Physica A 515 (2019) 558–565

Contents lists available at ScienceDirect

Physica A

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

Cold-start link prediction in multi-relational networks based on network dependence analysis

Shun-yao Wu [a], Qi Zhang [a,*], Chuan-yu Xue [a], Xi-yang Liao [b]

[a] Qingdao University, Qingdao 266071, China
[b] Ping An Technology (Shenzhen) Co.Ltd., Shenzhen 518000, China

HIGHLIGHTS

- An efficient method for network dependence analysis is proposed via projection correlation.
- Two kind of methods with multiple interactions are proposed for cold-start link prediction.
- It is promosing for cold start link prediction to establish regression between latent factors of sub-networks.

ARTICLE INFO

Article history:
Received 11 May 2018
Received in revised form 14 August 2018
Available online xxxx

Keywords:
Cold-start link prediction
Network dependence analysis
Projection correlation statistics
Robust principle component analysis

ABSTRACT

Cold-start link prediction has been a hot issue in complex network. Different with most of existing methods, this paper utilizes multiple interactions to predict a specific type of links. In this paper, multiple interactions are abstracted as multi-relational networks, and robust principle component analysis is employed to extract low-dimensional latent factors from sub-networks. Then a distribution free independence test, projection correlation, is introduced to efficiently analyze dependence between target and auxiliary sub-networks. Furthermore, associated auxiliary networks are exploited for cold-start link prediction, which aims to forecast potential links for new/isolated nodes in target sub-networks. Experimental results on 8 bioinformatics datasets validate rationality and effectiveness of the method.

1. Introduction

Many complex systems can be elegantly described as networks, in which nodes and links represent individuals/factors and interactions/relationships respectively. Link prediction in complex networks could detect potential links by taking full advantages of topological structure, and has been widely used in many fields, such as protein interaction prediction [1] and movie recommendations [2]. Recently link prediction has already become the focus of complex network.

Due to limited data availability, previous studies in complex network mainly managed single-relational networks containing only one kind of links. Local similarity indices, such as Common Neighbors index and related variations (Jaccard index, Adamic–Adar index and Resource Allocation index) [3,4], were widely applied for single-relational link prediction. Despite simplicity and efficiency, the above methods were ineffective sometimes. Thus, quasi-local indices and global similarity indices, such as Katz index and Local Random Walk [5], were proposed to exploit global topological structure. Despite these achievements, how to precisely predict missing links based on sparse information is a challenging issue,

* Corresponding author.
E-mail address: qizhang@qdu.edu.cn (Q. Zhang).

https://doi.org/10.1016/j.physa.2018.09.082
0378-4371/

since most of entries in the adjacency matrices of many real-world networks are zeros [6]. Recently, matrix completion was introduced to solve this challenge, which achieved great increase compared to many state-of-the-art algorithms [6–8].

However, "cold-start" problem imposed great challenges and has not been conquered yet. Generally, most existing work explore this problem in two scenes: (i) partial cold start with low-degree nodes; (ii) pure cold start with isolated nodes [9]. And this paper focuses on the latter scene. For partial cold start, Zhu et al. [10] proposed a parameter-dependent index to uncover missing links. For pure cold start, content-based strategy [11] provided a promising solution by utilizing nodal attributes, such as gender and hobbies in online social networks. Recently, an promising strategy for the pure cold start is to utilize multiple interactions to predict a specific type of interactions. For instance, Wu et al. [12] utilized chemical structure similarities to predict functional interactions between new drugs and marked drugs.

Link prediction with multiple interactions aims to exploit useful auxiliary interactions for predicting target interactions. Thus, It is necessary to eliminate irrelevant auxiliary interactions and select associated auxiliary interactions. However, efficiently analyze network dependence is a big challenge especially for large-scale networks. Degree–degree correlation and link overlap analysis are traditional methods to analyze network correlation by exploiting local topological structures. Degree–degree correlation analysis utilizes correlation coefficients, such as Pearson correlation coefficient [13], Spearman's rank correlation coefficient and Kendall's Tau correlation coefficient [14], to calculate correlation between degree distributions. Link overlap analysis employs link-weight vectorization rather than original adjacency matrix to explore network correlation [15,16]. Different with the above methods, Dai et al. [17] evaluated node importance by a variation of random walk to capture global topology of networks, and adopted Pearson's coefficient to estimate similarities of node importance between different networks. In order to efficiently handle large-scale networks, Wu et al. [12] extracted low-dimensional factors via the latent space network model, and examined correlation between factors through the likelihood ratio test. However, the assumptions of the likelihood ratio test appears strong, and the test cannot examine nonlinear associations between networks. Fortunately, distribution-free test of independence has achieved remarkable improvement. For instance, Cui et al. [18] proposed a new test of independence between a categorical variable and a continuous variable based on mean variance index. Lately, Zhu et al. [19] used projection correlation to examine dependence between two random vectors, which is instructive to analyze non-linear associations between latent factors.

In this work, robust Principal Component Analysis (robust PCA) is introduced to extract low-dimensional latent factors. Pech et al.'s work [6] demonstrated the effectiveness of robust PCA, which could discover not only missing links but also spurious links. And it is necessary to deal with spurious links, since measurement, instrumental, computational, and even human errors, interaction data collected from many fields usually mix with noise ineluctably. With the latent factors, a distribution free independence test with projection correlation statistics [19] is adopted to select relevant auxiliary interactions. Finally, random forest are employed to model the associated latent factors for predicting target interactions. Experiments on 4 protein interaction datasets and 4 drug interaction datasets validate rationality and effectiveness of the proposed method.

2. Cold-start link prediction in multi-relational networks

Multi-relational network $G = (V, E)$ is applied to abstract complex system with multiple interactions. Herein, $V = \{v_i\}_{i=1}^n$ represents the node set, and the link set E contains m types of links, $E = E_1 \bigcup E_2 \ldots \bigcup E_m$. Links in E_j constitute sub-network G_j, and $\mathbf{A}_j$ denotes the adjacency matrix of G_j, $\mathbf{A}_j = [a_{xy}^j]_{x,y=1}^n$. For unweighted networks, $a_{xy}^j = 1$ if $(v_x, v_y) \in E_j$ and $a_{xy}^j = 0$ otherwise; for weighted networks, a_{xy}^j denotes the weight of the link between nodes v_x and v_y and $a_{xy}^j \in \mathbb{R} - \{0\}$ if $(v_x, v_y) \in E_j$. Without loss of generality, G_1 is target sub-network while $G_2 \ldots G_m$ are auxiliary sub-networks.

2.1. Network dependence analysis with projection correlation

We extract low dimensional factors as well as denoise interaction data by means of robust PCA at first. The adjacency matrix of G_j is factorized as

$$\begin{aligned} \mathbf{A}_j &= \mathbf{L}_j + \mathbf{S}_j + \mathbf{E}_j \\ &= \mathbf{U}_j \mathbf{V}_j^T + \mathbf{S}_j + \mathbf{E}_j, \end{aligned} \tag{1}$$

where $\mathbf{L}_j$ is a low-rank matrix, $\mathbf{S}_j$ is a sparse matrix and $\mathbf{E}_j$ is the white noise. The positive entries in $\mathbf{S}_j$ represent spurious links while negatives stand for potential links. $\mathbf{L}_j = \tilde{\mathbf{U}}_j \tilde{\mathbf{\Gamma}}_j \tilde{\mathbf{V}}_j^T = (\tilde{\mathbf{U}}_j \tilde{\mathbf{\Gamma}}_j^{1/2})(\tilde{\mathbf{V}}_j \tilde{\mathbf{\Gamma}}_j^{1/2})^T = \mathbf{U}_j \mathbf{V}_j^T$. $\tilde{\mathbf{\Gamma}}_j$ is k_j order singular values diagonal matrix, and $\mathbf{U_j}, \mathbf{V_j}, \tilde{\mathbf{U}}_j^T, \tilde{\mathbf{V}}_j^T \in \mathbb{R}^{n \times k_j}$. Denote λ_l as the lth decreasing singular value, and k_j could be determined through accumulation contribution rate restriction $C_j = \sum_{l=1}^{k_j} \lambda_l^2 / \sum_{l=1}^{n} \lambda_l^2 \geq \theta$ $(0 < \theta < 1)$. And the latent factors of network G_j are organized as $\mathbf{N}_j = [\mathbf{U}_j, \mathbf{V}_j]$. Specially, $\mathbf{N}_j = \mathbf{U}_j$ when G_j is symmetrical.

Then the dependence between G_1 and $G_2, \ldots, G_m$ could be efficiently analyzed via the low-dimensional latent factors. In this paper, a novel distribution free independence test, projection correlation, is introduced to examine dependence between the latent factors. Practically speaking, given two networks G_i and G_j, the null and alternative hypotheses respectively are

$$H_0 : \mathbf{N}_i \text{ and } \mathbf{N}_j \text{ are independent} \quad \textit{versus} \quad H_1 : \text{otherwise}. \tag{2}$$