Physica A 482 (2017) 699–712

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

An evidential link prediction method and link predictability based on Shannon entropy

Likang Yin [a,b], Haoyang Zheng [a], Tian Bian [a], Yong Deng [a,c,*]

[a] School of Computer and Information Science, Southwest University, Chongqing, 400715, China

[b] School of Hanhong, Southwest University, Chongqing 400715, China

[c] Institute of Integrated Automation, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xian, Shaanxi, 710049, China

HIGHLIGHTS

- A novel method for predicting missing links by fusing attribute and structural similarity is proposed.
- This paper compares other nine indexes in both unweighted and weighted networks.
- A new method to measure the link predictability based on Shannon entropy is proposed.

ARTICLE INFO

Article history:
Received 5 December 2016
Received in revised form 18 April 2017
Available online 3 May 2017

Keywords:
Complex networks
Link prediction
Dempster–Shafer theory
Belief function
Predictability

ABSTRACT

Predicting missing links is of both theoretical value and practical interest in network science. In this paper, we empirically investigate a new link prediction method base on similarity and compare nine well-known local similarity measures on nine real networks. Most of the previous studies focus on the accuracy, however, it is crucial to consider the link predictability as an initial property of networks itself. Hence, this paper has proposed a new link prediction approach called evidential measure (EM) based on Dempster–Shafer theory. Moreover, this paper proposed a new method to measure link predictability via local information and Shannon entropy.

1. Introduction

Complex network is ubiquitous in both nature and society [1,2]. Link prediction is a field of great potential deserving research and application. It is used to predict missing links which will exist in the future or already exists but have not been observed. Link prediction is an effective math tool to handle the uncertainty and potential relationship between nonadjacent node in complex networks [3]. Link prediction conducts research in a wide variety of areas generally related to, but not restricted to network science, such as recommendation on social media [4,5], biology analysis [6], network reconstruction [7,8] and other hot fields [9,10]. In the beginning, researchers used probabilistic models to predict missing links, such as Markov chain model [11,12], statistical relational learning [13], naive Bayes model [14], degree-related clustering [15] and other mathematical tools [16,17]. However, the attributes of nodes are easier to know than the local structure of the nodes. Therefore, some other studies focus on modeling the problem based on the network structure, such as subgraph-based ranking model [18], local information model [19,20], link direction [21] and other research fields [22–24]. Moreover, some studies

* Corresponding author at: School of Computer and Information Science, Southwest University, Chongqing, 400715, China. Fax: +86 023 68254555.
E-mail addresses: ydeng@swu.edu.cn, prof.deng@hotmail.com (Y. Deng).

http://dx.doi.org/10.1016/j.physa.2017.04.106
0378-4371/

consider both structural features and the attributes of nodes to build local conditional probability models to predict missing links [25].

Recently, most of the studies focus on the similarity of nodes [26]. It is easy to understand how it works: if we are good friends, there must be some interests that we share together, and another friend of mine has more chance to make friend with you than a stranger [27,28]. From the online social network, we find an interesting phenomenon that people tend to hide their true interests on the online social networks. One reason for this phenomenon is the fact that people want to stay their comfort zone and try to maintain their relationships. Base on that, we consider the similarity between nodes may be either obvious (e.g. the nodes have same neighbors) or hidden (e.g. the nodes have same attributes). This paper proposed a new method base on the Dempster–Shafer theory [29,30] to combine the two kinds of similarity and present the general formula to calculate the similarity degree between nodes. Moreover, previous study [31,32] consider the unpredictability is also a property of network and proposed the predictability of network based on structural consistency. In this paper, we consider not only the whole network but also every single node has its unique unpredictability. To address this issue, we transfer each node in the network to a belief function to represent each node based on Dempster–Shafer theory.

The paper is organized as follows. The related works are briefly presented in Section 2. Section 3 introduces the benchmarks method. Proposed Evidential Measure (EM) and some numerical examples are used to illustrate the procedure of EM in Section 4. The experiment results are shown in Section 5. Discussion about predictability of nodes based on Shannon entropy is stated in Section 6. Finally, the conclusion is presented in Section 7.

2. Related work

With the development of big data and data mining fields, link prediction method became an effective analysis tool in communities of physics and computer science [33–35]. The common framework of link prediction methods is the similarity-based algorithm [36,37]. The problem can be simplified as how to measure the probability of any pairs of two individuals become friends in the social network [38]. To solve this problem, the first idea comes to minds is to measure the similarity or connection between the two individuals since people who are alike often become friends [31]. Moreover, the algorithm of similarity index can be extremely simple and easy to understand, such as the Common Neighbors index (CN), it considers that the more the more similar the two nodes. Based on that, in order to improve the accuracy of the prediction algorithm, Adamic–Adar (AA) [39] and Resource Allocation (RA) [19] index have been proposed, both AA and RA index refine the simple counting of common neighbors by assigning the less connected neighbors more weights. The main difference between CN and RA is that RA index punishes the high-degree common neighbors more heavily than AA. Meanwhile, people tend to hide their true interests in the online social network [40]. In order to better explain this phenomenon, it is necessary to focus on other aspects to investigate the similarity, and in this paper, we consider this similarity call structural similarity since it is based solely on the network structure. The structural similarity indices can be classified in various ways, such as using local information, regular equivalence and structural equivalence [41]. Those prediction methods are based on similarity, some studies focus on the attributive similarity while others focus on the structural similarity [38].

However, one link prediction method may work very well for some networks while failing for others. One possible reason for this phenomenon is that there is the huge difference between different networks in both attributive aspect and structural aspect. In order to investigate the predictability of each network, Lü has proposed a fancy method to measure the predictability of networks via structural perturbation method [32]. Qi et al. proposed a novel method to predicting missing links by biased cross-network sampling [42]. Moreover, Shannon entropy [43] is widely used to measure the uncertainty in information science. Furthermore, Xu et al. have proposed the link prediction method based on the path entropy [44], and the entropy-based method was also applied in weighted networks [45]. Shang et al. have proposed a novel method to model the evolving networks based on the past structure [46].

To sum up, most of existing link prediction method only consider either attributive similarity or structural similarity. Based on the previous studies, this paper proposed a new method to fuse those two parts. Moreover, link predictability becomes more and more important for predicting missing links, however, most of the existing method do not consider the link predictability, this paper proposed a fancy method to measure the predictability of node based on Shannon entropy and Dempster–Shafer evidence theory.

2.1. Introduction to Dempster–Shafer evidence theory

Theory of the evidence is an efficient tool to handle uncertain information between many information sources [29,30]. Dempster–Shafer theory is often applied to uncertain decision making [47–49], fuzzy information processing [50], Z-numbers modeling [51,52], D numbers modeling [53–55], information fusion [56–58] and other hot fields [59]. In order to get a better explanation of this method, some basic concepts are introduced follows.

The frame of discernment (FOD) is used to represent the set of all observed events. Let ϕ be the set of mutually exclusive and collectively exhaustive events E_i, namely

$$\phi = \{E_1, E_2, \ldots, E_i, \ldots, E_n\}. \tag{1}$$