlapping network community structure and hierarchical structure at the same time (Lancichinetti et al., 2009). The different choices of the initial nodes will result in different community partition results. Unlike the traditional clustering algorithm, fuzzy clustering does not divide a node into a community. It calculates the membership value of the node for the community. Gregory proposed a fuzzy algorithm to compute the membership vector of nodes for all network communities. Nepusz et al. proposed an optimization algorithm based on the combination of fuzzy clustering and simulated annealing (Nepusz et al., 2008).

The edge-based overlapping network community partition algorithms divide the edges in the network into the non-overlapping communities. If the different communities have a common node, the node is an overlapping node. Ahn et al. proposed an overlapping network community partition algorithm based on a line graph and link partition, which obtains the linked network through hierarchical clustering and then transforms the connected network into an overlapping node community (Ahn, Bagrow, & Lehmann, 2010). Shi et al. proposed a network community partition algorithm based on genetic algorithm (Shi, Cai, Fu, Dong, & Wu, 2013). Pan et al. proposed a network community partition algorithm based on local edge diffusion, which defines a local fitness function based on the edge community.

Recently, network community partition algorithms based on NFM have been proposed. Zhang et al. first applied the NMF model to overlapping communities, and divided the overlapping communities by the membership matrix. However, this algorithm needs to initialize a membership threshold, and the different membership thresholds differed greatly in community partition results (Zhang et al., 2007). In 2011, Wang et al. proposed SNMF, which decomposes the adjacency matrix into two symmetric low-dimensional nonnegative matrices (Wang et al., 2011). Cao et al. proposed a new model based on NMF algorithms, which discover overlapping communities, central nodes, and fringe nodes at the same time (Shi et al., 2013). The overlapping network community partition algorithm based on the NMF model mentioned above is an unsupervised algorithm. At present, some researchers have proposed the semi-supervised matrix factorization method to make full use of the priori knowledge in the network. Ma et al. proposed an SMNF algorithm, which combines the domain knowledge of the network to strengthen the original feature matrix, and establishes the must-link matrix M_{ml} and the cannon-link matrix M_{cl} to supervise the initial matrix (Ma et al., 2010). Similarly, Sun et al. proposed an enhanced semi-supervised matrix factorization algorithm. The algorithm puts forward the concept of transfer constraint through logical reasoning. The must-link constraint is used to improve the results of community partition more efficiently than the cannot-link constraint (Ma et al., 2010). Yang et al. used the priori knowledge to construct the regular constraint matrix into the matrix factorization optimization model (Cao, Wang, Jin, Cao, & He, 2013).

In most network community partition algorithms based on the matrix factorization mentioned above, the feature matrix X to be decomposed is usually represented by an adjacency matrix. This representation can only present the relationship between nodes directly connected to the network. The relationship between nodes that are not directly connected cannot be represented, so the resulting feature matrix contains limited network information. The random walk model can calculate the transition probability between any two nodes in the network. In this paper, we use the random walk model to describe the relationship between any two nodes in the network and constitute the feature matrix X. The priori knowledge is utilized in the network to supervise the feature matrix X, and the overlapping nodes are divided according to the decomposed matrix.

3. A network community partition algorithm based on node convergence degree and random walk

In this section, the node probability based on node aggregation degree is defined. The probability is added to the iterative equation of the random walk to obtain the probability of nodes in the entire network. To discover better community structure, we construct the feature matrix based on the random walk and design matrix decomposition using priori network information. Through model learning, the optimal parameters can be obtained and used to identify the overlapping community. Finally, the community par-