Link Prediction in Network by a Modified Mutual Information Model

Yuling Yang
College of Information Systems and Management
National University of Defense Technology
Changsha, China
yulingyoung@yeah.net

Guangquan Cheng
College of Information Systems and Management
National University of Defense Technology
Changsha, China
yyl9505@126.com

Kuihua Huang
College of Information Systems and Management
National University of Defense Technology
Changsha, China
yang_ma_cn@163.com

Zhong Liu
College of Information Systems and Management
National University of Defense Technology
Changsha, China
phillipliu@263.net

Abstract— **Link prediction in a network refers to predicting the possibility of connection between two nodes. A traditional method, Local Baysian Method, based on nodes' common neighbors, achieves high prediction accuracy as well as has low computing complexity. However, the method ignores the Mutual Information between the common neighbors. So, we take mutual information model into consideration, while the algorithm has high computing complexity. In this paper, we will modify the model and make it more efficient.**

Keywords- link prediction; Mutual Information; baysian network.

I. Introduction

Real-world systems can be modeled by complex networks in most cases. A typical network is composed of nodes and links, where nodes represent different individuals in the system, and links represent relationships between individuals. If there is a connection between two nodes, edges are joined, and vice versa. Two nodes connected by an edge are considered neighbors in the network. The nervous system of nematode worms, for example, can be thought of as a network of neurons connected by synapses. The American aviation network can be seen as a network formed by airports connected with each other through existing direct flight routes. Similarly, there are computer networks, social networks, logistics networks and so on.

Link prediction in the network refers to predicting the possibility of connection between two nodes that have not yet generated edges or whose connection has not yet been discovered [1] through known network structure and other information, which is actually a process of data mining. For example, A is a friend of B's, B is a friend of C's, then there may be a connection between A and C. The traditional link prediction method is to use Markov chain or machine learning to predict nodes using nodes' attributes. The prediction accuracy of this method is high, but its computational complexity and non-universal parameters limit its uses. Another method is mainly based on similarity and likelihood analysis, which uses the network structure characteristics. Among various similarity-based indices, Common Neighbors (CN) is undoubtedly the precursor with low computing complexity. This paper mainly adopts this method.

II. Problem Description

Considering an undirected network $\mathrm{G(V,E)}$, where V is the set of nodes and E is the set of links. Multiple links and self-connections are not allowed. Denote by U the universal set containing all $|\mathrm{V}|\cdot(|\mathrm{V}|-1)/2$ possible links, where $|\mathrm{V}|$ denotes the number of elements in set V, and $|\mathrm{E}|$ denotes the number of edges in set E. Then, the set of nonexistent links is $\mathrm{U}-\mathrm{E}$. We assume there are some missing links (or the links that will appear in the future) in the set $\mathrm{U}-\mathrm{E}$, and the task of link prediction is to find out these links. Generally, we do not know where the missing or future links are, otherwise we do not need to do prediction. Therefore, to test the algorithm's accuracy, the observed links, E, is randomly divided into two parts: the training set, E^{T}, which is treated as known information, while the probe set (i.e., validation subset), E^{P} , is used for testing and no information in this set is allowed to be used for prediction. Clearly, $\mathrm{E}^{\mathrm{T}}\cup\mathrm{E}^{\mathrm{P}}=\mathrm{E}$ and $\mathrm{E}^{\mathrm{T}}\cap\mathrm{E}^{\mathrm{P}}=\emptyset$. Considering a simple undirected network denoted as $\mathrm{G(V,E)}$, the given network can be represented by an $\mathrm{N}\times\mathrm{N}$ (N represents the number of the nodes) adjacency matrix A, where the element $\mathrm{A_{ij}}=1$, if nodes i and j are connected and $\mathrm{A_{ij}}=0$ otherwise.

III. Modified Mutual Information (MI) Approach

In probability theory and information theory, the Mutual Information (MI) of two random variables is a measure of the mutual dependence between the two variables. More specifically, it quantifies the "amount of information" (in units such as shannons, commonly called bits) obtained about one random variable through observing the other random variable. The concept of mutual information is intricately linked to that of entropy of a random variable, a fundamental notion in information theory that quantifies the expected "amount of information" held in a random variable.

A. *Mutual Information (MI) Approach*

Considering a random variable X related to the outcome x_k and probability $p(x_k)$, its self-information $I(x_k)$ can be denoted as

$$I(x_k) = log\frac{1}{p(x_k)} = -\,log\,p(x_k) \qquad (1)$$

where the base of the logarithm is specified as (1), thus the unit of self-information is bit. This is applicable for the following if not otherwise specified. The self-information indicates the uncertainty of the outcome x_k. Obviously, the higher the self-information is, the less likely the outcome x_k occurs.

Consider two random variables X and Y with a joint probability mass function $\mathrm{p(x,y)}$ and marginal probability mass functions p(x) and p(y). The mutual information $\mathrm{I(X;Y)}$ can be denoted as follows:

$$\begin{aligned} I(X;Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\,log\frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y)\,log\frac{p(x|y)}{p(x)} \qquad (2) \end{aligned}$$

Thus, in the network, the mutual information between x_i and x_j can be represented as:

$$\begin{aligned} I(x_i, y_j) &= log\frac{p(x_i|y_j)}{p(x_i)} \\ &= -\,log\,p(x_i) - \left(-\,log\,p(\chi_i|y_j)\right) \qquad (3) \end{aligned}$$

Mutual information is a measure of the dependency between two variables. $I(x_i, y_j) = 0$ represents that x_i and y_j are independent to each other. Considering link prediction method, we want to use local structure information to improve the prediction. For this purpose, we use Γ(x) to represent the set of adjacent nodes of node x. For node pairs (x,y), the set of their common neighborhoods is denoted as $\mathrm{O}_{\chi y} = \Gamma(\mathrm{x}) \cap \Gamma(\mathrm{y})$Given an unconnected node pair (x,y), if the set of its common neighbor $\mathrm{O}_{\chi y}$ is available, the likelihood score of node pair (x,y) is defined as

$$I(L_{xy}^1|O_{xy}) = I(L_{xy}^1) - I(L_{xy}^1; O_{xy}) \qquad (4)$$

$I(L_{xy}^1)$ is the self-information of that node pair (x,y) is connected. $I(L_{xy}^1; O_{xy})$indicates the reduction in uncertainty of the connection between nodes x and y due to the information given by their common neighbors.

If the elements of Oxy are assumed to be independent of each other, then

$$I(L_{xy}^1; O_{xy}) = \sum_{z\in O_{xy}} I(L_{xy}^1; z) \qquad (5)$$

$$I(L_{xy}^1; z) = \frac{1}{|\Gamma(z)|(|\Gamma(z)|-1)} \sum_{m,n\in\Gamma(z)} I(L_{mn}^1; z) \qquad (6)$$

$$I(L_{mn}^1; z) = I(L_{mn}^1) - I(L_{mn}^1|z) \qquad (7)$$

Here $I(L_{xy}^1; z)$ is defined as the average mutual information over all node pairs connected to node z. $I(L_{mn}^1|z)$ is the conditional self-information of that node pair (m,n) is connected when node z is one of their common neighbors, and $I(L_{mn}^1)$ denotes the self-information of that node pair (m,n) has one link.

B. *A Modified Model*

Since the computation of the Mutual Information of pair nodes costs much time, we want to simplify it. In formula (2), it is easy to relate the sum of possibility $p(x,y)\,log\frac{p(x|y)}{p(x)}$ to the expectation of $log\frac{p(x|y)}{p(x)}$, so we change the formula (2) into (8)

$$\begin{aligned} I(X;Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\,log\frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y)\,log\frac{p(x|y)}{p(x)} \\ &= E_{x,y\leftarrow p(x,y)}\left[log\frac{p(x|y)}{p(x)}\right] \qquad (8) \end{aligned}$$

We can sample the network nodes, and calculate the $log\frac{p(x|y)}{p(x)}$ of them. When the sample is big enough,The expectation in formula (8) is close to the real $I(X;Y)$.

IV. Conclusion

By modifying the model, we repair the big bug in the traditional Baysian method in network link prediction. The method is simple and fast. It approximates the real value with simulation results, and saves a lot of computing time.

References

[1] F. Tan, Y. Xia, and B. Zhu, “Link Prediction in Complex Networks: A Mutual Information Perspective” Plos One, 2014, 9(9):e107056.

[2] Z. Boyao, X. Yongxiang, and S. N. Irene , “Link Prediction in Weighted Networks: A Weighted Mutual Information Model” PLOS ONE, 2016, 11(2):e0148265-.

[3] H. Shakibian and N.M. Charkari. “Mutual information model for link prediction in heterogeneous complex networks” Scientific Reports, 2017, 7:44981.

[4] A.V.D. Oord, Y. Li, and O. Vinyals, “Representation Learning with Contrastive Predictive Coding”, 2018.

[5] L. Lü et al. “Toward link predictability of complex networks”, Proceedings of the National Academy of Sciences, 2015, 112(8):2325-2330