Knowledge-Based Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A fusion probability matrix factorization framework for link prediction

Zhiqiang Wang, Jiye Liang*, Ru Li

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China

ARTICLE INFO

Keywords:
Network data analysis
Probability matrix factorization
Link prediction
Fusion model

ABSTRACT

Link prediction is a fundamental research problem in network data analysis. Networks usually contain rich node-to-node topological metrics and their effective use is crucial to solve the link prediction problem. Despite significant advances, the existing metric-based link prediction methods usually only consider one single topological metric and thus show some limitations in different types of networks; the existing matrix factorization-based models mainly focus on modeling the adjacent matrix of a network, and this is hard to ensure the modeling of those topological metrics that can play an important role in link prediction. This study develops effective approaches by fusing the adjacent matrix and some key topological metrics in a unified probability matrix factorization framework. In these approaches, we consider not only the symmetric metrics but also the asymmetric metrics which are usually not taken into consideration in the related work. In our probability matrix factorization framework, we first present two fusion models by fusing two kinds of metrics respectively, and based on the fusion models, we put forward the final fusion models which fuse the two kinds of metrics simultaneously. To verify the performance of all the fusion models, we conduct the experiments with six directed networks and six undirected ones, and the extensive experiments show that the proposed models provide impressive predicting performance for link prediction.

1. Introduction

Link prediction is a fundamental and important problem in network data analysis [1]. The solution to the problem is essential to explain the reason of network structure generation, to help us explore the law of network evolution [2], and to understand the mechanism of complex systems [3,4]. Furthermore, it is also of great significance for many applications, such as finding friends in social networks [5], recommending items in user-item networks [6], finding experts in academic networks [7], and discovering unknown interactions in protein-protein networks [8].

Research on link prediction has draw increasing attention in recent years. Many methods have been proposed by researchers from physics, biology, sociology, and computer science [9–14], including the metric-based methods, the classification-based methods, the probabilistic graph model (PGM)-based methods and the matrix factorization (MF)-based methods. As one kind of important link prediction methods, MF-based methods solve the link prediction problem mainly through focusing on modeling the low-rank approximation of the adjacent matrix of a network. Some existing work [15,16] has shown the advantages of MF-based methods in solving link prediction problem such as its robustness to the networks from different domains and its scalability to large datasets.

Though powerful, existing MF-based methods still have some problems that could limit their applicability and prediction accuracy. The core of MF is to get the low-rank approximation of the adjacent matrix of a network, but compared with the network itself, the presented information of the adjacent matrix is insufficient. This is because the adjacent matrix only present the observed links of a network, but actually a network still contains rich topological metrics like the common neighbors between nodes and the path length between nodes. No evidence indicating that the existing MF-based models can give dual attention to modeling the observed network links and those topological metrics which can play an important role in link prediction.

Based on the above considerations, we have the following motivations:

- Whether a MF-based model can be built to fuse the adjacent matrix and the topological metrics between nodes in a network?
- Whether a MF-based model should consider both the symmetric metrics and the asymmetric metrics between nodes in a network?
- How can we take the two factors into account in one MF-based model if both the symmetric metrics and the asymmetric metrics are related to the links formation in a network.

* Corresponding author.
E-mail addresses: zhiq.wang@163.com (Z. Wang), ljy@sxu.edu.cn (J. Liang), liru@sxu.edu.cn (R. Li).

https://doi.org/10.1016/j.knosys.2018.06.005
Received 27 November 2017; Received in revised form 4 June 2018; Accepted 7 June 2018

Therefore, the objective of this study is to develop fusion models which fuse the adjacent matrix with some topological metrics together in one unified probability matrix factorization framework. In the framework, we first present two fusion models by fusing the symmetric metrics and the asymmetric metrics respectively, and based on the fusion models, we put forward the final fusion models which fuse the two kinds of metrics simultaneously.

This paper makes the following contributions:

- We propose fusion models to fuse two sides of network information, i.e. the adjacent matrix and some key topological metrics, in one unified probabilistic matrix factorization framework.
- Our fusion models consider not only the symmetric metrics but also the asymmetric metrics which are usually not taken into consideration in the related work.
- We conduct experimental evaluations with various directed and undirected network datasets, and our models get impressive predicting performance for link prediction.

The rest of the paper is organized as follows: Section 2 introduces the related work; Section 3 presents the building of the fusion models in the probabilistic matrix factorization framework, where the symmetric metrics and asymmetric metrics are fused respectively, and a final fusion model is proposed to fuse the two kinds of metrics simultaneously. In Section 4, we conduct a series of experiments to evaluate the proposed methods on various directed and undirected network datasets. Finally we conclude our work in Section 5.

2. Related work

For link prediction, there have been several excellent surveys [17–20] from different research perspectives. Liben-Nowell and Kleinberg [19] provided useful information for the prediction problem, especially some classical prediction measures based on topological information of networks. Lü and Zhou [21] summarized recent progress about link prediction algorithms, emphasizing the contributions from physical perspectives and approaches. Wang et al. [17] investigated the link prediction from the perspective of computer science, and systematically summarized all typical work on the link prediction in social networks. Martínez et al. [22] discussed a large number of proposed techniques focusing on undirected and unweighed networks. In this section, we will first give a brief summary of the related work following the researching line of MF-based methods which are related to the methodology of this paper, and after that we will provide a brief review of the work done in the area of link prediction.

2.1. Matrix factorization based link prediction

Matrix factorization is a type of technique to get the low-rank approximation (LAR) and global information of the adjacent matrix of a network. The classical matrix factorization methods such as singular value decomposition (SVD) [23], non-negative matrix factorization (NNMF) [24] and probabilistic matrix factorization (PMF) [25] can be directly used for solving the link prediction problem. Liben-Nowell and Kleinberg [19] investigated various types of link prediction methods, and among them the LAR-based link prediction methods are implemented by using SVD. Chen et al. [26] put forward a link prediction algorithm based on NNMF. Zhu et al. [27] proposed a scalable temporal link prediction model via NNMF. Yang et al. [28] combined link prediction method by convex NNMF with block detection to predict potential links using both of global and local information. In a word, the extensive experiments showed that the classical MF-based methods were effective in solving the link prediction problem.

In addition to the classical MF-based methods, some factorization methods also have been designed according to the characteristics of the link prediction problem to improve the prediction performance. Menon and Elkan [15] proposed a MF-based method to address the class imbalance problem by directly optimizing for a ranking loss, which is optimized with stochastic gradient descent and scales to large graphs. Zhai and Zhang [16] attempted to solve the link prediction problem by combining MF and Autoencoder (AE), and they utilized dropout to train both the MF and AE parts and the results showed that it could significantly prevent overfitting by acting as an adaptive regularization. Song et al. [29] proposed a rank-one alternating direction method of multiplier (ADMM) for nonnegative matrix factorization, and their experiment results demonstrated that rank-one ADMM is more effective than multiplicative update rule, alternating least square, and traditional ADMM.

Despite these significant advances, current state-of-the-art MF-based models mainly focus on modeling the adjacent matrix of a network, and this is hard to ensure the modeling of those topological metrics that can play an important role in link prediction. Intuitively, it is of certain potential to improve the performance of link prediction if the MF-based methods can give dual attention to modeling the adjacent matrix and some key topological metrics. Therefore, this paper intends to deal with the problem, which is the starting point of the study.

2.2. The other link prediction methods

Apart from the MF-based methods, the metric-based methods, the classification-based methods, and the PGM-based methods are also the mainstream methods in link prediction.

The metric-based methods address the link prediction problem by measuring the similarity between nodes, such as the neighbors-based metrics [30–37], path-based metrics [38–41], and random walk-based metrics [42–46]. David Liben-Nowell and Kleinberg [19] tested several topological metrics on social collaboration networks, and the results showed that the Katz [41] metric and its variants performed consistently well, and that some of the very simple metrics including common neighbors and the Adamic-Adar metric [33] also performed surprisingly well. Zhou et al. [31] compared a number of topological metrics on disparate networks which included the protein-protein interaction network, the electronic grid, the Internet, and the US airport network. The extensive experimental results showed that the Resource Allocation [33] metric performed best, while common neighbors and Adamic–Adar metric [33] had the second-best performance. Also, other topology-based metrics were proposed to solve the link-prediction problem [38–40,42,47–49]. Despite those significant advances, the effectiveness of the metrics depends on the domain, the specific network, and the available information.

The classification-based methods treat link prediction as a binary classification problem. In a classification-based link prediction model, the features are defined on each pair of nodes, and these features can be constructed in topological or non-topological. The topological features (such as the neighbors-based metrics and the path-based features) are the commonly-used features in a classification-based link prediction model [50–52]. Except for the topological features, the non-topological features (such as users' location, interests, and educational backgrounds) are often selected to improve the classification-based link prediction models [53–55]. In the classification-based methods, it is still a challenge to predict links because the class imbalance problem can be difficult to deal with and most models are prone to yield biased results.

The PGM-based methods solve the link prediction problem by building a statistical network model. The hierarchical network model [56] models a network as a hierarchical random graph and the linking possibility between nodes can be calculated by the probability expectation. Stochastic block models [57,58] assume that the network nodes can be partitioned into some blocks, and that the linking probability between any two nodes depends on which block the nodes belong to. Latent-feature models [59–62] are kinds of probabilistic generative model, where the nodes' latent-features and the edges in a