

Link Prediction in Metabolic Networks using Topology-based Mixture Models

Akira Ninagawa¹

akiraninagawa@cs25.scitec.kobe-u.ac.jp

Koji Eguchi¹

eguchi@port.kobe-u.ac.jp

¹ Department of Computer Science and Systems Engineering, Kobe University, 1-1 Rokkodai, Nada-ku, Kobe 657-8501, Japan

Abstract

This paper focuses on the task of link prediction in networks using a hierarchical Bayesian model, which assumes unobservable prior distributions over a mixture model based on network topology. We demonstrate that our method is remarkably effective for this task in a biological metabolic network.

Keywords: link prediction, mixture models, network topology, metabolic networks

1 Introduction

Modeling of complex networks is a crucial task in a wide variety of fields, such as in biology and social sciences. A large number of researches have been conducted; however, most of them required explicit, specific prior knowledge on target networks. On the other hand, a few recent works on probabilistic mixture models do not require such explicit prior knowledge and turned out to be effective for the task of group detection of vertices such as in social networks [3]. This paper focuses on the task of link prediction in biological metabolic networks using a hierarchical Bayesian model, which assumes unobservable prior distributions over a mixture model based on network topology and is estimated using Bayesian inference via Gibbs sampling. This is the first work investigating the task of link prediction using the topology-based mixture modeling approach, to our knowledge. We demonstrate that, using this model, link prediction performance was significantly improved compared to three conventional methods through experiments using a real metabolic network.

2 Methods and Results

We start with a network G that consists of a set of vertices or entities $\mathbf{v} = \{v_i\}$ ($i = 1, \dots, N$) and a set of edges or links $\mathbf{E} = \{\mathbf{e}_i\}$ ($i = 1, \dots, N$), in which $\mathbf{e}_i = \{e_{ij}\}$ ($j = 1, \dots, M_i$) indicates a set of all edges from vertex v_i to others. \mathbf{E} is equivalent to the adjacency matrix of the network. We assume that network G is comprised of a set of underlying groups $\mathbf{g} = \{g_k\}$ ($k = 1, \dots, K$), each of which group is defined as a distribution over vertices. Let z_{ij} to be the group assigned to vertex v_i 's adjacent vertex v_j . Therefore, $z_{ij} = g_k$ represents that group g_k is assigned to vertex v_j adjacent from vertex v_i . Moreover, $\mathbf{Z} = \{\mathbf{z}_i\}$ ($i = 1, \dots, N$) can be defined where $\mathbf{z}_i = \{z_{ij}\}$ ($j = 1, \dots, M_i$). We then consider a probabilistic mixture model, where each vertex is represented as a mixture of the groups. $P(\mathbf{z}_i|\theta_i)$ indicates per-vertex mixture distribution over groups; in other words, the probability of repeatedly sampling a group that an arbitrary vertex adjacent from vertex v_i belongs to. Moreover, $P(\mathbf{E}|\mathbf{Z}, \phi_k)$ indicates per-group multinomial distribution over edges; in other words, the probability of repeatedly sampling an edge having a vertex that belongs to group g_k . Parameters θ_i and ϕ_k are sampled from Dirichlet distributions specified by given hyperparameters α and β , respectively. We denote $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ as the entire sets $\{\theta_i\}$ ($i = 1, \dots, N$) and $\{\phi_k\}$ ($k = 1, \dots, K$), respectively. The probabilistic mixture model above is a simple hierarchical Bayesian model in the sense that parameters θ_i and ϕ_k are sampled from the respective conjugate prior distributions. In contrast, a mixture model used by Newman et al. [3] does not use prior distributions and thus robust, accurate estimation of model parameters is hard to achieve. The hierarchical Bayesian model above is a ‘‘generative’’ model of network, and the process of generating a network is formalized as follows:

- (1) For all v_i vertices sample $\theta_i \sim \text{Dirichlet}(\alpha)$
- (2) For all g_k groups sample $\phi_k \sim \text{Dirichlet}(\beta)$
- (3) For each of the M_i vertices v_j adjacent from vertex v_i :
 - (a) Sample a group $z_{ij} \sim \text{Multinomial}(\theta_i)$
 - (b) Sample a vertex $v_j \sim \text{Multinomial}(\phi_{z_{ij}})$

Table 1: Evaluation results on link prediction task.

	MAP (%)	MP@10 (%)		MAP (%)	MP@10 (%)
Adamic/Adar	0.03014	0.7273	proposed ($K = 80$)	21.44	43.40
Jaccard	0.00236	0.1818	proposed ($K = 90$)	21.25	35.60
Katz	3.587	9.636	proposed ($K = 100$)	22.05	42.40

where (v_i, v_j) corresponds to an edge e_{ij} . Given hyperparameters α and β , the full joint distribution over all variables and parameters is: $p(\mathbf{E}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\phi} | \alpha, \beta) = p(\boldsymbol{\phi} | \beta) \prod_{i=1}^N p(\theta_i | \alpha) P(\mathbf{z}_i | \theta_i) P(\mathbf{e}_i | \mathbf{z}_i, \boldsymbol{\phi})$. This can be transformed into the following equation [1]:

$$p(\mathbf{E}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\phi} | \alpha, \beta) = \prod_{i=1}^N \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^K \theta_{ik}^{\alpha-1+n_{i,k}} \times \prod_{k=1}^K \frac{\Gamma(N\beta)}{\Gamma(\beta)^N} \prod_{j=1}^N \phi_{kj}^{\beta-1+n_{j,k}}$$

where n_{ijk} indicates the count that group g_k is assigned to vertex v_i 's adjacent vertex v_j , and ‘ \cdot ’ means a corresponding index is marginalized. In other words, $n_{jk} = \sum_i n_{ijk}$ and $n_{i,k} = \sum_j n_{ijk}$. N and K indicate the number of vertices and the number of underlying groups in a target network, respectively. We used Gibbs sampling to estimate unknown parameters of the model [1].

Next, we describe experiments on the task of link prediction. We introduce three existing methods to compare with the proposed method. Those methods are well accepted and well investigated [2]. The first existing method we used is Jaccard coefficient: $(|\mathbf{a}_i| \cap |\mathbf{a}_j|) / (|\mathbf{a}_i| \cup |\mathbf{a}_j|)$, where \mathbf{a}_i denotes a set of adjacent vertices of vertex v_i . The second is Adamic/Adar: $\sum_{k \in |\mathbf{a}_i| \cap |\mathbf{a}_j|} 1 / \log |\mathbf{a}_k|$. The last is $Katz_{\mu}$: $\sum_{\ell=1}^{\infty} \mu^{\ell} |paths_{ij}^{(\ell)}|$.

The notation $paths_{ij}^{(\ell)}$ indicates the number of paths from vertex v_i to vertex v_j of which length is ℓ . The network data used in our experiments is the metabolic pathway of an organism of ‘‘S.Cerevisiae’’ and was constructed by Yamanishi et al. [4]. The number of vertices is 668, the number of edges is 2782, and the proportion of the edges to all vertex pairs is 0.0125. We used 80% of all the vertex pairs as training data, 10% as development data and the remainder as test data. It is necessary for our method to determine some parameters that will be described later, and thus the development data was used to determine the optimal parameters. We conducted experiments on the task of link prediction using 50 sets of training data, development data and test data that were randomly sampled in the same proportion from the entire set of vertices, and using each of the data sets, we compared the proposed method with the three existing methods. We determined the following four parameters: hyperparameters α and β of Dirichlet prior distributions, the number of latent groups K , and the number of iterations for Gibbs sampling, so that development-set log-likelihood is maximized.

We used mean average precision (MAP) as an evaluation metric of the task of link prediction. MAP is defined as $(1/|\mathbf{data}|) \sum_{d \in \mathbf{data}} \left\{ (1/|\mathbf{true}_d|) \sum_{r \in \mathbf{rank}_d} prec(r) \right\}$, where ‘‘ \mathbf{data} ’’ denotes a series of test data ($|\mathbf{data}| = 50$), ‘‘ \mathbf{true}_d ’’ indicates the entire set of ‘‘true’’ links in test data d (i.e., all appeared links in d), and ‘‘ \mathbf{rank}_d ’’ indicates the entire set of links predicted by a method using the training data and the development data corresponding to d . The notation $prec(r)$ indicates precision at rank r in the ranking of predicted links, where precision is defined as the proportion of predicted true links out of r top-ranked predicted links. Here, the link prediction ranking is achieved according to test-set log-likelihood in the case of our method, or according to a similarity measure in the case of the other existing methods.

3 Discussions

According to Table 1, the link prediction performance of our method is more than 18% higher than that of the other three methods, in terms of MAP. Its percentage improvement (i.e., the ratio of the increase to a baseline) was 515%, compared with Katz measure. The improvement obtained by our method was statistically significant over either Katz, Adamic/Adar or Jaccard coefficients, where $p < 0.01$ with the two-sided Wilcoxon signed-rank test. According to the other evaluation metric, mean of precision at the 10th rank, our methods remarkably outperformed the baselines, as well.

References

- [1] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc. of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.
- [2] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proc. of the 12th ACM International Conference on Information and Knowledge Management*, pages 556–559, New Orleans, Louisiana, USA, Nov. 2003.
- [3] M.E.J.Newman and E.A.Leicht. Mixture models and exploratory analysis in networks. *Proc. of the National Academy of Sciences of the United States of America*, 104(23):9564–9569, 2007.
- [4] Y. Yamanishi, J.-P. Vert, and M. Kanehisa. Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics*, 21(1):468–477, 2005.