Contents lists available at ScienceDirect

Physica A

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

Q1 A class-based link prediction using Distance Dependent Chinese Restaurant Process

Q2 Azam Andalib, Seyed Morteza Babamir*

Department of Computer, University of Kashan, Iran

HIGHLIGHTS

- In this paper, we proposed a novel method for link prediction enabling the exploitation of topological structure information of networks. We also proposed a new Gibbs sampling algorithm for computing the posterior distribution of the hidden variables based on the training data.
- We developed a novel MCMC algorithm for inferring the hidden variables and parameters of the proposed model.
- Empirical results on several benchmark real world datasets showed our proposed model has more performance than other state of the art probabilistic relational models.

ARTICLE INFO

Article history:
Received 29 January 2016
Available online xxxx

Keywords:
Chinese Restaurant Process
Dirichlet Process
Distance Dependent Chinese Restaurant Process
Link prediction

ABSTRACT

One of the important tasks in relational data analysis is link prediction which has been successfully applied on many applications such as bioinformatics, information retrieval, etc. The link prediction is defined as predicting the existence or absence of edges between nodes of a network. In this paper, we propose a novel method for link prediction based on Distance Dependent Chinese Restaurant Process (DDCRP) model which enables us to utilize the information of the topological structure of the network such as shortest path and connectivity of the nodes. We also propose a new Gibbs sampling algorithm for computing the posterior distribution of the hidden variables based on the training data. Experimental results on three real-world datasets show the superiority of the proposed method over other probabilistic models for link prediction problem.

1. Introduction

An efficient way to model the relations between the people in a group or community is through networks. Recently, there has been a great interest to use a graph structure for visualizing such relations and interactions. In such graphs, a vertex (node) represents a person in some community and an edge demonstrates some association between the corresponding people. Social networks are useful patterns for efficient modeling the relations between community groups and people. To analyze such networks, relations and interactions between individuals are visualized using the graph structure where vertexes (nodes) and edges represent community people and association between individuals respectively.

In recent years, because of the growth of the availability and popularity of the social networks, a great number of ideas have been developed to analyze them. Among others, the link prediction is used to predict the association between two

* Corresponding author.
E-mail addresses: azam_andalib@kashanu.ac.ir (A. Andalib), babamir@kashanu.ac.ir (S.M. Babamir).

http://dx.doi.org/10.1016/j.physa.2016.03.067
0378-4371/© 2016 Published by Elsevier B.V.

individuals (nodes). To this end, we exploit the observation of the relationships between individuals in a network and try to predict latent (unobserved) links. In a social network, for instance, we may know which people friended/not friended someone and we want to know which people are likely to friend some others.

The rest of this paper is organized as follows. Section 2 introduces the related work on the link prediction problem. Section 3 briefly describes the Dirichlet Process (DP) and its representation, called Chinese Restaurant Process (CRP). Section 4 explains the Dependent Chinese Restaurant Process (DDCRP) model, from which we use to model the latent cluster memberships. In Section 5, we propose a supervised link prediction method based on DDCRP model. Section 6 deals with the posterior distribution, which is used to predict the latent links. Section 7 addresses the prediction of the missing links in the network based on the inferred posterior distributions. In Section 8, the effectiveness of our proposed method is shown by experimental results. Finally, in Section 9 we draw conclusions.

2. Related work

Existing link prediction methods fall into two categories: unsupervised [1,2] and supervised methods [3–6]. Most of link prediction methods are unsupervised, which assign scores to potential links. Each score is interpreted as the similarity between two end nodes of a link; then, all non-observed links are listed in descending score order. The links that scored more points (i.e. connected to more similar nodes) are supposed to be more prospective links in future.

The unsupervised methods are categorized into three main classes:

(1) Those predict prospective links based on the adjacency of the network nodes such as Preferential Attachment [7] and Common Neighbors [8]. In Common Neighbors, the score of link (i, j) is the number of the link neighbors in an undirected network and in a directed network, the score is computed as the number of the link out-degree neighbors, which are shared by nodes i and j (the node *out-degree* is the number of tail endpoints adjacent to the node and the node *in-degree* is the number of head endpoints adjacent to the node). The Preferential Attachment link prediction score is the product of the degrees of the nodes i and j.

(2) Those utilize all paths as an ensemble, such as Katz [9]. The Katz score method sums a set of paths where the weight of short paths exponentially was reduced by value (Relation 1).

$$\text{Score}_{Katz}(i, j) = \sum_{l=1}^{\infty} \gamma^l |P_{ij}^l| \tag{1}$$

where P_{ij}^l is the set of all paths having the length of l from node i to j, and γ^l is the damping parameter for the set of such paths. (3) Finally, those consist of high level methods such as matrix factorization and clustering.

Although the unsupervised link prediction methods are so simple to implement, they can utilize features of the network topology only; moreover, they are domain-specific.

Because of the problems stated above, some researchers have used *supervised* machine learning techniques to solve link prediction problems where each data point corresponds to a pair of nodes in the network graph. To train such models, we use the link data in the *training* interval $[t_0, t'_0]$ and predict future links in the *test* interval $[t_1, t'_1]$. Formally speaking, given that $u, v \in V$ are two nodes of graph $G(V, E)$, y_{uv} is defined as a label of the data point (u, v) (Relation 2). We assume the interaction between u and v is symmetric; so, the pair of (u, v) and (v, u) represent the same data point and hence $y_{uv} = y_{vu}$.

$$y_{uv} = \begin{cases} +1 & \text{if } (u, v) \in E \\ -1 & \text{if } (u, v) \notin E. \end{cases} \tag{2}$$

Based on Relation 2, for labels y_{uv} we build a *classification* model using a set of training data points and then we predict the unknown label for a pair of nodes $(u', v') \notin E$ of the graph G in the interval $[t_1, t'_1]$. Since it is a typical binary classification, the well-known supervised classification tools, such as support vector machines (SVM), Gaussian Process, Decision Tree, Logistic Regression, Naive Bayes, Neural Networks and K-Nearest Neighbors may be used.

Although the results of supervised methods are much better than unsupervised methods, selecting a set of appropriate features for the classification task is very difficult. Most of the existing supervised link prediction methods have used the topological structure of networks to extract proper features.

Recently, a number of methods have proposed probabilistic models to learn features of the network entities from training data [10–15]. Some methods assume that there exists a set of latent *clusters* where each entity (node) belongs to a single cluster or has a distribution over a number of clusters [10–14]. On the other hand, some methods use the features to describe the network nodes where the relationship between two entities is determined by their common features [15,16].

We used the *Dirichlet* Process to cluster the network labels. To show why we based this process, we first explain three other methods in the Sections 2.1–2.3. Then, we state that because of a problem with these methods, we cannot exploit them despite their advantages (in the methods, it is assumed that there are K latent clusters (features), where K need not be fixed a priori).