Fuzzy Systems and Data Mining IX
A.J. Tallón-Ballesteros and R. Beltrán-Barba (Eds.)

doi:10.3233/FAIA231034

Link Prediction Method Combining Node Labels with Common Neighbors

Xinyang Li[a], Yayun LIU[a], Zhijian ZHANG[a,1], Lin JIANG[a], and Hui ZHONG[a]
[a]*Faculty of Science, Kunming University of Science and Technology, China*

Abstract. The link prediction method that only considers topology information without considering node labels does not achieve a good prediction result for label networks. The paper proposes a link prediction method combining Node Labels with Common Neighbors (NL-CN) to solve this problem. First, a similarity index based on node labels (NL) is defined. The similarity of two nodes is measured by the cosine of the angle between the label feature vectors of the two nodes. Secondly, the NL index and the common neighbor index are combined to obtain the binding index and link prediction for social networks using the binding index. Finally, experiments are conducted in six label networks, and the experimental results show that the method can effectively improve prediction accuracy. In the Gene, Citeseer and Cora networks, the AUC values of the NL-CN index were improved by 10.84, 4.76, and 0.22 percentage points, respectively, compared to the traditional Cos+ index with better performance.

Keywords. label networks, link prediction, similarity index, NL(Node Labels) index

1. Introduction

Improving the accuracy of link prediction in different networks is one of the core issues in this field. Link prediction can predict the likelihood that pairs of nodes that are not currently linked will produce connected edges in the future by using known topology and node information, and this prediction includes the prediction of unknown connected edges between two nodes in the network and the prediction of future connected edges[1].

The existing link prediction methods are broadly classified into three categories: based on node attributes, based on network topology, and based on maximum likelihood. However, in some special networks, not only the topology information of the network itself but also the node attribute information is available, and in this class of networks, only considering the topology information of the network does not effectively predict the similarity of connecting edges between two nodes. Therefore, this paper proposes a link prediction method for label networks that combines Node Labels and Common Neighbor (NL-CN).

The NL-CN method first defines an NL index based on the nodes' and neighbors' nodes, which uses the labels to calculate the similarity between nodes. Second, the NL

[1] Corresponding Author: Zhijian ZHANG, E-mail: zhijian@kust.edu.cn. This work is supported by the National Natural Science Foundation of China (11461037), the Scientific Research Foundation of Yunnan Provincial Department of Education (2017ZZX133), and the Introduction of talent scientific research start-up Foundation Kunming University of Science (KKZ3202207025).

index is combined with the CN index to obtain the NL-CN index. In this paper, a new similarity index is defined using node labels. Nodes in real networks have labels or attributes, so the NL-CN method is more applicable to realistic scenarios. Compared with the traditional link prediction methods, the NL-CN method makes more use of the information in the network to have relatively good prediction performance.

2. Related Work

In recent years, many advances have been made in research on link prediction, and in its early studies, most researchers have used link prediction methods based on node attributes. If two nodes share the same attributes, such as the more similar the interests of two people, then the more likely there is a link between these two people. Bu Xinyi et al.[2] used Sina Weibo social network for empirical analysis, defined a new similarity index using the fusion of network structure and node attribute labels, and performed link prediction for nodes.

The most basic link prediction method based on local information is the common neighbor (CN)[3], which describes the similarity between two nodes by defining the number of common neighbor nodes. In addition, many indexes evaluate the similarity between nodes based on local information: Jaccard coefficient[4], Sorensen index, and LHN-I index[5] based on common neighbors, concatenation of neighbors, and degree of nodes, respectively. Adamic and Adar[6] proposed the AA index, which mainly considers the importance of nodes to common neighbors through the degree of common neighbors. Zhou et al.[7] proposed the resource allocation (RA) index with higher prediction accuracy based on the previous index based on local information. The path-based method has three main indexes: the LP index[8], the LHN-II index, and the Katz index[9]. The main methods based on random wandering are the Cos+ index[10], the randomized wandering index (RWR), the restarted randomized wandering index (RWR)[11], and the SimRank index[12], Clauset, Moore, and Newman[13] proposed a link prediction method based on the maximum likelihood value, which is based on the hierarchical model, the stochastic chunking model, and the closed circuit model. Guimera and Sales-Pardo[14] used a stochastic chunking model to divide the nodes in the network into sets. The probability of connecting an edge between two nodes is only related to the corresponding settings in which the node is located.

3. Link prediction method combining node labels with common neighbors

3.1. Link prediction problem description

In link prediction methods for social networks, the social networks are usually regarded as topological graphs. In this paper, we can represent the social network as an undirected graph $G=(V,E)$, which V denotes the set of nodes and E denotes the set of edges. If the node set V contains N nodes, then the graph G contains most $N(N-1)/2$ edges. The link prediction method is mainly given a similarity index S_{xy} of any two nodes x, y.

In performing link prediction, the known edge set E is usually divided into a training set E^R and a test set E^T. In this paper, the ratio of splitting the edge set E is 9:1.

3.2. NL index based on node labels

Consider that in a label network, if two nodes have the same label, the probability of adjacent edges between these two nodes will be higher. In addition, this paper also considers that the labels of node neighbors can also provide information about the similarity between nodes.

In this paper, we use One-Hot Encoding to process the label information of nodes, and then we can obtain a label embedding matrix B, which contains the label embedding vector of each node. B_i denotes the label embedding vector of the node i, which is the first i row of the matrix B. Since both the node's label embedding vector and its neighbor's label embedding vector have an impact on the similarity between nodes, this paper defines the label feature vector as

$$\mathrm{A}_i = \lambda \mathrm{B}_i \oplus (1-\lambda) \sum_{j \in \Gamma(i)} \mathrm{B}_j \tag{1}$$

Where A is the label feature matrix; A_i denotes a row i of the matrix A denotes the label feature vector of node i; $\Gamma(i)$ denotes the set of neighboring nodes of the node i; and the prescribed symbol $\oplus$ is a first and last splicing rule between vectors. Parameter λ is set to assign the effect of the node label embedding vector and its neighbor node label embedding vector on the similarity between nodes.

The matrix of the number of products between the label eigenvectors is

$$\mathrm{Z} = \mathrm{A}\mathrm{A}^T \tag{2}$$

Each node in the network has a corresponding label feature vector, and this paper considers that the similarity between two sets of vectors is related to the angle between the vectors θ, if the angle between two sets of vectors θ is smaller, then they are more similar. Therefore, for any nodes x and y in the network, the similarity index based on the node labels is defined as

$$S_{xy}^{NL} = \cos\theta = \frac{|\mathrm{Z}_{xy}|}{|\mathrm{A}_x||\mathrm{A}_y|} \tag{3}$$

3.3. NL-CN method

In label networks, the label information of nodes is essential for link prediction of the network. However, from a global perspective, topology information in the network should be preserved. Therefore, this section combines the NL index with the most widely used common neighbor (CN) index in network topology-based link prediction methods to form the NL-CN index. The NL-CN index will be defined as

$$S_{xy}^{NL-CN} = \omega |\Gamma(x) \cap \Gamma(y)| + (1-\omega) \frac{|\mathrm{Z}_{xy}|}{|\mathrm{A}_x||\mathrm{A}_y|} \tag{4}$$

Therefore, the proposed algorithm NL-CN is described as follows.

Algorithm 1 NL-CN algorithm

Input:

network diagram G ; node embedding matrix B

Output:

AUC、Precision、RS value

1: **for** each node i **do**

2: Calculate the label feature vector A_i using Equation 1

3: **end for**

4: Calculate Z between label feature matrices using Equation 2

5: **for** $\omega = 0$ to $\omega = 1$ **do**

6: Calculate the similarity between two nodes according to

$$S_{xy} = \omega \left| \Gamma(x) \cap \Gamma(y) \right| + (1-\omega) \frac{\left| \mathrm{Z}_{xy} \right|}{\left| \mathrm{A}_x \right| \left| \mathrm{A}_y \right|}$$

7: Calculate the AUC、Precision、RS value using Equation 5,6,7

8: **end for**

9: Return the optimal AUC、Precision、RS value from Step 7

Suppose the number of nodes in the network is N , and the average degree of nodes is k . For the CN algorithm, its time complexity is $O(N^2 k)$. The proposed link prediction method combining node labels and common neighbors in this paper first embeds node labels into node label feature vectors using unique thermal coding and then makes the number product of label feature vectors of two nodes. Assuming the number of labels in the network is m , then the time complexity of this step is $O(2N^2 m)$, so the time complexity of the NL-CN algorithm is $O(N^2(2m+k))$. Since we consider the label information in the network based on the traditional algorithm, the time complexity of our algorithm increases. However, our algorithm is more applicable to real networks and has better prediction results.

4. Experiment

4.1. Evaluation indexes

AUC index[15] is done by randomly drawing an edge from the test set and then randomly drawing an edge from the unconnected edge set and comparing the size of the connected edge scores between them. Compare independently n times, if there are n' times when the score of the edge in the test set is greater than the score of the edge in the unconnected edge set, and there are n'' times when both scores are equal, then the AUC is calculated as

$$AUC = \frac{n' + 0.5n''}{n} \tag{5}$$

The AUC index should range from [0.5,1]. When the AUC is 0.5, it means that the scores of the edges are randomly assigned. When the AUC indicator is 1, it means that the full prediction is accurate.

Precision index[16] considers edges with scores ranked in the top L. It defines the link prediction accuracy as the proportion of edges in the test set among the top L edges. If m of the top L edges are in the test set, then Precision is computed as

$$\Pr ecision = \frac{m}{L} \tag{6}$$

In this paper, 10% of the edge set of the dataset is chosen as the value of L.

Ranking Score index[17] considers the position of the edges in the test set in the final ranking. Let the set $H = M \cup E^T$, where M is the set of unconnected edges. Let γ_i be the position of the edges $i(i \in E^T)$ in the final ranking, then the Ranking Score of an edge i is $RS_i = \gamma_i / |H|$. By traversing all the edges in the test set, the Ranking Score can be calculated as

$$RS = \frac{1}{\left|E^T\right|} \sum_{i \in E^T} RS_i = \frac{1}{\left|E^T\right|} \sum_{i \in E^T} \frac{\gamma_i}{\left|H\right|} \tag{7}$$

Where $|E^T|, |H|$ denotes the number of elements in the set, and smaller RS values indicate higher predictive accuracy.

4.2. Experimental data set

In this paper, we selected six label networks in the network repository[2], which are internet-industry-partnerships (IIP), webkb-wisc (WEB), ENZYMES8 (ENZ), Citeseer, Cora, and Gene. IIP is a dataset based on internet industry partnerships. WEB is a dataset based on world wide knowledge base project. ENZ is a graph data collection built on the structure of biomolecular proteins. Citeseer is a multi-disciplinary dataset consisting of papers from 10 research fields. Cora is a dataset based on citations between scientific papers. Gene is a database of genetic information for all species. The details of the dataset are shown in Table 1.

Table 1. Data set information

Dataset	Number of nodes	Number of consecutive sides	Average degree	Average clustering coefficient	Number of labels
ENZ	88	133	3	0.02	2
Gene	1103	1672	3	0.21	2
IIP	219	631	5	0.18	3
WEB	262	510	3	0.18	5
Citeseer	3264	4536	2	0.14	6
Cora	2708	5429	4	0.25	7

4.3. Experimental Results and Analysis

The link prediction of the network is first performed by the NL index defined in this paper to derive the optimal λ value in each data set. The optimal λ value and the

[2] https://networkrepository.com/

AUC、Precision、RS are different in different datasets, as shown in Table 2. Data in parentheses are the corresponding λ values.

Table 2. AUC, Precision, RS and corresponding λ values for the NL index in the six data sets

Dataset	ENZ	Gene	IIP	WEB	Citeseer	Cora
AUC	0.7106(0.65)	0.7783(0.55)	0.7425(0.53)	0.6293(0.26)	0.8997(0.49)	0.8945(0.39)
Precision	0.71(1)	0.69(0.67)	0.83(1)	0.55(0.27)	0.91(1)	0.89(1)
RS	0.5857(0.47)	0.3691(0.54)	0.3527(0.43)	0.4763(0)	0.3098(0.5)	0.3205(0.33)

Parameter λ in Equation 1 should be chosen for different data sets as its optimal value. It combines the NL index with the most widely used CN index based on network topology. The prediction accuracy and the corresponding ω values for the six networks using the NL-CN index for link prediction are shown in Table 3. Data in parentheses are the corresponding ω values.

Table 3. AUC, Precision, RS and corresponding ω values for the NL-CN index in the six datasets

Dataset	ENZ	Gene	IIP	WEB	Citeseer	Cora
AUC	0.7554(0.17)	0.8875(0.8)	0.8121(0.37)	0.7001(0.6)	0.9338(0.92)	0.9332(0.55)
Precision	0.84(1)	0.69(0.06)	0.83(0)	0.67(0.72)	0.91(0.5)	0.89(0.5)
RS	0.5769(0.5)	0.304(0.64)	0.3309(0.08)	0.4297(0.46)	0.2913(0.74)	0.2982(0.72)

As seen in Table 3, the prediction accuracy of the combined index improves on both the original index compared to the index considering only the network topology or the node labels. Therefore, it is necessary to consider node labels in the link prediction of label networks. After that, the prediction accuracy of the NL-CN index, NL index, and index based on network topology are compared in different networks.

Table 4. Prediction accuracy of different indexes in different data sets

index	ENZ			Gene			IIP			WEB			Citeseer			Cora		
	AUC	P	RS	AUC	P	RS	AUC	P	RS	AUC	P	RS	AUC	P	RS	AUC	P	RS
NL-CN	0.7554	**0.84**	0.5769	**0.8875**	**0.69**	**0.304**	**0.8121**	**0.83**	**0.3309**	**0.7001**	**0.67**	**0.4297**	**0.9338**	**0.91**	0.2913	**0.9332**	**0.89**	0.2982
NL	0.7106	0.71	0.5857	0.7783	0.69	0.3691	0.7425	0.83	0.3527	0.6293	0.55	0.4763	0.8997	0.91	0.3098	0.8945	0.89	0.3205
CN	0.5472	0.65	0.9822	0.8215	0.67	0.4614	0.7434	0.81	0.6077	0.6706	0.45	0.6826	0.7564	0.78	0.5154	0.7704	0.75	0.5521
RA	0.5483	0.63	0.9822	0.8212	0.65	0.4614	0.7522	0.81	0.5671	0.6829	0.65	0.6494	0.7570	0.76	0.5154	0.7714	0.79	0.5518
AA	0.5478	0.63	0.9822	0.8220	0.65	0.4614	0.7529	0.75	0.5744	0.6839	0.65	0.6494	0.7564	0.76	0.5154	0.7718	0.82	0.5517
LHN-I	0.5444	0.58	0.9822	0.8198	0.63	0.4614	0.6858	0.72	0.5923	0.6331	0.63	0.7122	0.7560	0.82	0.5154	0.7701	0.85	0.5528
Sorenson	0.5438	0.54	0.9822	0.8208	0.64	0.4614	0.6971	0.68	0.5858	0.6441	0.62	0.7137	0.7557	0.83	0.5154	0.7701	0.74	0.5524
Jaccard	0.5441	0.61	0.9822	0.8210	0.67	0.4614	0.6986	0.75	0.5858	0.6430	0.64	0.7137	0.7561	0.73	0.5154	0.7703	0.69	0.5524
Cos+	**0.7991**	0.75	**0.4408**	0.8007	0.66	0.3551	0.6466	0.81	0.3861	0.6745	0.59	0.4977	0.8913	0.89	**0.2869**	0.9311	0.88	**0.287**

From Table 4, it can be seen that the prediction accuracy of the NL-CN index is higher than most indexes. It is slightly lower than the Cos+ index in the ENZ data set. Otherwise, the NL-CN index outperformed all other indexes. Also, the AUC of the NL-CN index is higher in the Citeseer and Cora datasets where the data are more extensive, and the number of labels is more elevated, reaching about 0.93.

Precision for the NL-CN index outperforms the other indexes in all six datasets. Meanwhile, the RS of the NL-CN index is second only to the Cos+ index in the ENZ, Citeseer, and Cora datasets. In addition to this, the RS of the NL-CN indicator outperforms most of the indicators.

5. Conclusion

In this paper, we propose a node labels-based similarity index, which uses the label information of nodes to calculate the similarity of node pairs. Then the NL index is combined with the common neighbors index to obtain the NL-CN index. This method makes up for the deficiency that the traditional index only considers the network topology and ignores the node attribute information. Among the six label networks selected, the prediction accuracy of the NL-CN index is improved to varying degrees compared to the index based on the network topology.

In future work, we will consider how to reduce the time complexity of the method to reduce the time to run in large networks, where the framework also works better. In addition, we will consider the case where a node has multiple labels, extending the current one-dimensional to multi-dimensional, to improve the usability of the framework in link prediction studies.

References

[1] Daud N N, Ab Hamid S H, Saadoon M, et al. Applications of link prediction in social networks: A review. Journal of Network and Computer Applications, 2020, 166: 102716.
[2] Bu X Y, Chen M L. A study of link prediction in social networks. Research on Library Science, 2016(17): 17-21.
[3] Liben-nowell D, Kleinberg J. The link prediction problem for social networks//Proceedings of the twelfth international conference on Information and knowledge management. 2003: 556-559.
[4] Qin X, Han X, Chu J, et al. Density peaks clustering based on Jaccard similarity and label propagation. Cognitive Computation, 2021, 13: 1609-1626.
[5] Yue Z H, Xu H Y, Wang Q F. Dynamic Link Prediction of Knowledge Diffusion in Disciplinary Citation Networks Based on Local Information. Information studies: Theory & Application, 2020, 43(2): 84.
[6] Kumar A, Singh S S, Singh K, et al. Link prediction techniques, applications, and performance: A survey. Physica A: Statistical Mechanics and its Applications, 2020, 553: 124289.
[7] Zhou T, Lü L, Zhang Y C. Predicting missing links via local information. The European Physical Journal B, 2009, 71: 623-630.
[8] Lü L, Jin C H, Zhou T. Similarity index based on local paths for link prediction of complex networks. Physical Review E, 2009, 80(4): 046122.
[9] Bai M, Hu K, Tang Y. Link prediction based on a semi-local similarity index. Chinese Physics B, 2011, 20(12): 128902.
[10] Zhang Y X, Feng Y X. A review of methods and developments in link prediction. Measurement & Control Technology, 2019, 38(2): 8-12.
[11] Shao H, Wang LW, Deng J. A link prediction algorithm based on density peak clustering. Small Microcomputer Systems, 2020, 41(5): 1007-1012.
[12] Jiang W C, Lu M M. A link prediction algorithm with superimposed random wandering gravity model. Journal of Chongqing University of Technology (Natural Sciences), 2022, 36(5): 137-146.
[13] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks. Nature, 2008, 453(7191): 98-101.
[14] Guimerà R, Sales-pardo M. Missing and spurious interactions and the reconstruction of complex networks. Proceedings of the National Academy of Sciences, 2009, 106(52): 22073-22078.
[15] Zeng G, Zeng E. On the three-way equivalence of AUC in credit scoring with tied scores. Communications in Statistics-Theory and Methods, 2019, 48(7): 1635-1650.
[16] Wu Z, Lin Y, Zhao Y, et al. Improving local clustering based top-L link prediction methods via asymmetric link clustering information. Physica A: Statistical Mechanics and Its Applications, 2018, 492: 1859-1874.
[17] Zhou T, Ren J, Medo M, et al. Bipartite network pro-jection and personal recommendation. Physical review E, 2007, 76(4): 046115.