Link Prediction Techniques in Complex Networks

M. USMAN AKHTAR
University of Engineering and Technology, Peshawar, Pakistan
Email: ua@uetpeshawar.edu.pk

M. IMRAN KHAN KHALIL
University of Engineering and Technology, Peshawar, Pakistan

Abstract

In this paper we discuss how recently emerged machine learning approach, and conventional graph theoretic approaches used for the prediction of missing links in real world complex networks. Future projected plan can be build based on prediction results. This paper shows that the machine learning approach is significantly good as a newly emerged field. If any real-world situation can be mapped in to complex graph. where the nodes in the graph represent different objects of real world, and the links in the graph denotes the link, then a subset of these links is given as an input to the algorithm for prediction. It will also be describe the mechanism of different structural based link prediction algorithms.

Keywords: *Graph Theoretic Approach, Machine Learning, Complex Networks, Neural Networks, Link Prediction.*

Introduction

Due to the evolving nature of world real world complex networks also evolve over time, new links and edges are added (Dorogovtsev & Mendes, 2002). The structural based link prediction approaches such as graph theoretic and machine learning link prediction approach uses past data to predict the future structure of a complex network (Zhu & Xia, 2015). Define the links that may appear in the near future and yet not part of the complex network. The missing link prediction have variety of uses in different areas generally used in online social networks such as Linkeidn, google plus, twitter, facebook new friends (Liben-Nowell & Kleinberg, 2007)and events are recommended on the bases of different parameters likes event or friend suggestion on the bases of common friends and event occurring near residence. Utilization of missing link prediction can also be observed in recommender systems (Schafer, Konstan, & Riedl, 2001), such as new items on daraz online store is recommended on the bases of previous grocery shopping list, in the same way our user previously requests related books and movies are recommended on amazon online book store and netflex an online movie store. Link prediction also used to predict medicine formulation for any newly appeared fatal health care problem. Conventional methods based on the similarity score for missing link between pair of nodes, such as nodes degree, common neighbors, shortest distance between pair of nodes, where a missing link score compared against threshold value for the missing links and edges. with the highest scores are considered the most likely edges.

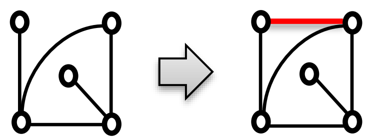
In this paper, we present different structural based link prediction algorithms with supervised learning algorithms for the link prediction problem. Current network state used as an initial input to train a learning algorithm to predict links which may appear in the near future but are absent in the input network. Then, for each missing link, we use the threshold value of the learning algorithm as our link prediction heuristic.

Problem Setting

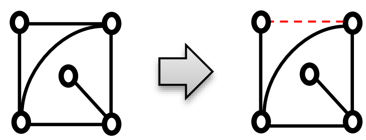
Link prediction problem can be understood in following way. Let a graph $G(V, E)$ where V represents vertices and E represents edges in a graph at a given time t. The network may evolve over time and we are interested to know how the network will look at time t+1, by using the network structure at time t as an input as given in Figure 1(a).

Scientists with mathematical background utilizes given topological structure of the given complex network for forecasting the future structure of complex network. Scientists with latest machine learning background focused to solve the missing link prediction problem by utilizing the graph nodes structural information specific to different complex networks individually.

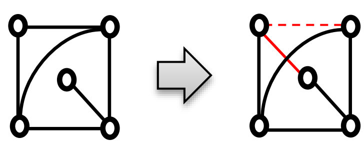
There are three categories of link prediction issue. (i) Predicting new links in current network. (ii) Eliminating existing links from network and (iii) both, eliminating and adding links in complex network.



(a) ***Prediction new links***



(b) Eliminating existing links



(c) Adding and removing links links

Figure 1: Link prediction problems

Predicting new Links

In upcoming time one or more new links will be created between nodes. See Figure 1(a).

Eliminating existing links

Means we predict about the links that will disappear in the near future. Parallel to the condition of adding new links. See Figure 1(b).

Inserting new links and removing old links

It's the mixture of two earlier described link prediction problems. See Figure 1(c).

In this research, we will be working only on first type of link prediction problem. Which is to predict the future structure of graph with addition of new links. Motive for selecting type one problem is mainstream work focuses on it.

Approaches

All over the research paper i and j represents nodes, N represents number of nodes in graph. Γ(i) is the neighbor set of node i, $\Gamma(i) \cap \Gamma(j)$ is the set of common neighbors of node i and node j. where k is the degree of node. We will discuss two approaches of missing link prediction in complex networks at first we will discuss Graph theoretic approach and secondly we will explain some machine learning algorithms for link prediction.

Graph Theoretic Approaches

We documented a brief report of generally used link prediction algorithms that use topology structure of networks in the link prediction process.

Common Neighbor: Ranking of the likelihood of edges is based on number of common neighbors between two distinct nodes(Lü, Jin, & Zhou, 2009).

$$\delta_{ij} = |\Gamma(i) \cap \Gamma(j)|$$

Common Neighbor and Distance: When there are no common neighbors between them $\Gamma(i) \cap \Gamma(j) = \emptyset$, distance d_{ij} between nodes is used for score (Yang & Zhang, 2016).

$$\delta_{ij} = \begin{cases} \frac{CN_{ij} + 1}{2}, & \Gamma(i) \cap \Gamma(j) \neq \emptyset, \\ \frac{1}{d_{ij}}, & otherwise, \end{cases}$$

Preferential Attachment: Ranking of scores is based on degree multiplication of two distinct nodes (Liben-Nowell & Kleinberg, 2007).

$$\delta_{ij} = k_i * k_j$$

Sorensen Index: Is the ratio of common neighbor to that of their submission of their degrees (Gao, Musial, Cooper, & Tsoka, 2015).

$$\delta_{ij} = \frac{2|\Gamma(i) \cap \Gamma(j)|}{k_i + k_j}$$

Adamic Adar: Improves the scoring by weighting lesser degree nodes heavily (Liben-Nowell & Kleinberg, 2007).

$$\delta_{ij} = \sum z \in \Gamma(i) \cap \Gamma(j) \frac{1}{log\ k_z}$$

Hub Promoted Index: Score is computed by computing the Ratio of common neighbors to that of minimum degree of connected nodes (Gao et al., 2015).

$$\delta_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{\min\{k_i, k_j\}}$$

Machine Learning Approaches

Logistic Regression: It's a standard machine learning algorithm. The algorithm try to find θ to get optimum outcome. This learner uses the logistic function, $g(z)\frac{1}{1+exp-z}$ which maps $\theta^T x$ to values between 0 and 1. which trains a binary machine learning classifier (Julian & Lu, 2016).

The supposition signifies the possibility that $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ the training set, y, is true. Using random gradient increased result with the update rule, $\theta_j = \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)})x_j^{(i)})$. The θ can be updated according to need to achieve improved link prediction.

Random Forest: Algorithm combines different trees (groups in network) and combines there mutual characteristics by taking mode. Every time take a random training set E^T as an input and trains the algorithm on it (Hristova, Noulas, Brown, Musolesi, & Mascolo, 2016).

Softmax. It trains a neural network. The softmax activation converts the link possibility in to probability and thus we get a probability distribution.

The higher the probability confirms the link $h(z)_j = \frac{e^{z_j}}{e^{z_0} + e^{z_1}}\ j \in \{0,1\}$ formation chances in near future (Julian & Lu, 2016).

Conditional Temporal Restricted Boltzmann Machine (ctRBM)

The algorithm is capable to handle time-based variations in neighbor properties in the input training set E^T and makes predictions projected according to the current time frame. The neighbor influence is described as
$\eta_p^t = \frac{1}{U_p^t}\sum_{q=1}^{n} l(x_p^t, x_q^t) * P(\tilde{V}_q^t | H_q^t).P(H_q^t | V_p^t, V_q^{<t}; \theta_q)$ where $U_p^t = \sum_{q=1}^{n} l(x_p^t, x_q^t)$ and the indicator function l is 1 if node q is linking to node p at time t, and 0 otherwise. θ_q is the parameter of model M_q. Neighbors make their predictions for node p base on their own experience (historical data) (Li et al., 2014).

Algorithm 1 Predicting G_{t+1} by ctRBM

Input: A trained $\mathbb{M}$ for all nodes, in which $M_p \in \mathbb{M}$ is parameterized by $\theta_p : \{W_{Ap}, W_p, a_{Ap}, a_p, b_p\}$.
A series of snapshots $\{G_{t-N+1}, ..., G_t\}$, where N is the window size.

Output: G_{t+1}

$p \Leftarrow 1$, $G_{t+1} \Leftarrow zeros(length(\mathbb{V}), length(\mathbb{V}))$
for $p < length(\mathbb{V}) + 1$ **do**
  $V_p^{<t+1} \Leftarrow \{G_{t-N+1}^p, ..., G_t^p\}$
  $V_p^{t+1} \Leftarrow ones(1, length(\mathbb{V})) \times 0.5$
  Get neighbor indices: $Idx \Leftarrow find(G_t^p == 1)$
  Get neighbor models list: $\mathbb{M}_{nbr} \Leftarrow \mathbb{M}(Idx)$
  Calculate η_p^{t+1} by Eq.(3.6) given $\mathbb{M}_{nbr}$
  $a_p^{t+1} \Leftarrow a_p + \eta_p^{t+1}$
  Calculate $\tilde{V}_p^{t+1}$ by Eq.(3.4) & (3.7) substituting V^t and $V^{<t}$ by V_p^{t+1} and $V_p^{<t+1}$
  $G_p^{t+1} \Leftarrow \tilde{V}_p^{t+1}$
end for

Conclusion

Through machine learning approach, we are capable to train the algorithms on different datasets to predict network-specific missing links. which outperform traditional similarity techniques. Notably, the use of only distance, degree and other features is enough to have higher accuracy than our baseline of the Adamic/Adar coefficient. Because future network structure is based on the topological structure of variety of different networks and do not use similar features all the time, it is possible to investigate similarities among network structures by cross-checking the performance of the learning algorithms.

It should be noted that networks growth is simulated by randomly removing few links. Unfortunately, Networks from different domains exhibited different network structures which degraded the performance of link predictors.

References

Dorogovtsev, S. N., & Mendes, J. F. (2002). Evolution of networks. *Advances in physics, 51*(4), 1079-1187.

Gao, F., Musial, K., Cooper, C., & Tsoka, S. (2015). Link prediction methods and their accuracy for different social networks and network metrics. *Scientific programming, 2015*, 1.

Hristova, D., Noulas, A., Brown, C., Musolesi, M., & Mascolo, C. (2016). A multilayer approach to multiplexity and link prediction in online geo-social networks. *EPJ Data Science, 5*(1), 24.

Julian, K., & Lu, W. (2016). Application of Machine Learning to Link Prediction.

Li, X., Du, N., Li, H., Li, K., Gao, J., & Zhang, A. (2014). *A deep learning approach to link prediction in dynamic networks.* Paper presented at the Proceedings of the 2014 SIAM International Conference on Data Mining.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology, 58*(7), 1019-1031.

Lü, L., Jin, C.-H., & Zhou, T. (2009). Similarity index based on local paths for link prediction of complex networks. *Physical Review E, 80*(4), 046122.

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data mining and knowledge discovery, 5*(1-2), 115-153.

Yang, J., & Zhang, X.-D. (2016). Predicting missing links in complex networks based on common neighbors and distance. *Scientific reports, 6*, 38208.

Zhu, B., & Xia, Y. (2015). An information-theoretic model for link prediction in complex networks. *Scientific reports, 5*, 13707.