Physics Letters A 381 (2017) 3405–3408

Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

Cold-start link prediction in multi-relational networks

Shun-yao Wu, Qi Zhang*, Mei Wu

Qingdao University, Qingdao 266071, China

ARTICLE INFO

Article history:
Received 1 July 2017
Received in revised form 20 August 2017
Accepted 21 August 2017
Available online 24 August 2017
Communicated by C.R. Doering

Keywords:
Cold start
Link prediction
Latent space network model
Correlation test

ABSTRACT

During the last decade, interaction data have accumulated exponentially in many fields and provide a new opportunity for cold start link prediction. It seems necessarily to take full advantages of diversified information. However, correlation between different interactions has to be pre-tested. Therefore, this paper abstracts complex systems as multi-relational networks, and employs latent space network model to extract low-dimensional factors of sub-networks and adopts likelihood ratio test to examine correlation between factors. Then, regression between target sub-networks and correlated auxiliary sub-networks could be established for cold start link prediction. Experiments on 8 bioinformatic data sets validate the effectiveness and potential of our strategy for network correlation analysis and cold-start link prediction.

1. Introduction

Complex network provides some computationally convenient tools for modeling complex systems related to social, biological and traffic fields, in which nodes represent entities or factors and links represent relationships or interactions between nodes. In many cases of interest, potential links detected with well-known network information is termed as "Link Prediction"[1]. To date, link prediction has been a hot issue in complex network, and widely applied into biomolecular interaction prediction [2], social relationship recommendation [3] and so on.

Previously link predictions mainly face the topological structures of single-relational networks, which consist of one simplex links [4]. However, this procedure leaves link prediction for isolated or new nodes in suspense. The above problem is so-called cold start link prediction [5], and one common strategy in recommendation system is content-based recommendation [6] by using the nodal attribute information. For example, it doesn't make sense to exploit network information to recommend reasonable friends for a new registered WeChat user (analogue of an isolated node in the online social network). However, additional information like age, gender and hobbies could start a viable way.

Fortunately the era of big data is affluent in various large-scale interaction data to cold start, just like life sciences and web mining data. For example, a loyal Tencent QQ user tries to register for WeChat, and the relationship records of Tencent QQ supply powerful covariant for describing his/her new social performance. Here is another example. It is promising to utilize various kinds of interactions, such as drug chemical structure similarities, to predict drug–drug interactions for downstream experimental validation. Herein, drug–drug interactions denote adverse and synergistic interactions between drugs, and a new drug's function validation usually cost expensive experiments [7]. Therefore, intuition suggests that multiple interactions could be another potential assistant for cold start.

However, it's non-trivial to design effective and efficient cold-start link prediction methods with multiple interactions. First of all, whether correlation among multiple interactions exists should be examined. Recently, complex systems with multiple interactions were abstracted as multi-layer network [9], multi-dimensional network [1] or multi-relational network [8] (multi-relational network is adopted in this paper). A multi-relational network contains several sub-networks, and each sub-network characterizes one type links. Although lots of work studied the relatedness between sub-networks depends on degree–degree correlation [10] or link overlap analysis [11], how to efficiently infer correlation in large-scale is still a big challenge. Secondly, provided that there exists correlation between sub-networks, how to take advantage of the correlation to predict links for isolated or new nodes of the target sub-network is another challenging issue.

In this work, latent space network model [15] is introduced into link prediction. Firstly, low-dimensional latent factors are extracted from the adjacency matrices of multiple sub-networks via single value decomposition (denoted as *SVD*). Then, likelihood ratio test is employed to examine whether correlation exists between sub-networks via their latent factors. Finally, regression based on correlated factors is established. Experimental results on 8 bioinformatic data sets validate the effectiveness and potential of our approaches.

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

http://dx.doi.org/10.1016/j.physleta.2017.08.046
0375-9601/© 2017 Elsevier B.V. All rights reserved.

2. Network correlation analysis

Denote multi-relational network as $G=(V,E)$. $V=\{v_i\}_{i=1}^{n}$ is the n nodes set, and E includes q types links $E=E_1\bigcup E_2 \ldots \bigcup E_q$. Each type links constitute sub-network $G_j=(V,E_j)$, where A_j denotes the adjacency matrix of G_j, $1\leqslant j\leqslant q$. Generally speaking, G_1 is considered as the target sub-network, and our subject is to predict its potential links for new or isolated nodes based on the other auxiliary sub-networks. Quantities of link prediction methods concerned multi-relational networks have been developed [12,13], especially Dai et al. [8] proposed a belief propagation algorithm for measuring sub-network correlation without model reduction but costing super complexity.

In recent years complex network values correlation much more. Degree–degree correlation and link overlap analysis are mainstream strategies. The first one analyzes the correlation between degree distributions of any two networks. Nicosia et al. [10] employed Spearman and Kendall's tau rank for two degree distributions. The other strategy usually emphasizes link-weight vectorization rather than original adjacency matrix. Barigozzi et al. [14] contrasted different international-trade networks by covariance. Szell et al. [11] utilized jaccard coefficient to quantify interdependencies between online social networks. In sum, both strategies merely extract local network information. Moreover, link overlap analysis works hardly in case of large-scale. Recently, Fosdick et al. [15] proposed low-dimensional latent variable model to explain network structure and relative dependency test for latent variable and nodal attributes. Inspired by this work, this paper intends to construct similar latent variable models for each network, and perform likelihood ratio test to examine their correlation.

Noted by directed networks, such as twitter and gene regulatory networks, we concern networks whether symmetric or not. According to the latent model of previous paragraph, adjacency matrix of network G_j could be expressed as

$$A_j=\mu_j\mathbf{1}_n\mathbf{1}_n^T+\mathbf{a}_j\mathbf{1}_n^T+\mathbf{1}_n\mathbf{b}_j^T+U_jV_j^T+E_j, \tag{1}$$

where $\mathbf{1}_n=(1,\cdots,1)_{n\times 1}^T$; E_j is $n\times n$ white noise; μ_j is overall mean; n-dimensional $\mathbf{a}_j$ and $\mathbf{b}_j$ show all nodes' sender (outgoing) and receiver (incoming) additive effect; especially, $n\times k_j$ U_j and V_j compress high order dependence to multiplicative interaction $U_jV_j^T$. The above parameters are organized as latent factor $N_j=[\mathbf{a}_j,\mathbf{b}_j,U_j,V_j]$, which can be directly estimated through *SVD* algorithm as

$$\begin{aligned}
\hat{A}_j&=\hat{U}_j\hat{\Gamma}_j\hat{V}_j{}^T=\breve{U}_j\breve{V}_j{}^T\\
&=(\tilde{U}_j+\mathbf{1}_n\boldsymbol{\mu}_{U_j}^T)(\tilde{V}_j+\mathbf{1}_n\boldsymbol{\mu}_{V_j}^T)^T\\
&=\tilde{\mu}_j\mathbf{1}_n\mathbf{1}_n^T+\tilde{\mathbf{a}}_j\mathbf{1}_n^T+\mathbf{1}_n\tilde{\mathbf{b}}_j^T+\tilde{U}_j\tilde{V}_j^T\\
\tilde{U}_j&=\breve{U}_j-\mathbf{1}_n\boldsymbol{\mu}_{U_j}^T \qquad \tilde{V}_j=\breve{V}_j-\mathbf{1}_n\boldsymbol{\mu}_{V_j}^T\\
\tilde{\mathbf{a}}_j&=\breve{U}_j\boldsymbol{\mu}_{U_j} \qquad \tilde{\mathbf{b}}_j=\breve{V}_j\boldsymbol{\mu}_{V_j} \qquad \tilde{\mu}_j=\boldsymbol{\mu}_{U_j}^T\boldsymbol{\mu}_{V_j}
\end{aligned} \tag{2}$$

Here, $\hat{U}_j$ and $\hat{V}_j$ are $n\times k_j$ nonsingular matrices, and $\hat{\Gamma}_j$ is k_j order singular values diagonal matrix. $\breve{U}_j=\hat{U}_j\hat{\Gamma}_j{}^{1/2}$, $\breve{V}_j=\hat{V}_j\hat{\Gamma}_j{}^{1/2}$. n-dimensional vectors $\boldsymbol{\mu}_{U_j}$ and $\boldsymbol{\mu}_{V_j}$ are columns means of $\breve{U}_j$ and $\breve{V}_j$ respectively. Finally, latent factors of network G_j could be extracted by $\tilde{N}_j=[\tilde{\mathbf{a}}_j,\tilde{\mathbf{b}}_j,\tilde{U}_j,\tilde{V}_j]$. Since dimension of $\tilde{N}_j$ is $2k_j+2$, complex network cannot be reduced significantly unless $k_j<<n/2$. Meanwhile the accumulation contribution rate restriction $\sum_{l=1}^{k_j}\lambda_l^2/\sum_{l=1}^{n}\lambda_l^2\geq\theta$ is adopted to select small k_j, where λ_l is the l-th decreasing singular value and threshold value $0<\theta<1$.

After estimating low-dimensional factors $\mathbf{n}_{ix}=(a_{ix},b_{ix},\mathbf{u}_{jx}^T,\mathbf{v}_{jx}^T)^T$ of node x in networks G_j, latent factors are assumed to follow multivariate normal distribution

$$(\mathbf{n}_{ix},\mathbf{n}_{jx})\sim_{i.i.d.}\mathrm{N}_{2k_i+2k_j+4}\left(\mathbf{0},\Sigma_{N_iN_j}=\begin{pmatrix}\sum_{N_i} & \sum_{N_i,N_j}\\ \sum_{N_j,N_i} & \sum_{N_j}\end{pmatrix}\right) \tag{3}$$

Herein, $\mathbf{0}$ is $(2k_i+2k_j+4)$-dimensional zero mean vector; $\Sigma_{N_iN_j}$ denotes covariance matrix. Then we substitute networks' correlation test to factors' correlation test,

$$H_0:\sum\nolimits_{N_i,N_j}=0 \qquad H_1:\sum\nolimits_{N_i,N_j}\neq 0 \tag{4}$$

our likelihood ratio test statistic is formulated as follows

$$\begin{aligned}
\Lambda&=\frac{\max_{\sum_{N_iN_j}}L(\sum_{N_iN_j}|N_i,N_j)}{\max_{\sum_{N_i},\sum N_j}L_0(\sum_{N_i},\sum_{N_j}|N_i,N_j)}\\
&=\prod_{l=1}^{(2k_i+2)\wedge(2k_j+2)}(1-r_l^2)^{-n/2}
\end{aligned} \tag{5}$$

Here, L_0 and L stands for the likelihood function under H_0 and whole parameter space respectively. r_l is the l-th ordered eigenvalue of

$$(N_i^TN_i)^{-1/2}(N_i^TN_j)(N_j^TN_j)^{-1}(N_j^TN_i)(N_i^TN_i)^{-1/2}.$$

Under null hypothesis, $W=\Lambda^{-2/n}$ follows Wilks' Lambda distribution, which could be characterized as product of beta random variables

$$\begin{aligned}
W=\Lambda^{-2/n}&\sim U(2k_j+2,2k_i+2,n-(2k_i+2))\\
&=\prod_{l=1}^{2k_j+2}Beta(\frac{n-2k_i-2k_j-4+l}{2},\frac{2k_i+2}{2})
\end{aligned} \tag{6}$$

The rejected region of null hypothesis contains W smaller than α-quantiles of sampled distribution which could be calculated using Monte Carlo simulation. Specifically, supposing network G_i and G_j are both symmetrical, merely half factors $N_i=[\mathbf{a}_i,U_i]$ and $N_j=[\mathbf{a}_j,U_j]$, need to be tested.

3. Cold start link prediction in multi-relational networks

For given auxiliary sub-networks $G_{s1},G_{s2},\ldots,G_{sm}$ in multi-relational network G, provided their correlation relation with target sub-network G_1 has been tested, regression models could be established for cold start link prediction directly. It's worth noting that, auxiliary sub-networks may supply more interactions information for additional nodes than target G_1. For instance, cost of a new drug's development usually exceeds over 800 million dollars [16] requiring plenty of downstream experimental validations, such as drugs' adverse and synergistic interactions research. However, drug chemical structure similarities knowledge are more convenient and helpful to predict functional interaction intuitively.

The latent factors of above m auxiliary and target sub-network are denoted by explanatory $[N_{s1},N_{s2},\ldots,N_{sm}]$ and response N_1 respectively. The explanatory variables are abbreviated to X. For the sake of potential multicollinearity of X, *SVD* is utilized again to extract low-dimensional factors N_X for dimension reduction. Then, we construct multivariate regression model as follows

$$f(N_X)=N_1 \tag{7}$$

Model $f(.)$ is quite flexible to linear, nonlinear, even nonparametric structure, etc. After modeling multi-relationship among old nodes, potential interactions associated with new nodes in the target sub-network could be quantified based on deducible interactions between new and old nodes. Similarly, We denote the latent factors