An Index Hybrid Method Based on Improved Logistic Model for Link Prediction

Yi-xi XIE*, Hong-chang CHEN, Lei CHEN, Yi-teng WU and Si-hao LI
National Digital Switching System Engineering & Technological Research Center, Zhengzhou, China
*Corresponding author

Keywords: Complex network, Link prediction, Hybrid index, Logistic model.

Abstract. In link prediction, the single index prediction effect depends on whether the method can reasonably describe the target network topology characteristics. What’s more, the network features are only described from a certain perspective, this limitation leads to the low robustness of a single index for the prediction of different networks. Based on this, we proposed a hybrid method based on Logistic model, considering the similarity index of the prediction results of complementary characteristics and importance in different networks, adaptive fusion gives each index weight reasonably. Experiments show that the AUC and Precision of the hybrid method on each target network are higher than those of the baseline.

Introduction

Complex network link prediction refers to inferring unknown links between implied node pairs in the network. The link prediction method is divided into three types [1]: link prediction method based on network structure similarity, link forecasting method based on maximum likelihood estimation and link prediction method based on probability model. The link prediction method based on the similarity of network structure is simple and only uses the network topology information which is easy to obtain. This kind of method starts from the origin of the network evolution, and proposes a link forecasting method with certain physical meaning and interpretability.

In general, the research of the convergence problem of link prediction algorithm is still in the initial stage. This method can integrate the advantages of multiple similarity indicators, effectively improve the prediction accuracy, and in different data sets on the performance of stability. Based on this, this paper presents a new idea of integration of indicators. Firstly, the non-linear expansion of the Logistic model is carried out. Secondly, the model is optimized for the specific evaluation index. Then, the complementarity of the prediction results between the structural similarity indexes is considered. The weights of each index in the target network are adjusted adaptively Road prediction.

Related Work

Link Prediction Problem Description

Is a non-acyclic network \( G(V, E) \), where \( V \) represents a collection of all nodes in the network and \( E \) represents the set of all edges. Let \( U \) be a complete set of \( N(N-1)/2 \) nodes in a non-directional network. For the accuracy of the test algorithm, the known edge \( E \) is divided into training set \( E^T \) and test set \( E^P \).

Structural Similarity Index in Link Prediction

The link prediction method based on network structure similarity is simple. Its core idea is that the structure of the two nodes in the network is more similar, the greater the likelihood that the node will be on the edge.

The most basic similarity index based on local information is CN index [3].
Where $|\cdot|$ represents the potential of the set, $\Gamma(\bullet)$ representing all nodes of the node.

Adamic [4] proposed AA index, which is defined as (2).

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(|\Gamma(z)|)}$$

The RA index is proposed in [5], and its idea is similar to AA index.

$$RA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|}$$

The PA index is proposed in [6], as shown in equation (4)

$$PA(x, y) = k_x \times k_y$$

Considering the influence of the third-order path on the similarity, the similarity index LP based on local path is proposed in [7].

$$LP = A^2 + \alpha A^3$$

Where $\alpha$ is the weight coefficient of the third-order path and $A$ is the adjacency matrix of the network. When $\alpha = 0$, the local path index is the form of CN index.

Considering the higher order path, add higher order terms to the LP metric. The Katz index [8] takes into account all the paths in the network, as defined in equation (6).

$$Katz(x, y) = \sum_{l=1}^{\infty} \beta^l \cdot |path_{x,y}^l| = \beta A_{xy} + \beta^2 A_{xy}^2 + \beta^3 A_{xy}^3 + \ldots$$

Algorithm Overview

In the case of network structure similarity link forecasting indicators, where a single indicator exhibits better predictive performance in the target network, but in other Performance in the network may not be ideal. Taking into account the different types of indicators between the wrong sample set may be different, so different indicators can give complementary information, through the integration of a variety of indicators can improve the predictive performance have a positive impact. In this paper, the improved logistic model is used to integrate the indexes. The following focuses on the optimization strategy of the model.

Development of Nonlinear Logistic Model

The logistic regression model with linear function fusion structure similarity index has the characteristics of easy description, good explanation. However, the linear model is only an approximation of the real function, its prediction accuracy is limited. The linear function is extended to the nonlinear function, which can effectively improve the prediction accuracy. There are $N$ to be fusion indicators, each indicator of the link prediction score is $x_i, i = 1, 2, \ldots, N, M_i, i = 1, 2, \ldots, M_i$. $\Phi_i(x_i)$ is basis function representing the variable of $x_i$. The linear logic regression model is extended to the function of the predictive variable $x_i$ as a nonlinear logic regression model of the new predictor.

$$p = \frac{\exp\left(\sum_{i=1}^{N} \sum_{j=1}^{M_i} \beta_{ij} \Phi_i(x_i)\right)}{1 + \exp\left(\sum_{i=1}^{N} \sum_{j=1}^{M_i} \beta_{ij} \Phi_i(x_i)\right)}$$

56
Where the basis function \( \Phi_{ik}(x_i) \) is a function of known form, the basis of the selection of a wide range of functions, such as polynomial function, step function, and wavelet function. Since the form of the logistic regression function does not change, the maximum likelihood method or the least squares method can still be used to fit the objective function to obtain the fitting coefficient \( \beta_{ik} \). If there are \( n \) sets of training data, the logical regression model of each set of training data is as shown in equation (8).

\[
p_j = \frac{\exp\left(\sum_{i=1}^{N} \sum_{k=1}^{M_i} \beta_{ik} \Phi_{ik}(x_{ij})\right)}{1 + \exp\left(\sum_{i=1}^{N} \sum_{k=1}^{M_i} \beta_{ik} \Phi_{ik}(x_{ij})\right)}, \quad j = 1, 2, \ldots, n
\]

The optimization problem of solving is shown in equation (9).

\[
L = \sum_{j=1}^{n} (y_j - p_j)^2
\]

\[
\text{minimize } L, \quad k = 0, 1, 2, \ldots, M_i; \quad i = 1, 2, \ldots, N
\]

Where, if and only if \( y_j = 1 \) there is a side, otherwise \( y_j = 0 \). Optimization of the objective function can be used to reduce the random gradient algorithm, set the step \( \theta \), the parameter update process in equation (10).

\[
\beta_{ik}^{t+1} = \beta_{ik}^{t} - \theta \frac{\partial L}{\partial \beta_{ik}}, k = 0, 1, 2, \ldots, M_i; \quad i = 1, 2, \ldots, N
\]

The logistic regression model from linear model to nonlinear model is realized. Taking the polynomial basis function as an example, the maximum number of polynomial basis functions is 2 times, then the basis function is shown in Eq. (11).

\[
p_j = \frac{\exp\left(\sum_{i=1}^{N} \sum_{k=0}^{2} \beta_{ik} x_i^k\right)}{1 + \exp\left(\sum_{i=1}^{N} \sum_{k=0}^{2} \beta_{ik} x_i^k\right)}, \quad j = 1, 2, \ldots, n
\]

**Optimized Logistic Model for AUC Optimization**

Whether it is a linear logic regression model or a nonlinear logic regression model, the optimal objective function is to make the two types of sample scores as close as possible, so as to indirectly achieve the goal of maximizing the AUC index of the fusion index. If the optimized objective function is set as the AUC value of the fusion index on the training set, the accuracy of the link prediction can be further improved.

If \( x_j = (x_{j1}, x_{j2}, \ldots, x_{jn})^T \), the link prediction score for each set of training data is shown in Equation (12).

\[
p_j(x_j) = \frac{\exp\left(\sum_{i=1}^{N} \sum_{k=1}^{M_i} \beta_{ik} \Phi_{ik}(x_{ij})\right)}{1 + \exp\left(\sum_{i=1}^{N} \sum_{k=1}^{M_i} \beta_{ik} \Phi_{ik}(x_{ij})\right)}, \quad j = 1, 2, \ldots, n
\]

The objective function of directly optimizing the AUC index is shown in equation (13).
\[ L_{AUC} = \frac{1}{a^+a^-} \sum_{i,j} I \left( p_i \left( x_i^+ \right) > p_j \left( x_j^- \right) \right) + \frac{1}{2} I \left( p_i \left( x_i^+ \right) = p_j \left( x_j^- \right) \right) \]

\[ \text{minimize} L_{AUC}, \ k = 0,1,2,\ldots,M_j; \ i = 1,2,\ldots,N \]

(13)

So that is equivalent to solving, if \( L_{AUC} = -L_{AUC} \):

\[ \text{minimize} L_{AUC}, \ k = 0,1,2,\ldots,M_j; \ i = 1,2,\ldots,N \]

(14)

Where \( a^+ \) are number of samples in the training sample, \( a^- \) the number of samples without edge in the training sample; the exponential function is \( I(x) = \begin{cases} 1, & \text{true} \\ 0, & \text{false} \end{cases} \).

The optimization objective function can use the random gradient descent algorithm to set the step size \( \theta \). The parameter update process is shown in equation (15).

\[ \beta_{ik}^{n+1} = \beta_{ik}^{n+1} - \theta \frac{\partial L_{AUC}}{\partial \beta_{ik}}, \ k = 0,1,2,\ldots,M_j; \ i = 1,2,\ldots,N \]

(15)

Optimized logistic model for Precision optimization is similar to AUC.

**Experiment**

**Introduction to Experimental Datasets**

The experiment uses six open complex network data sets for algorithm testing, including: (1) FWFB: The food chain network of the Florida Gulf rainy season. (2) USA air: American Airlines network, where each node represents an airport, and each side represents a direct flight between the two airports connected. The network does not contain the frequency of flight, flight direction and other specific information. (3) Router: The router hierarchy network, where the node represents the router, such as the node is connected to the router through the cable and other means of direct exchange of data packets. (4) Yeast: protein interaction network, where nodes represent proteins, and sides represent protein-protein interactions. (5) Email: contains a factory employee e-mail situation. (6) Jazz: Jazz musicians cooperation network.

**Evaluation Index**

Experiments are used to evaluate the two used link prediction algorithms AUC and Precision [9].

AUC is the abbreviation of area under the receiver operating figure curve, the area under the ROC curve. If the test set of points is high adding 1 point; if the two sides are equal adding 0.5 points.

\[ AUC = \frac{n^+ + 0.5n^-}{n} \]

(16)

The precision index is in accordance with the test scores from large to small, the exact ratio of the predicted edge of the front L bit. The formula is shown in equation (17).

\[ \text{Precision} = \frac{m}{L} \]

(17)

Where \( m \) indicates that the score is in the first \( L \) and there are \( m \) in the test set, \( m \) is predicted correctly.

**Experiment Result**

Based on the combination strategy, this paper considers the performance complementarity between the benchmark indexes and selects CN, RA, AA and PA. In the experiment, the training set and the
test set were divided according to the ratio of 9:1, and the average of 10 independent and repeated experiments was taken as the result. The values marked in bold in Table 1 are the best results. Figure 1 for the precision index test L value of 20,40,60,80,100 when the link to determine the accuracy of the composition of the polyline.

As can be seen from Figure 1 and Table 1, the indicators with too few fusion indexes are limited for the prediction accuracy. For the logistic model, too many features will bring too high computational overhead and lead to over-fitting. In the AUC experiment, the Katz index based on global information with high prediction accuracy was added as a contrast. Both AUC and Precision were significantly improved before the fusion.

![Figure 1. Link prediction Precision curve.](image1)

**Table 1. HLP algorithm and the structural similarity index in the data set AUC value.**

<table>
<thead>
<tr>
<th>Network</th>
<th>FWFB</th>
<th>USAir</th>
<th>Router</th>
<th>Yeast</th>
<th>Email</th>
<th>JAZZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN</td>
<td>0.603</td>
<td>0.954</td>
<td>0.673</td>
<td>0.914</td>
<td>0.919</td>
<td>0.955</td>
</tr>
<tr>
<td>RA</td>
<td>0.607</td>
<td>0.971</td>
<td>0.672</td>
<td>0.913</td>
<td>0.924</td>
<td>0.971</td>
</tr>
<tr>
<td>AA</td>
<td>0.605</td>
<td>0.962</td>
<td>0.664</td>
<td>0.915</td>
<td>0.923</td>
<td>0.963</td>
</tr>
<tr>
<td>PA</td>
<td>0.761</td>
<td>0.914</td>
<td>0.948</td>
<td>0.863</td>
<td>0.857</td>
<td>0.783</td>
</tr>
<tr>
<td>Katz$^{0.01}$</td>
<td>0.677</td>
<td>0.951</td>
<td>0.978</td>
<td>0.971</td>
<td>0.934</td>
<td>0.936</td>
</tr>
<tr>
<td>LRF</td>
<td>0.776</td>
<td>0.979</td>
<td>0.933</td>
<td>0.941</td>
<td>0.966</td>
<td>0.977</td>
</tr>
</tbody>
</table>

According to the four similarity indexes used in this paper, CN, AA, RA and PA are local similarity index based on local information. Their definitions are based on the description of network characteristics from a certain point of view. The time complexity is low, but in different networks the prediction effect is not stable. It can be seen from the experimental data that the prediction effect of the fusion algorithm depends largely on the individual prediction accuracy. The low prediction accuracy of a single indicator can affect the overall performance of the fusion algorithm. Therefore, it is the key to the design of the algorithm to combine the a priori knowledge and to select the similarity index to be fused and to maximize the accuracy of the prediction.

**Summary**

In this paper, four kinds of link predictive indexes based on structural similarity are selected from the point of view of combinatorial strategy, and the Logistic model is improved from the non-linear expansion of the model and the specific optimization of the evaluation index, so as to fuse the four
similarities index. According to the importance of each index, the method adaptively adjusts the weight of each index on a specific network, so as to achieve better forecasting effect. The choice of indicators can be adjusted for a particular network. In particular, the proposed method and optimization strategy can not only be used to integrate the structural characteristics of the network, but also can be used to integrate the attributes of the network.

References


