Knowledge-Based Systems 148 (2018) 47–54

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Community aware random walk for network embedding

Mohammad Mehdi Keikha[a,b], Maseud Rahgozar[a,*], Masoud Asadpour[a]

[a] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
[b] University of Sistan and Baluchestan, Zahedan, Iran

ARTICLE INFO

Article history:
Received 14 October 2017
Revised 21 January 2018
Accepted 18 February 2018
Available online 20 February 2018

Keywords:
Representation learning
Network embedding
Community detection
Skip-gram model
Link prediction

ABSTRACT

Social network analysis provides meaningful information about behavior of network members that can be used for diverse applications such as classification, link prediction. However, network analysis is computationally expensive because of feature learning for different applications. In recent years, many researches have focused on feature learning methods in social networks. Network embedding represents the network in a lower dimensional representation space with the same properties which presents a compressed representation of the network. In this paper, we introduce a novel algorithm named "CARE" for network embedding that can be used for different types of networks including weighted, directed and complex. Current methods try to preserve local neighborhood information of nodes, whereas the proposed method utilizes local neighborhood and community information of network nodes to cover both local and global structure of social networks. CARE builds customized paths, which are consisted of local and global structure of network nodes, as a basis for network embedding and uses the Skip-gram model to learn representation vector of nodes. Subsequently, stochastic gradient descent is applied to optimize our objective function and learn the final representation of nodes. Our method can be scalable when new nodes are appended to network without information loss. Parallelize generation of customized random walks is also used for speeding up CARE.

We evaluate the performance of CARE on multi label classification and link prediction tasks. Experimental results on various networks indicate that the proposed method outperforms others in both Micro and Macro-f1 measures for different size of training data.

1. Introduction

There has been remarkable growth in online social networks and the number of their users. Valuable information can be extracted from social networks by analyzing both their structure and content. Machine learning techniques are used as a way to extract valuable features from social networks for different analysis tasks such as classification [1–3], recommendation [4,5] and link prediction [6–9]. These learning methods can be both supervised and unsupervised. Supervised learning algorithms are able to extract features better for a specific task on social networks but their scalability would be challenging for large networks. On the other hand, unsupervised methods can handle scalability of feature learning methods; however, the extracted features show low accuracy in different network analysis tasks. They are too general to give valuable information for a specific task. [10–16].

Network embedding, as an unsupervised representation learning task, tries to extract informative lower dimensional representation of network nodes. It learns social relationships of network nodes in a low dimensional space to preserve both microscopic and macroscopic network structure including various proximity orders, community membership and their inherent properties. These representation vectors can be used in different social network analysis tasks such as classification [17], recommendation [18] and link prediction [6]. Some of classic network embedding methods use eigenvectors of affinity graph as feature vectors [10,15,19,20]. Graph factorization is another technique which is used for network embedding [21]. The aforementioned approaches suffer from scalability for large social networks.

In recent years, deep learning as an unsupervised method is widely used in natural language processing which a detailed description of these researches can be found in [11]. There are also many researches that have used deep learning for social network embedding [22–25]. Network embedding methods try to represent graph nodes with some informative feature vectors. DeepWalk [22], LINE [23] and Node2vec [24] are the most important methods that are proposed in the recent years. Though, these methods

* Corresponding author.
E-mail addresses: mehdi.keikha@ut.ac.ir (M.M. Keikha), rahgozar@ut.ac.ir (M. Rahgozar), asadpour@ut.ac.ir (M. Asadpour).

https://doi.org/10.1016/j.knosys.2018.02.028
0950-7051/

show good performance in comparison to other graph representation methods such as Spectural clustering, but they attempt to extract only local structural information from each node, and then employ them to learn final representation of the node. However, communities are important structural information ignored by these methods [26].

Community structure imposes constraints in a higher structural level on the nodes' representation. The representation of nodes within a community should be more similar than those belonging to different communities. Furthermore, for two nodes within a community, even if they only have weak relationship in local structure due to the data sparsity issue, their similarities will also be strengthened by the community structure constraint. Thus, incorporating community structure in network embedding can provide effective and rich information to solve data sparsity issues in global structures and moreover, make the learned nodes' representation more discriminative [25].

In this paper, we propose a new network embedding method called "CARE," which utilizes community information of network nodes to capture more structural information of networks. Some previous researches tried to embed community information on nodes' representation. For instance, Grover et al. in [24] only consider the community members that their distance to the source nodes is less than 2. However, in real-world networks which communities have thousands of members, Node2vec would not be able to consider information about the nodes that their distance is more than two from the source of random walk because Node2vec creates second order random walks. CARE can also produce the representation vector of nodes for arbitrary type of networks such as weighted, complex and directed.

CARE, firstly, extracts communities of the input network. We prepare this information with the Louvain method [27] which has effective performance on different social networks. To learn final representations, we generate some community aware random walks that consider both first and higher order proximities as well as community membership information for each node. The customized paths contain the nodes that are in the same neighborhood structure as well as nodes that belong to the same community. CARE makes several customized paths for each network node to embed different structural information into final representation. Finally, the customized random walks are used as contextual information to learn final representation of nodes by the Skip-gram learning model.

CARE is evaluated with two social network analysis tasks: multi label classification and link prediction. The experimental results show that CARE outperforms Node2vec with a gain of 50% on multi label classification with BlogCatalog dataset and 3% on the link prediction task for PPI dataset.

To summarize, we make the following contributions:

- We present a novel network embedding algorithm named CARE that learns the representation of nodes for different types of networks such as: weighted, directed and complex networks.
- Our method can preserve community information of the network in the learned representation vectors while the previous researches are not able to define an optimization function considering this information explicitly.
- CARE preserves all properties of the network structure through the generation of customized paths for each node, independently. Therefore, it spends less time to learn final representations of nodes because of parallel path generation.
- We empirically evaluate the algorithm on multi label classification and link prediction problems with different real world social networks. The experimental results indicate the efficiency of CARE in contrast to other network embedding methods.

The rest of paper is organized as follows: In Section 2, we summarize related works to network embedding. We explain details of CARE in Section 3. Section 4 outlines the experimental results on two network analysis tasks. Finally, Section 5 presents conclusion and future works.

2. Related works

In this section, we review recent researches related to unsupervised representation learning of network nodes. Some feature learning approaches use adjacency matrix of the network and try to preserve the first order proximity of nodes. These researches act as dimension reduction methods and find the best eigenvectors of network matrices [10,15,16,19–21,28] to use as the feature vector of networks. Eigenvector decomposition is usually computationally expensive. Furthermore, they only consider immediate neighborhood of nodes and do not use higher order proximities and community information. So, they are unable to preserve the global structure of networks. As a result, the learned representations would not provide an appropriate performance on diverse network analysis tasks.

In recent years, deep learning is used as an alternative to learn feature vector of network nodes. These methods have utilized deep learning to learn representation vectors. They generate random walks with different graph exploration strategies and have embedded them as contextual information into the Skip-gram model. DeepWalk was the first method that used the Skip-gram model [22]. It treats DFS like search strategy to generate random walk. Despite the good performance on multi label classification, this method failed to preserve global network structure because it does not consider community information of network nodes. LINE uses first and second order proximities to learn nodes' representation, but it also preserves local information of the networks [23]. The authors in [23] define two independent functions for first and second order proximities but they ignore community information. LINE and DeepWalk also fail to learn representation vector for network edges.

Node2Vec makes random walks based on DFS and BFS like strategies [24]. While Node2vec uses two controlled parameters to consider both homophily [29] and structural equivalences [30] of networks, it does not guarantee to reach different nodes of a community. The main reason for this problem is that these algorithms only consider second order proximities and cannot reach the nodes that their distance is more than 2 from the start node of random walk. Because in real networks, there are many nodes in a community and obviously their distance is greater than two, thus Node2vec would not consider all the community members during creation of random walks for a node. SDNE proposes a semi-supervised deep model, which has multiple layers of non-linear functions, thereby being able to capture the highly non-linear network structure [31]. It exploits the first and second-order proximity jointly to preserve the network structure, but it doesn't use community information.

The proposed method in [25] uses modularized non negative matrix factorization to preserve both microscopic and macroscopic information of networks. The authors in [25] define two independent model to embed local and community information independently and then optimize the joint function to learn the representation of nodes. They learn local and community structure separately. Consequently, they combine the final representations. Their final representation is not general enough to be used in different network analysis tasks because It also has some local structure information loss because it combines first and second order proximities in a unified matrix. Unification of matrices leads to missing information about different proximities during representation learning. Their method also suffers from scalability when the networks