2019 International Conference on Communications, Information System and Computer Engineering (CISCE)

A Link Prediction Algorithm by Unsupervised Machine Learning

SHAO HAO
College of Electronic Engineering
National University of Defense Technology
Hefei, China

DENG JIAN
Second Department of Shijiazhuang Campus
Army Engineering University of PLA
Shijiazhuang, China

WANG Lunwen
College of Electronic Engineering
National University of Defense Technology
Hefei, China

Abstract—**In order to solve the problem that single link prediction index cannot be applied to all networks because of the diversity of network structure, this paper proposes a link prediction algorithm which can adapt to network structure. Considering the complementarity of traditional prediction indexes, we use different prediction indexes as multi-dimensional data of unknown links, and use clustering analysis to transform the link prediction into classification. By clustering classification results, we consider the nature of unknown links comprehensively. The simulation results show that the proposed algorithm can adapt to different network structures and has good prediction accuracy in each network on the basis of comprehensive consideration of various traditional prediction indicators.**

Keywords-component; link prediction; clustering; cmplex network.

I. INTRODUCTION

The network topology is one of the most important properties of networks [1]. Link prediction not only predicts lost links in networks, but also can be used to predict future links that may appear in evolutionary networks. In social networks, link prediction can recommend potential friendships between users to help they find new friends [2]. The most commonly used link prediction method is to measure the similarity of node pairs, which calculate and rank the values of each node similarity. The node pairs with higher similarity are more likely to have unknown links [3]. In recent years, link prediction based on network structure has gradually become a research hotspot. Tan F [4] proposed using mutual information of nodes to improve prediction accuracy. Beyza Ermis [5] proved coupling decomposition can improve prediction performance. Pech R [6] brought robust principal component analysis to link prediction.

The proposed prediction indexes are considered to be dominant in all networks. In fact, different networks have distinct structural characteristics. Even in the same network, the structure of different parts is also significantly different. Therefore, a single similarity index cannot be applied to all networks. Ma C [7] proposed a method combining multiple structural features, defined the index function combining multiple structural features, and then the weight of each feature in the function was determined by using the known structural information. In this paper, we propose an adaptive link prediction method (ALP), which combines multiple similarity indexes. In this method, unsupervised machine learning is used to classify the unknown links according to the classification results. The experimental results show that the proposed adaptive link prediction method has better prediction performance in different networks.

II. RELEVANT KNOWLEDGE

A. Basic Properties of Networks

For network $G(N,E)$, N is the node set and E is the link set. a_{ij} represents link information between nodes. If the link between node i and j exists, $a_{ij}=1$; otherwise, $a_{ij}=0$. $\Gamma(i)$ represents the set of neighbor nodes of node i, k_i represents the degree of the node i.

Clustering coefficient denote the tightness of connections between neighbor nodes of node i, which is defined as the ratio of the number of links between all neighbor nodes of a node to the maximum number of links that may be generated.

$$c_i = \frac{2l_i}{k_i(k_i-1)} \tag{1}$$

B. Density Peak Clustering (DPC)

Clustering [8] is a typical unsupervised machine learning algorithm. In this paper, The DPC is used to classify the unknown links in networks has two reasons. First, the input parameters of DPC are less. Secondly, the structure of different networks is significant different, compared with other clustering algorithms such as K-means clustering algorithm, the DPC can find clustering center of different sizes and non-spherical data, which can improve the applicability of the algorithm.

978-1-7281-3681-3/19/$31.00 ©2019 IEEE
DOI 10.1109/CISCE.2019.00145

The DPC considers that the points in the clustering center have the following properties [8]: (1) the nodes with high density are surrounded by other nodes; (2) there is a greater distance between the cluster center and other nodes with higher density.

In clustering, data points with large distances and densities are used as clustering centers.

III. A Link Prediction Method Based on DPC

A. Problem description

Set U represents a set of all possible links in an unauthorized undirected network.

$$|U| = \frac{|N|*(|N|-1)}{2} \tag{2}$$

We consider set V represents a set of the links which has been found, and the set $Z = U - V$ represents a set of the unknown links. The purpose of link prediction is finding these missing links. The traditional link prediction algorithm calculates the similarity indexes s_{ij} between all node pairs, and arranges s_{ij} from large to small. The node pairs whose s_{ij} are larger have greater possibility of links existing between these node pairs. Some local similarity indexes based on neighbor nodes are shown below.

TABLE I. LOCAL SIMILARITY INDEXES BASED ON NEIGHBOR NODES

Index	Definition	Index	Definition
CN	$s_{ij} = \lvert\Gamma(i) \cap \Gamma(j)\rvert$	AA	$s_{ij} = \sum_{x \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log k_x}$
JC	$s_{ij} = \frac{\lvert\Gamma(i) \cap \Gamma(j)\rvert}{\lvert\Gamma(i) \cup \Gamma(j)\rvert}$	RA	$s_{ij} = \sum_{x \in \Gamma(i) \cap \Gamma(j)} \frac{1}{k_x}$

The above similarity indexes consider the properties of neighbor nodes. In addition to the above indexes, this paper also considers the local random walk (LRW). π is defined as the probability of a random walk particle arriving at node x at time x and node y at time $t+1$.

$$\pi_x(t+1) = P^T \pi_x(t) \quad t \geq 0 \tag{3}$$

Where π is a vector with the value of 1 for the first x element and 0 for the rest. P is a Markov probability transition matrix.

The LRW index can be defined be equation (4).

$$s_{ij}(t) = q_x \bullet \pi_{xy}(t) + q_y \bullet \pi_{yx}(t) \tag{4}$$

In this paper, we consider that the initial resource distribution of network nodes is average. So,

$$s_{ij}^{LRW}(t) = \frac{k_x \bullet \pi_{xy}(t)}{|E|} + \frac{k_Y \bullet \pi_{yx}(t)}{|E|} \tag{5}$$

Considering that the traditional path similarity index Katz needs the all paths of the network, it is difficult to apply to large-scale networks. In this paper, we used a local path similarity index LP.

$$s_{ij}^{LP}(t) = A^2 + \alpha \times A^3 \tag{6}$$

In equation (6), α is an adjustable parameter, A is the adjacency matrix of networks, A^3 is the number of paths with length of three between nodes.

B. Algorithm thinking and steps

Different similarity indexes get different prediction results. The prediction performance of different similarity indicators cannot be improved by simple merging or intersection. For example, the left and right circles in Figure 1 represent the set of predicted results of CN and JC indices on Hep-ph networks.

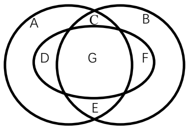

Figure 1. Prediction result

TABLE II. MERGING AND INTERSECTION OF PREDICTION

	CN	JC
The correct part	$D \cup G$	$G \cup F$
Prediction result	$A \cup C \cup D \cup E \cup G$	$B \cup C \cup F \cup E \cup G$
Prediction accuracy	$\frac{D \cup G}{A \cup C \cup D \cup E \cup G}$	$\frac{G \cup F}{B \cup C \cup F \cup E \cup G}$
The correct part after merging	$D \cup G \cup F$	
Accuracy after merging	$\frac{D \cup G \cup F}{A \cup B \cup C \cup D \cup E \cup F \cup G}$	
The correct part after intersection	G	
Accuracy after intersection	$\frac{G}{C \cup G \cup E}$	

From Table.2, we can see that there is no direct relationship between prediction accuracy, merging accuracy and intersection accuracy. If the prediction results of different similarity indexes are merged randomly, the false alarm rate will be greatly increased (misjudged as having links for node pairs without links). If the prediction results of different similarity indicators are intersected, the false alarm rate will be greatly increased (the nodes with links will be neglected).

Different from merging or intersection, this paper proposes a link prediction algorithm based on peak density clustering. Different similarity indexes are taken as one-dimension parameters of nodes, and the nodes are classified by DPC and the corresponding parameters are synthesized. Following, we discuss how to select the similarity index as the parameter index of the node, and how to coordinate the efficiency and prediction accuracy.

In this algorithm, clustering is used to classify network nodes. Considering the accuracy of classification, relatively independent indexes are used as the data of each dimension of nodes as far as possible. For this reason, JC, RA, LRW and LP are used as the parameters of node pairs. The reason reasons are chiefly as follows. (1) JC and RA are based on local similarity of nodes, especially the properties of neighbor nodes at both ends of nodes; LWR is based on local random walk model; LP considers the path and node information between node pairs, and each dimension contains as many information as possible. (2) JC is the improvement index of CN. Both RA and AA take the degree of common neighbor as penalty index, so choose one of the two parameters to avoid duplication.

Through these four indexes, clustering analysis is carried out for all unlinked node pairs. The links exists or not can be judged by referring to two criteria: (1) In general, the networks have natural sparsity. In link prediction, many links have been known before, so the number of nodes with links is much smaller than the number of nodes without links. (2) For some pairs of nodes with high index values, it can be considered that there are links between them. Finally, according to the clustering results, the existence of links between pairs of nodes is judged.

To sum up, the steps of the adaptive Link Prediction Method are as follows:

Input: An undirected network $G(N,E)$, unknow link set Z (set element z_{ij} represents node pairs composed of node i and j).
Output: The link between node i and j exists or not.
Setp1: for $i=1\ to\ N$, for $j=1\ to\ N$
Setp2: Calculate $s_{ij}^{JC}=\frac{|\Gamma(i)\cap\Gamma(j)|}{|\Gamma(i)\cup\Gamma(j)|}$, $s_{ij}^{RA}=\sum_{x\in\Gamma(i)\cap\Gamma(j)}\frac{1}{k_x}$, $s_{ij}^{LRW}(t)=\frac{k_x\bullet\pi_{xy}(t)}{|E|}+\frac{k_Y\bullet\pi_{yx}(t)}{|E|}$, $s_{ij}^{LP}(t)=A^2+\alpha\times A^3$
Setp3: for $m=1\ to\ |Z|$, for $n=1\ to\ |Z|$
Setp4: $\rho_m=\chi(dist(x_m,x_n)-dist_{cut})$, $\delta_m=\min_{n:\rho_n>\rho_m}(dist(x_m,x_n))$ and classify node pairs by density peak clustering
Setp5: The node pairs which are classified into link existing are outputs according to the clustering results,

IV. EXPERIMENTAL ANALYSIS

A. Experimental data process

In order to analyze the performance of the algorithm, four networks are used for experimental analysis. They are US Air Network, Political Blogs Network, Jazz Network and NS Network. The parametric properties of each network are shown in Table 2.

TABLE III. PARAMETERS OF NETWORKS

Networks	$\|N\|$	$\|E\|$	$\langle k\rangle$	$\langle c\rangle$
US Air	332	2128	12.81	0.75
PB	1222	16717	27.36	0.36
Jazz	198	2742	27.70	0.63
NS	379	941	4.97	0.79

In the experiment, the links in networks are divided into two parts: the training set E^T and the test set E^P . $E^T\cup E^P=E$, $E^T\cap E^P=\varnothing$ 。The indexes of CN, JC, RA, AA, LRW and LP of all node pairs from set $U\text{-}E^T$ in each network are calculated respectively. Then, JC, RA, LRW, LP are used respectively as one-dimensional data of each node pair. After standardization and normalization, we use link prediction algorithm proposed in this paper, and obtain the link properties of each node pair. The value similarity index of node pairs classified as having links is 1, otherwise it is 0.

AUC is a link prediction algorithm evaluation standard. This paper also compares the AUC values of each algorithm to analyze their performance. AUC indicates that the probability of randomly selecting a link from the test set is higher than that of randomly selecting a link from the set of predicted nodes.

$$AUC=\frac{n^{'}+0.5n^{''}}{n} \tag{7}$$

Where n is the number of random selection, $n^{'}$ is the number of times that the test set score value is greater than that of the set $U\text{-}E$, and $n^{''}$ is the number of times that the values are equal. In particular, the qualitative binary value (0/1) is obtained by the proposed algorithm, so the AUC of the proposed algorithm can be expressed as the correct probability of prediction in the set $U\text{-}E$.

B. Experimental results and analysis

Ninety percent of the links in networks are used as training sets, and the remaining ten percent are used as test sets. According to the experimental steps introduced before, the AUC of each algorithm on different networks are obtained to measure the prediction performance.

TABLE IV. AUC OF EACH PREDICTION ALGORITHM IN DIFFERENT NETWORKS WHEN 90% TRAINING SET

	CN	JC	RA	AA	LRW	LP	ALP
US Air	0.942	0.924	0.962	0.961	0.964	0.948	0.979
PB	0.927	0.917	0.937	0.935	0.931	0.945	0.955
Jazz	0.949	0.965	0.967	0.956	0.968	0.967	0.974
NS	0.970	0.971	0.978	0.978	0.976	0.979	0.990

Table.4 shows the prediction accuracy AUC of the algorithm in different networks. The prediction accuracy of ALP algorithm proposed in this paper are better than traditional prediction algorithms in each network. Compared with CN, JC, RA and AA algorithms, the proposed algorithm improves by 2.97%, 3.30%, 1.42% and 1.77% respectively; compared with LWR algorithm based on random walk model, it improves by 1.59%. The ALP algorithm achieves the best link prediction performance in all networks. It shows that ALP algorithm can synthetically consider different similarity indexes to obtain more comprehensive link prediction information, thus effectively improving the prediction accuracy.

In order to measure the performance of the algorithm comprehensively, the proportion of training set is changed from 90% to 80%. The experimental process is re-carried out and the prediction accuracy AUC of each algorithm is obtained. The AUC is shown in Table 5.

TABLE V. AUC OF EACH PREDICTION ALGORITHM IN DIFFERENT NETWORKS WHEN 80% TRAINING SET

	CN	JC	RA	AA	LRW	LP	ALP
US Air	0.9251	0.892	0.945	0.942	0.950	0.939	0.970
PB	0.9034	0.891	0.924	0.918	0.927	0.937	0.949
Jazz	0.9251	0.943	0.942	0.942	0.941	0.954	0.968
NS	0.9345	0.955	0.961	0.964	0.962	0.967	0.982

Table.5 shows the prediction accuracy AUC of each algorithm in different networks when training set is 80%. When the training set is reduced from 90% to 80%, the A UC index of all prediction algorithms decreases. This is because the proportion of link training set decreases, which leads to the reduction of network information in prediction, such as the reduction of node neighbors, which leads to the reduction of the number of common neighbors of nodes in the CN algorithm and the decrease of prediction accuracy. Whether the training set is 80% or 90%, the prediction accuracy of ALP algorithm is best Moreover, when the training set is reduced, the ALP algorithm has the lowest reduction rate of accuracy.

V. CONCLUSION

In this paper, we propose the ALP algorithm based on unsupervised machine learning, which calculates the similarity indexes of node pairs, synthesizes different indexes through density peak clustering, and improves the accuracy of link prediction by utilizing the complementarity of different indexes. The simulation results show that ALP algorithm has high recognition accuracy in different networks compared with traditional prediction indexes. The main contribution of this paper is to introduce unsupervised machine learning into link prediction. Next, we will consider introducing other weak classifiers to further improve link prediction accuracy.

REFERENCES

[1] Shahrampour S, Rakhlin A, Jadbabaie A. Distributed Detection: Finite-Time Analysis and Impact of Network Topology[J]. IEEE Transactions on Automatic Control, 2014, 61(11):3256-3268.

[2] Gong N Z, Talwalkar A, Mackey L, et al. Joint Link Prediction and Attribute Inference Using a Social-Attribute Network[J]. Acm Transactions on Intelligent Systems & Technology, 2014, 5(2):1-20.

[3] Yang Y, Lichtenwalter R N, Chawla N V. Evaluating link prediction methods[J]. Knowledge & Information Systems, 2015, 45(3):751-782.

[4] Tan F , Xia Y , Zhu B . Link Prediction in Complex Networks: A Mutual Information Perspective[J]. PLOS ONE, 2014, 9.

[5] Beyza Ermiş, Acar E , Cemgil A T . Link prediction in heterogeneous data via generalized coupled tensor factorization[J]. Data Mining and Knowledge Discovery, 2013, 29(1):203-236.

[6] Pech R , Hao D , Pan L , et al. Link prediction via matrix completion[J]. EPL (Europhysics Letters), 2017, 117(3):38002.

[7] Ma C , Bao Z K , Zhang H F . Improving the performance of link prediction by adaptively exploiting multiple structural features of networks[J]. Physics Letters A, 2016, 381.

[8] Wang B, Jian Z, Ding F, et al. Multi-document news summarization via paragraph embedding and density peak clustering[C]// International Conference on Asian Language Processing. 2018.