ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1730

Link Prediction by Incidence Matrix Factorization

Sho Yokoi[1] and **Hiroshi Kajino**[2] and **Hisashi Kashima**[3]

Abstract. Link prediction suffers from the data sparsity problem. This paper presents and validates our hypothesis that, for sparse networks, incidence matrix factorization (IMF) could perform better than adjacency matrix factorization (AMF), which has been used in many previous studies. A key observation supporting the hypothesis is that IMF models a partially-observed graph more accurately than AMF. A technical challenge for validating our hypothesis is that, unlike AMF approach, there does not exist an obvious method to make predictions using a factorized incidence matrix. To this end, we newly develop an optimization-based link prediction method adopting IMF. We have conducted thorough experiments using synthetic and real-world datasets to investigate the relationship between the sparsity of a network and the performance of the aforementioned two methods. The experimental results show that IMF performs better than AMF as networks become sparser, which strongly validates our hypothesis.

1 Introduction

Link prediction attempts to predict missing links based on other observed links and attributes of nodes [6, 2, 10, 13]. We focus on link prediction based on a graph structure, which is formulated as follows: given a partially-observed graph $G=(\mathcal{V},\mathcal{E}_\mathrm{P})$ with the set of nodes $\mathcal{V}$ and the set of *positive links* (observed links) $\mathcal{E}_\mathrm{P}\subset\mathcal{V}\times\mathcal{V}$, its goal is to learn a scoring function $s:\mathcal{V}\times\mathcal{V}\to\mathbb{R}$ to predict a new link on an *unlabeled pair of nodes* in a set $\mathcal{E}_\mathrm{U}:=(\mathcal{V}\times\mathcal{V})\setminus\mathcal{E}_\mathrm{P}$.

As pointed out by many researchers, one of the central issues in link prediction is the *sparsity* of positive links [5, 12, 9]. Our idea to counter the problem is to employ incidence matrix factorization (**IMF**) as a building block of a link prediction method, instead of adjacency matrix factorization (**AMF**), which has been used in various previous studies [8, 1, 7, 11, 4]. A key observation supporting the idea is that IMF can model a partially-observed graph more accurately than AMF.

First of all, We briefly introduce the previous AMF-based approach (Fig. 1 [TOP]). Given a partially-observed graph $G=(\mathcal{V},\mathcal{E}_\mathrm{P})$, AMF learns latent feature vectors $\{\boldsymbol{x}_k\}_{v_k\in\mathcal{V}}$ of nodes using both positive links and unlabeled node pairs such that $\langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle\approx 1$ if $(v_i,v_j)\in\mathcal{E}_\mathrm{P}$, 0 otherwise holds in its simplest instantiation. This modeling has a little flaw. Let us consider a pair of nodes (v_i,v_j) that is not linked in a partially-observed graph but is actually positive in its fully-observed graph. In the ideal case, latent vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ obtained from the fully-observed graph satisfy $\langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle\approx 1$, while those obtained from the partially-observed graph satisfy $\langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle\approx 0$. As the observed part of a graph becomes sparser, the number of such node pairs will increase, and therefore, this inconsistency issue can lead to the poor performance. On contrary, IMF can avoid the inconsistency issue because it learns a model by utilizing only positive links. IMF learns latent feature vectors of nodes $\{\boldsymbol{x}_k\}_{v_k\in\mathcal{V}}$ and those of positive links $\{\boldsymbol{y}_l\}_{e_l\in\mathcal{E}_\mathrm{P}}$ such that $\langle\boldsymbol{x}_i,\boldsymbol{y}_j\rangle\approx 1$ if $v_i\in e_j$, 0 otherwise holds in its simplest instantiation. Since this modeling does not utilize unlabeled node pairs, the model obtained from a partially-observed graph is consistent with that obtained from its fully-observed one; therefore, the performance of IMF is expected to be robust to the sparsity of a graph. In this light, we arrive at the hypothesis that IMF can counter the sparsity problem better than AMF.

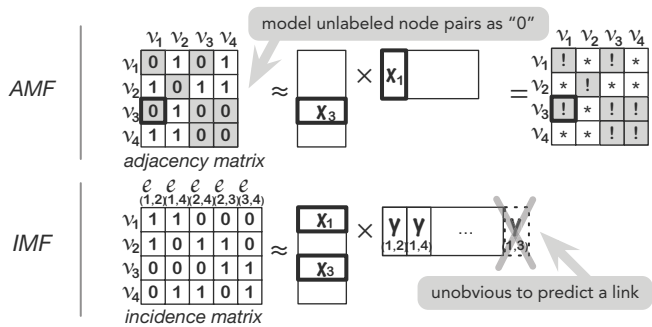

Figure 1. [TOP] Link prediction using AMF. [BOTTOM] It is not trivial to predict a link using IMF. We learn a latent vector $\boldsymbol{y}_{(i,j)}$ for any observed link $e_{(i,j)}\in\mathcal{E}_\mathrm{P}$ in addition to a latent vector $\boldsymbol{x}_k$ for any node. Predicting a link between v_1 and v_3 requires a latent vector of the unlabeled pair of nodes $(v_1,v_3)\notin\mathcal{E}_\mathrm{P}$, which we cannot obtain through IMF.

While the IMF approach is promising, it is not trivial to predict a new link using a factorized incidence matrix (Fig. 1 [BOTTOM]). The main purpose of this paper is (i) to develop a new link prediction method based on IMF and (ii) to confirm the hypothesis by thorough experiments with synthetic and real-world datasets.

2 IMF-based Link Prediction

In this section, we newly propose an optimization-based efficient link prediction method adopting the IMF approach.

Algorithm. Figure 2 illustrates the overview of our method. Given an incidence matrix $B\in\mathbb{R}^{|\mathcal{V}|\times|\mathcal{E}_\mathrm{P}|}$ of the graph $G=(\mathcal{V},\mathcal{E}_\mathrm{P})$, IMF first factorizes B into two matrices X and Y using truncated SVD:

$$B\approx U_k\Sigma_k{V_k}^\top=XY^\top, \tag{1}$$

where $X:=U_k\Sigma_k$ and $Y:=V_k$. It provides us latent vectors of nodes X and those of positive links Y. Here, for any positive link (v_i,v_j), $\boldsymbol{b}_{(i,j)}\approx X\boldsymbol{y}_{(i,j)}$ holds, where $\boldsymbol{b}_{(i,j)}:=(0,\dots,0,\underset{i}{1},0,\dots,0,\underset{j}{1},0,\dots,0)^\top$ is a column vector of B. Our idea is to predict a link on an unlabeled node pair (v_i',v_j') by how well we can recover its latent vector $\boldsymbol{y}_{(i',j')}$ that is consistent with the factorization, i.e., $\boldsymbol{b}_{(i',j')}\approx X\boldsymbol{y}_{(i',j')}$. This

[1] Tohoku University, Japan, email: yokoi@ecei.tohoku.ac.jp
[2] IBM Research - Tokyo, Japan, email: KAJINO@jp.ibm.com
[3] Kyoto University, Japan, email: kashima@i.kyoto-u.ac.jp

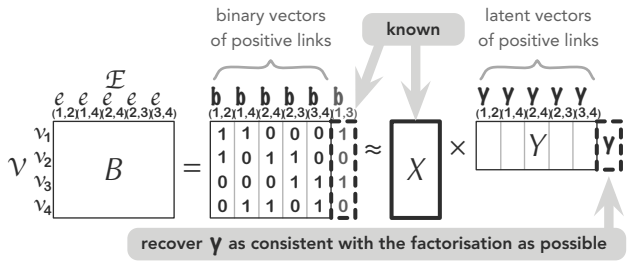

Figure 2. Overview of our link prediction method based on IMF.

idea boils down to the following scoring function, and this optimization problem can be solved in a closed form:

$$s_{\mathrm{IMF}}(v_i,v_j) := -\min_{\boldsymbol{y}\in\mathbb{R}^k} \left\|\boldsymbol{b}_{(i,j)} - X\boldsymbol{y}\right\|_2^2 = -\|\boldsymbol{w}_i+\boldsymbol{w}_j\|_2^2, \quad (2)$$

$$(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_{|\mathcal{V}|}) := X(X^\top X)^{-1}X^\top - I_{|\mathcal{V}|} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}. \quad (3)$$

Computational Efficiency. At first sight, the computational cost of the IMF-based method seems more expensive than the AMF-based method because the size of an incidence matrix is larger than that of an adjacency matrix generally. However, with a simple contrivance, the cost of the matrix factorization of our method can be as small as that of the AMF-based method. Observing that we only need the matrices U_k and Σ_k in the matrix factorization (Eq. (1)), it is sufficient to factorize the positive semi-definite symmetric matrix $BB^\top$ into $Q_k\Lambda_k{Q_k}^\top$ by truncated SVD to obtain $U_k = Q_k$ and $\Sigma_k = \Lambda_k^{1/2}$. Since the size of $BB^\top$ is the same as the adjacency matrix A, the computation time of matrix factorization of our method is the same as that of AMF.

Moreover, the construction of the matrix $BB^\top$ requires almost the same computation time as that of the adjacency matrix A because $BB^\top = A + D$ holds, where D denotes $\mathrm{diag}(d_1,\ldots,d_{|\mathcal{V}|})$, and each d_i corresponds to the degree of a node v_i.

3 Experiments

To demonstrate that IMF actually counters the sparsity problem better than AMF, we conducted comparative experiments with synthetic and real-world datasets.

Datasets. In the first experiment, we generated 10 synthetic graphs ($|\mathcal{V}| = 10^4, |\mathcal{E}_\mathrm{P}| \simeq 10^4,\ldots,10^6$) by the Barabási–Albert model [3], which possess scale-free and small-world properties. In the second experiment, we extracted all the unweighted and undirected real-world graphs from KONECT[4] and chose the 24 smallest graphs in terms of the size $|\mathcal{V}|$ ($|\mathcal{V}| \simeq 10^1,\ldots,10^4, |\mathcal{E}_\mathrm{P}| \simeq 10^1,\ldots,10^6$).

Performance Measure. We used ROC–AUC to evaluate the performance of the scoring function, which is known to be a proper performance measure in link prediction [9].

Experimental Procedure. We conducted five-fold cross validation to measure the performance of IMF and AMF by repeating the following process, and then reported the mean of AUC. First, given $G = (\mathcal{V},\mathcal{E}_\mathrm{P})$, we randomly divide $\mathcal{E}_\mathrm{P}$ into $\mathcal{E}_\mathrm{P}^{(\mathrm{train})}$, $\mathcal{E}_\mathrm{P}^{(\mathrm{dev})}$, and $\mathcal{E}_\mathrm{P}^{(\mathrm{test})}$ by a ratio of $3:1:1$. Second, with $\mathcal{E}_\mathrm{P}^{(\mathrm{train})}$, we learn a scoring function s_k for each $k \in \{2^0, 2^1, \ldots, \min\{2^{14}, 2^{\lfloor \log_2(\mathrm{rank}\, M)\rfloor}\}\}$, where k is the rank of truncated SVD, and M is the incidence or adjacency matrix. Then we select the best hyperparameter k in terms of AUC of s_k. Third, with $\mathcal{E}_\mathrm{P}^{(\mathrm{test})}$, we calculate AUC of $s_{\mathrm{best}k}$ as results.

[4] http://konect.uni-koblenz.de/

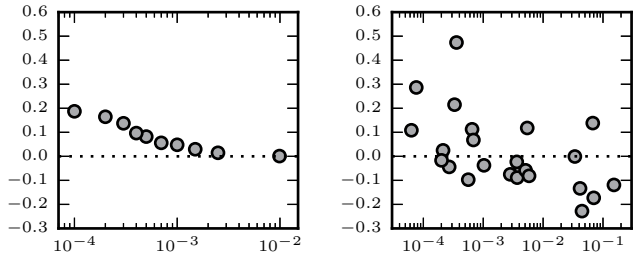

Figure 3. Scatter plot illustrating the relation between the sparsity (x-axis, $|\mathcal{E}_\mathrm{P}|/|\mathcal{V}|^2$) and the performance improvement of IMF over AMF (y-axis, $\mathrm{AUC}_\mathrm{IMF} - \mathrm{AUC}_\mathrm{AMF}$). Each point corresponds to each graph. [LEFT] Synthetic datasets. [RIGHT] Real-world datasets.

Experimental Results. Figure 3 [LEFT] shows the experimental results on the synthetic datasets. The Spearman's ρ between the sparsity measure and the AUC improvement of IMF over AMF is $-1.0 < 0$ ($p = 0.0 < 0.01$); i.e., the performance gain of IMF over AMF increases as the original graph becomes sparser, and the hypothesis is strongly supported. Furthermore, the AUCs of IMF on all the synthetic graphs are nearly constant (0.70), while that of AMF becomes worse as the graph becomes sparser ($0.72,\ldots,0.50$). It implies that the IMF approach is potentially capable of capturing scale-free or small-world properties of networks.

Figure 3 [RIGHT] shows the experimental results on the real-world datasets. Similar to the former experiments, Spearman's $\rho = -0.55 < 0$ ($p = 0.0054 < 0.01$), which also supports our hypothesis.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 15H01704.

REFERENCES

[1] E. Acar, D. M. Dunlavy, and T. G. Kolda, 'Link Prediction on Evolving Data Using Matrix and Tensor Factorizations', in *LDMTA at ICDM*, pp. 262–269, (2009).
[2] M. Al Hasan and M. J. Zaki, 'A Survey of Link Prediction in Social Networks', in *Social Network Data Analytics*, 243–275, (2011).
[3] A. L. Barabási and R. Albert, 'Emergence of Scaling in Random Networks', *Science*, **286**(5439), 509–512, (1999).
[4] E. Dong, J. Li, and Z. Xie, 'Link Prediction via Convex Nonnegative Matrix Factorization on Multiscale Blocks', *J. Appl. Math.*, **2014**, 786156:1–786156:9, (2014).
[5] L. Getoor, 'Link Mining: A New Data Mining Challenge', *SIGKDD Explor.*, **5**(1), 84–89, (2003).
[6] L. Getoor and C. P. Diehl, 'Link Mining: A Survey', *SIGKDD Explor.*, **7**(2), 3–12, (2005).
[7] J. Kunegis and A. Lommatzsch, 'Learning Spectral Graph Transformations for Link Prediction', in *ICML*, pp. 561–568, (2009).
[8] D. Liben-Nowell and J. Kleinberg, 'The Link-Prediction Problem for Social Networks', *J. Am. Soc. Inf. Sci. Technol.*, **58**(7), 1019–1031, (2007).
[9] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, 'New Perspectives and Methods in Link Prediction', in *KDD*, pp. 243–252, (2010).
[10] L. Lü and T. Zhou, 'Link prediction in complex networks: A survey', *Physica A*, **390**(6), 1150–1170, (2011).
[11] A. K. Menon and C. Elkan, 'Link Prediction via Matrix Factorization', in *ECML PKDD*, pp. 437–452, (2011).
[12] M. J. Rattigan and D. Jensen, 'The Case For Anomalous Link Discovery', *SIGKDD Explor.*, **7**(2), 41–47, (2005).
[13] P. Wang, B. Xu, Y. Wu, and X. Zhou, 'Link prediction in social networks: the state-of-the-art', *Sci. China Inform. Sci.*, **58**(1), 1–38, (2014).