Link Prediction in Social Network

Eric L. Lee, Shu-Hao Chang, Yu-Cheng Weng, Can Jiang

{lee3388,denielll, weng47, Jiang607}@purdue.edu.

Link prediction is one of the core problems of network-structure data. Some heuristics work pretty well in network kind of data such as katz index[1], common neighbor, ... etc. However, because the heuristic method highly depends on the assumption we make in the data set, it is not a general solution. What we want to do is propose a machine learning model that can work in the network-structured data. In Recent years, there are some works focusing on these solutions. Zhang and Chen [2] [3] focused on solving these problems by labeling nodes in an enclosed subgraph. We want to do experiments to test all the methods that proposed in recent years in different network and do some data analysis about these methods. Besides, we aims at proposing a new machine learning model that can achieve better result.

1 INTRODUCTION

Given two nodes in a graph. Link prediction is a task to predict the probability whether there will be a link between them. It can benefit lots of applications

such as friend recommendation in social network, movie recommendation, and metabolic network reconstruction.
Although it is an important problem, it is not a trivial problem for machine learning. One of the most difficult problems is to model the geometrical information of a certain node. Unlike common machine learning problems that already have specific feature vector, it is hard to find a feature vector to describe the situation of a node.

1.1 Related Works

There are already some existing works about link prediction. We organized these methods into three categories: Heuristics, latent feature models, and Labeling nodes in a enclosed subgraph.

1.1.1 Heuristic Method

The most straight forward method is to design heuristics. In network structured data, there are some famous heuristics for prediction. One of famous heuristics is common neighbors. It is proved to be a very successful and there are lots of heuristics extend common neighbors. For example, jaccard index is an intuitive and successful idea. In 2003, Adamic and Adar [6] proposed an index that decrease the contribution of the nodes having high degrees and achieve a better performance. In 2009, based on the same idea, Zhou et al. [7] propose Resource Allocation index for link prediction. There are also heuristics for other kind of link prediction problem. For example, Preferential Attachment [8] is a popular index for road network prediction.
The heuristic mentioned above only considered the node that are one hop away (common neighbor, Jaccard, Preferential Attachment) or two hop away(Adamic, Resource Allocation) from the node we want to predict. There are also some methods consider nodes that is high distance away from the target node. For example, Katz [1] apply exponential decay to distance to the nodes that has at least one path to arrive. Local pagerank uses random walk to find a stable state when the start point is the target node.

1.1.2 Latent Feature Model

One of the most successful solutions is to train latent features for each node. One successful example is matrix factorization. Matrix Factorization is a method that have a very good performance in the field of bipartite graph link prediction. Koren et al. proposed a matrix factorization [4] method minimizing square error and have a very good performance in terms of Root Mean Square Error(RMSE) in the Netflix competition. In 2012, Rendle et al. [5] proposed Bayesian Personalized Ranking to optimize Area Under receiver operating characteristic Curve(AUC) and solve ranking problem while using matrix factorization. This methods achieve a very good result in link prediction in bipartite graph. However, these methods are specifically design for link prediction in bipartite graph.

1.1.3 Labeling node in an enclosed subgraph

In 2017, Zhang et al. proposed a novel machine learning method. The hardest part to design a matching learning model is to extract geographical feature that can describe the geometry information for a certain node. Zhang proposed a novel method to solve the problems. He first divide the network into several enclosed subgraph. For each subgraph, he uses Weisfeiler-Lehman (WL) graph labeling to label nodes. The magic of WL graph labeling is that it preserves the order of each node in each iteration.
After labeling the nodes in the enclosed subgraph, we can enumerate all the nodes and get a adjacency matrix that can be used for feature vectors. The paper shows that their result outperforms most previous methods if they train their model with convolutional neural network.

2 Research Plan

2.1 Data Acquisition and Evaluation

The main data source of this project will be from SNAP (Stanford Network Analysis Project). For each node within each network we randomly remove 10% of edges as testing set. However, the network provides the time when each link is

established, we will remove latest 10% edges as testing set. For each node, we will propose a candidate list that ranks the remaining potentially linked nodes. In this way, we can calculate F1@k, Recall@k, Accuracy@k, AUC, MAP@k, ... metrics that are related to ranking.

2.2 Research Direction

We will first implement the method we mentioned in the related work section and do some data analysis of each model. After that, we will do some research topics listed below.

2.2.1 Ensembling Heuristics

The first step we want to do is to explore each heuristic node. In this way we not only do some data analysis but also build a very intuitive baseline. For the data analysis part, what we want to know is the following two things:
1. What heuristic is better in different network?
2. Is there any combination of the heuristics that can achieve better result?

For the baseline model part, we want to use a machine learning model and use each heuristics I mentioned above as features. However, it is not a trivial problem. Degrees of nodes are much smaller than the number of nodes in the graph which means the data is very imbalanced. This poses a challenge to a machine learning model.Negative data is much more than the positive one.
There are lots of topics that we want to explore:
1. How can we sample negative data? Will it affect result a lot?
2. If negative sampling affect a lot to performance, can we find a good way to sample negative data?
3. What machine learning models can achieve better performance?

2.2.2 Leverage latent feature models

To our best knowledge, there is no latent feature models applied on general link prediction problems directly. However, there are lots of latent feature models

aims at solving community detection, role detection problems in social network. For example, Yang and Leskovec [9] propose BIGCLAM that can do overlapping community detection.
One of our directions is to incorporate these kind of models and design a new model that considers the latent features.

2.2.3 USE META INFORMATION OF NODES

We believe the meta information in each node can further improve the result of link prediction. One simple and naive solution is to add this information as features. The other is to use the information to enhance the graph labeling. Currently, we have no concrete solution about this. But we will keep pondering over it in this semester.

2.2.4 MODELING RECIPROCALNESS

One interesting fact about social network is that, if there is a link between A and B, it is not only because A likes B but also B likes A. We believe that if we take this into account, we can further improve the performance.

REFERENCES

[1] Leo Katz. *A New Status Index Derived from Sociometric Analysis*

[2] Muhan Zhang, Yixin Chen. *Link Prediction Based on Graph Neural Networks*

[3] Muhan Zhang, Yixin Chen. *Weisfeiler-Lehman Neural Machine for Link Prediction*

[4] Yehuda Koren, Robert Bell and Chris Volinsky *Matrix Factorization for recommender system*

[5] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner and Lars Schimidt-Thieme *BPR: Bayesian Personalized Ranking from Implicit Feedback*

[6] Lada A Adamic and Eytan Adar *Friends and neighbors on the web. Social networks, 25(3):221-340, 2003*

[7] Tao Zhou, Linyuan Lu, and Yi-Cheng Zhang. *Predictiong missing links via local information. The European Physical Journal B, 71(4):623-630, 2009*

[8] Albert-Laszlo Barabasi and Reka Albert. *Emergence of scaling in random networks. Science, 286(5439):509-512, 1999.*

[9] Jaewon Yang, Jure Leskovec *Overlapping Community Detection at Scale: A Nonnegative Matrix Factorization Approach*