Link Prediction via Latent Factor BlockModel

Sheng Gao, Ludovic Denoyer, Patrick Gallinari
LIP6 - University Pierre et Marie Curie
4 place Jussieu, 75005 Paris, France
{sheng.gao, ludovic.denoyer, patrick.gallinari }@lip6.fr

ABSTRACT

In this paper we address the problem of link prediction in networked data, which appears in many applications such as social network analysis or recommender systems. Previous studies either consider latent feature based models but disregarding local structure in the network, or focus exclusively on capturing local structure of objects based on latent blockmodels without coupling with latent characteristics of objects. To combine the benefits of previous work, we propose a novel model that can incorporate the effects of latent features of objects and local structure in the network simultaneously. To achieve this, we model the relation graph as a function of both latent feature factors and latent cluster memberships of objects to collectively discover globally predictive intrinsic properties of objects and capture latent block structure in the network to improve prediction performance. Extensive experiments on several real world datasets suggest that our proposed model outperforms the other state of the art approaches for link prediction.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications—*Data Mining*; J.4 [**Social and Behavioral Sciences**]: Sociology

General Terms

Algorithms, Experimentation

Keywords

Latent feature model, latent blockmodel, link prediction

1. INTRODUCTION

Networked data has become ubiquitous, which consists of interrelated objects linked with multiple relation types, thus *link prediction* has arisen as the fundamental task in many applications (e.g. social network analysis, recommender systems and bioinformatics), which involves to predict the presence or absence of links between objects of a network. However, the complexity of network structure makes this a great challenging problem: the correlations among objects give rise to complex structural patterns (e.g. the local structure in the network is either dense or sparse), which deviates the classical dense cluster assumption, i.e., the strongly correlated data always forms dense clusters. Moreover, networked data is quite sparse and scalable since each relation graph generated from the data involves a large number of objects with each being connected to only a tiny proportion of the whole graph, which calls for the corresponding models capable of learning from rare, noisy and largely missing observations.

Recently, the latent feature based models have been successfully studied for link prediction task [2] [4], which consider link prediction as a matrix completion problem and employ latent matrix factorization framework to learn latent feature factors for each object, and make predictions by taking appropriate inner products. However, these models focus more on globally qualitative analysis and disregard the local structure in the network. In contrast, there also exist structure based models [1], sometimes referred to as latent blockmodels, which assign each object to a cluster or block with respect to the local structure, and then use the latent cluster memberships to solely predict the link structure. Since this kind of approaches are largely Bayesian, one important issue is the computational efficiency and scalability.

In this paper, we combine the benefits of previous work and propose a statistical model that extracts latent local structure representing the topological properties of individual objects from the observed data, and simultaneously incorporates the effect of latent features of objects based on latent matrix factorization framework. To achieve this, we model the relation variables as a function of latent feature factors and latent block structure to improve prediction performance.

Problem Statement: Suppose we have a set of objects $\{x_1, ..., x_n\}$. Observations consisting of links are represented by the relation matrix $\mathbf{S} = \{S_{ij} \in \{0, 1, ?\}, i, j = 1, ..., N\}$, where 1 denotes there is an observed present link, 0 denotes the absent link, and ? denotes the missing link. We then use the binary indicator matrix $\mathbf{W}$ to indicate whether or not the link is observed. We use $z_i \in \{1, ..., K\}$, where K is the number of latent clusters, to denote the latent cluster assignment of object x_i. We furthermore introduce $z_{ik} = [z_i = k]$ to indicate that x_i is in the kth cluster when $z_{ik} = 1$ and 0 otherwise. Latent cluster assignments matrix $\mathbf{Z} = \{z_{ik} : i \in 1, ..., N, k \in 1, ..., K\}$ includes the latent cluster memberships of all the objects in the network. Given such a relation graph, our goal is to predict the missing links between the unobserved pairs in the network.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

2. METHODOLOGY

We consider modeling the observed data based on latent feature model. We first assume the elements of the relation matrix S_{ij} as Bernoulli-distributed variables, which are conditionally independent given the latent parameter H_{ij} through the logistic function $\sigma(H) = \frac{1}{1+e^{-H}}$. Thus, the conditional distribution over the observations in the relation matrix $\mathbf{S}$ can be defined as follows:

$$p(\mathbf{S}|\mathbf{H}) = \prod_{i,j} [\sigma(H)_{ij}^{S_{ij}} (1 - \sigma(H)_{ij})^{1-S_{ij}}]^{W_{ij}} \quad (1)$$

To characterize the latent parameter matrix $\mathbf{H}$ in the framework of latent feature model, we consider that there are *latent feature factors* $\mathbf{U} \in \mathbb{R}_+^{N\times K}$ and $\mathbf{V} \in \mathbb{R}_+^{N\times K}$ for object pair x_i and x_j within a directed relation, where K is the dimension of latent feature factor, that can be used for encoding the observable attributes (e.g. a user's profile) or latent semantic topics (e.g. a movie's genre). Then based on the latent cluster membership z_i, we introduce a latent block matrix $\mathbf{C} \in \mathbb{R}_+^{K\times K}$ to explicitly capture the latent local structure, where C_{kk} denotes the probability of a link existing between objects within the same kth cluster, and C_{kl} denotes the probability of one object in kth cluster linking to the other within the lth cluster. Thus the latent parameter H_{ij} can be defined as follows:

$$H_{ij} = \mathbf{u}_i \mathbf{v}_j^T + \mathbf{z}_i \mathbf{C} \mathbf{z}_j^T + \epsilon \quad (2)$$

Here, $\mathbf{u}_i \in \mathbb{R}^K$ and $\mathbf{v}_i \in \mathbb{R}^K$ are row vectors from $\mathbf{U}$ and $\mathbf{V}$, the inner product term of $\mathbf{u}_i \mathbf{v}_j^T$ provides the probabilities of a link between the two objects based on their latent feature factors. $\mathbf{z}_i \in \mathbb{R}^K$ is considered as the latent cluster indicator *vector* for each object, which implies the object x_i associates with the kth cluster. Actually, the form of $\mathbf{z}_i \mathbf{C} \mathbf{z}_j^T$ provides a general model to discover various latent local structure in the relation graph, i.e., there may exist latent dense or sparse clusters among objects. More specifically, we can use the latent block matrix $\mathbf{C}$ to represent various types of dense or sparse cluster structures. ϵ denotes the sparsity of the relations in the network, which can also be considered as a bias term.

We thus far model the observed interactions by combining the benefits of latent feature factors of objects with their corresponding latent cluster assignments as well as the latent block structure, hence the integration of these multiple effects makes our proposed model better generalization in link prediction and more interpretable for network structure, which can be referred to as Latent Factor BlockModel (LFBM).

To make our proposed model more accurate, we can impose some prior distributions on the latent factors as in Bayesian learning. For example, the latent cluster indicator $\mathbf{z}_i$ for each object can be generated based on multinomial distribution. Gaussian priors can be put on the latent feature factors $\mathbf{U}, \mathbf{V}$ and the block matrix $\mathbf{C}$. Then for learning the latent factors and other model parameters from the observed data, we develop an optimization transfer algorithm based on the Generalized Expectation-Maximization (EM) method to alleviate the model complexity in optimization.

3. EXPERIMENTAL RESULTS

We evaluate our proposed model on two social network datasets compared to several methods: NMF model [5], MMSB model [1], MLFM model [2] as well as GLFM model [3]. We use LiveJournal and ArXiv coauthorship datasets for the link prediction experiments. LiveJournal consists of 3,773 users and 209,832 social links while ArXiv contains 2,403 authors and 21,397 coauthorship. We randomly choose 80% of the relations for training, and 20% as missing for test. We evaluate all the models by AUC (Area Under the ROC curve) values averaged over five times. We set the number of latent clusters to $k = 20$ and the dimension of latent feature factors to $d = 20$ for both datasets. The latent factors $\mathbf{U}, \mathbf{V}$ and $\mathbf{C}$ are initialized from Gaussian priors with zero-mean and fixed variance as 2 in the experiments. Experimental results are shown in Table 1. We find that our proposed LFBM model outperforms all the other models in both datasets, which suggests that integrating the effects of latent feature factors and latent cluster information among objects can lead to better performance compared to the models that learn the effect of latent characteristics separately. Taking LiveJournal as an example, comparing to latent blockmodel (MMSB), LFBM gains much higher improvement, which proves the excellent predictive performance of latent feature based methods on link prediction task. While with respect to the latent feature only based models (i.e. NMF, MLFM, GLFM), LFBM improves the performance about 10% $\sim$ 20%, indicating that exploiting latent local structure is quite effective to achieve better prediction performance.

Table 1: AUC performances on two datasets using different models in link prediction task.

	NMF	MMSB	MLFM	GLFM	LFBM
LiveJournal	0.7468	0.6512	0.8023	0.8319	**0.8720**
ArXiv	0.6801	0.6099	0.7345	0.7676	**0.8029**

4. CONCLUSION

In this paper, we have addressed the problem of link prediction in networked data. For that we proposed a novel model that simultaneously incorporated the effect of latent feature factors and the impact from the latent block structure in the network. The model can collectively capture globally predictive intrinsic properties of objects and discover the latent block structure, which shows the success of the coupled benefits of latent feature based approaches and latent blockmodels. We prove the efficacy of our proposed model through the link prediction task on several real world data sets.

5. REFERENCES

[1] M. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic block models. *JMLR*, pages 1981–2014, 2008.

[2] P. Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational & Mathematical Organization Theory*, pages 261–272, 2009.

[3] W.-J. Li, D.-Y. Yeung, and Z. Zhang. Generalized latent factor models for social network analysis. *IJCAI*, pages 1705–1710, 2011.

[4] A. Menon and C. Elkan. Link prediction via matrix factorization. *ECML-PKDD*, 2011.

[5] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. *SIGIR '03*, pages 267–273, 2003.