Journal of Computer Applications
计算机应用,2020,40(1):28-35
ISSN 1001-9081
CODEN JYIIDU
2020-01-10
http://www.joca.cn

文章编号:1001-9081(2020)01-0028-08
DOI:10.11772/j.issn.1001-9081.2019061008

融合聚集系数的链接预测方法

刘昱阳,李龙杰*,单　娜,陈晓云
(兰州大学 信息科学与工程学院,兰州 730000)
(*通信作者电子邮箱 ljli@lzu.edu.cn)

摘　要:许多基于网络结构信息的链接预测算法利用节点的聚集程度评估节点间的相似性,进而执行链接预测;然而,该类算法只注重网络中节点的聚集系数,没有考虑预测节点与共同邻居节点之间的链接聚集系数对节点间相似性的影响。针对上述问题,提出了一种融合节点聚集系数和非对称链接聚集系数的链接预测算法。首先,计算共同邻居节点的聚集系数,并利用共同邻居节点对应的两个非对称链接聚集系数计算该预测节点的平均链接聚集系数;然后,基于 Dempster-Shafer 证据理论将两种聚集系数进行融合生成一个综合性度量指标,并将该指标应用于中间概率模型(IMP),得到一个新的节点相似性指标(IMP_DS)。在 9 个网络数据上的实验结果表明,该算法的受试者工作特征(ROC)的曲线下方面积(AUC)与精度值(Precision)优于共同邻居(CN)、Adamic-Adar(AA)、资源分配(RA)指标和基于共同邻居的中间概率模型(IMP_CN)。

关键词:链接预测;复杂网络;Dempster-Shafer 理论;聚集系数;相似性指标
中图分类号:TP391　　**文献标志码**:A

Link prediction method fusing clustering coefficients

LIU Yuyang, LI Longjie*, SHAN Na, CHEN Xiaoyun
(*School of Information Science and Engineering, Lanzhou University, Lanzhou Gansu 730000, China*)

Abstract: Many network structure information-based link prediction algorithms estimate the similarity between nodes and perform link prediction by using the clustering degree of nodes. However, these algorithms only focus on the clustering coefficient of nodes in network, and do not consider the influence of link clustering coefficient between the predicted nodes and their common neighbor nodes on the similarity between nodes. Aiming at the problem, a link prediction algorithm combining node clustering coefficient and asymmetric link clustering coefficient was proposed. Firstly, the clustering coefficient of common neighbor node was calculated, and the average link clustering coefficient of the predicted nodes was obtained by using two asymmetric link clustering coefficients of common neighbor node. Then, a comprehensive measurement index was obtained by fusing these two clustering coefficients based on Dempster-Shafer(DS) theory, and by applying the index to Intermediate Probability Model (IMP), a new node similarity index, named IMP_DS, was designed. The experimental results on the data of nine networks show that the proposed algorithm achieves performance in terms of Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) and Precision in comparison with Common Neighbor (CN), Adamic-Adar (AA), Resource Allocation (RA) indexes and InterMediate Probability model based on Common Neighbor (IMP_CN).

Key words: link prediction; complex network; Dempster-Shafer (DS) theory; clustering coefficient; similarity index

0　引言

链接预测[1]是复杂网络分析中的重要研究方向,得到了越来越多的关注。链接预测根据网络中已知信息预测网络中丢失的链接或者未来可能出现的链接,在网络分析中起着非常重要的作用,可以用于指导生物实验、重建网络结构以及模拟网络演化等。链接预测在许多实际问题中都有很高的应用价值,不同领域的学者都可以利用其作为工具辅助本领域的研究。例如,在生物学领域,生物学家可以利用链接预测筛选潜在的蛋白质相互作用关系[2]和进行脑功能网络研究[3],能够减少实际实验的次数以及降低实验成本。对于在线社交网络[4]、电子商务网络[5]和航空运输网络[6],都可以利用链接预测来增加其商业价值。在在线社交网络中,链接预测可以发现用户的潜在朋友[7],并通过把预测结果推荐给用户的方式来增加用户关联,同时用于提高用户的活跃度与忠诚度。

截至目前,学者们提出了大量基于相似性的链接预测方法,这些方法给节点对分配相似性分数,利用所分配的分数估计两个节点间存在链接的可能性,节点间的相似性分数越高,它们之间存在链接的可能性就越大。基于网络结构特征评估节点间的相似性是目前的一个主要研究方向。共同邻居

收稿日期:2019-06-14;**修回日期**:2019-09-18;**录用日期**:2019-09-23。
基金项目:国家自然科学基金青年基金资助项目(61602225);中央高校基本科研业务费专项(lzujbky-2019-90)。
作者简介:刘昱阳(1996—),男,山西临汾人,硕士研究生,主要研究方向:网络信息挖掘、链接预测;　李龙杰(1981—),男,河南商丘人,工程师,博士,CCF 会员,主要研究方向:数据挖掘、机器学习、网络信息挖掘;　单娜(1996—),女,河北保定人,硕士研究生,主要研究方向:网络信息挖掘、链接预测;　陈晓云(1954—),女,吉林长春人,教授,博士生导师,CCF 高级会员,硕士,主要研究方向:数据挖掘、大数据分析、人工智能、网络信息挖掘。

(Common Neighbors,CN)指标[8]是其中最简单的一个,它基于两个节点共同邻居的数量进行预测,共同邻居数量越多,这两个节点间存在链接的可能性就越高。相关方法还有考虑对大度节点进行惩罚的资源分配(Resource Allocation, RA)指标[9]以及 AA(Adamic-Adar)指标[10]等。上述方法基于节点的共同邻居信息,而局部路径(Local Path, LP)指标[9]、Katz 指标[11]等考虑节点间的路径信息。除此之外还有基于随机游走的相似性指标,如平均通勤时间(Average Commute Time, ACT)指标[12]、基于随机游走的余弦相似性(Cos +)指标[13]、有重启的随机游走(Random Walk with Restart, RWR)指标[14]和 SimRank 指标[15]等。以及基于网络局部随机游走的 LRW(Local Random Walk)指标[16]和 SRW(Superposed Random Walk)指标[16]。

节点的聚集系数是一种常用的网络结构信息,用于度量节点的邻居之间链接的密度。许多链接预测模型使用节点的聚集系数来评估该节点的两个邻居之间存在链接的概率。基于聚集系数的链接预测方法(Clustering Coefficient for Link Prediction,CCLP)[17]将两个节点的共同邻居的聚集系数之和作为这两个节点的相似度值。局部朴素贝叶斯(Local Naive Bayes, LNB)模型[18]利用贝叶斯分类器理论计算节点间存在链接的可能性。该模型认为不同的邻居对相似性的计算可能有不同的贡献,并使用邻居的聚集系数来表示其贡献。中间概率(InterMediary Probability, IMP)模型[19]是一种广义的概率评估模型,它可以根据节点间的不同特征来评估其存在概率的可能性。IMP_CN(InterMediate Probability model based on Common Neighbor)是基于 IMP 模型衍生的链接预测算法[19],该算法将节点间的共同邻居作为特征,邻居的聚集系数作为中间概率。最近,Wu 等[20]定义了非对称链接聚集系数(Asymmetric Link Clustering Coefficient, ALCC),该聚集系数计算经过一条链接的三角形的概率。与文献[21]中提出的链接聚集系数不同,ALCC 将链接的一个端点定义为要预测链接的节点,另一个端点为邻居节点。用 ALCC 替换了 CCLP 方法、LNB 模型中节点的聚集系数,提升了链接预测的精度[20]。

节点的聚集系数与非对称链接聚集系数从不同的角度度量了两个节点间存在链接的可能性。本文考虑将两者进行结合以得到一个综合的度量指标,并使用该指标评估节点间存在链接的可能性。本文方法使用 Dempster-Shafer(DS)证据理论[22]将两种聚集系数进行融合,并且将融合后的度量指标引入到 IMP 模型[19]中设计了一个新的链接预测方法。为验证本文所提方法的性能,在多个网络数据上进行了实验,结果表明,与其他方法相比,本文提出的方法取得了较好的预测效果。

1　相关工作

1.1　IMP_CN 算法

IMP 算法是一种广义的概率模型,可以利用不同的网络特征评估节点间存在链接的可能性,IMP 模型公式如式(1):

$$p(l_{\Omega}^{xy}) = 1 - \prod_{w_i \in \Omega}(1 - p(l_{w_i}^{xy})) \tag{1}$$

其中:l_{Ω}^{xy} 表示在给定特征集的情况下,顶点 x 和 y 之间存在链接的事件;$p(l_{\Omega}^{xy})$ 表示这一事件发生的概率。Ω 可以被赋予不同的含义,当 Ω 为 x、y 的共同邻居集合时,可以得到与共同邻居有关的 IMP 算法如式(2):

$$S_{xy} = 1 - \prod_{z \in O_{xy}}(1 - p(l_z^{xy})) \tag{2}$$

在式(2)中,$p(l_z^{xy})$ 表示在已知共同邻居 z 的情况下,x 和 y 之间存在链接的概率。IMP_CN 算法使用 z 的聚集系数代替 $p(l_z^{xy})$,定义如下:

$$S_{xy} = 1 - \prod_{z \in O_{xy}}(1 - p(c_z)) \tag{3}$$

1.2　CN 指标

CN 指标认为:两个不连接的节点如果有更多的共同邻居,则它们更倾向于连边。CN 指标定义两个节点 x 和 y 的相似性为其共同邻居的数量,即:

$$S_{xy} = |O_{xy}| \tag{4}$$

1.3　AA 指标

AA(Adamic-Adar)指标认为度小的共同邻居的贡献大于度大的共同邻居,因此为每个邻居节点赋一个权重值。该权重等于该节点的度的对数的倒数,其定义为:

$$S_{xy} = \sum_{z \in O_{xy}} \frac{1}{\lg k_z} \tag{5}$$

1.4　RA 指标

RA(Resource Allocation)指标受网络中资源分配过程的启发,考虑网络中不相连的两个节点 x 和 y,从 x 可以传递一些资源到 y,在这个过程中,共同邻居就成为资源传递的媒介。假设每个媒介将得到的资源平均分配给它的邻居,则 y 可以接收到的资源数就定义为节点 x 和 y 的相似度,即:

$$S_{xy} = \sum_{z \in O_{xy}} \frac{1}{k_z} \tag{6}$$

2　Dempster-Shafer 证据理论

Dempster-Shafer(DS)证据理论,以其表示和处理不确定信息的能力而闻名,DS 融合规则可以使命题得到不同来源信息的综合支持度。最早应用于专家系统中,用于根据多个信息源的不确定性信息[23]作出决策。例如,针对供应商选择问题,Liu 等[24]提出了一种模糊拓展分析网络方法,该方法利用 DS 证据理论解决专家判断中的认知不确定问题。DS 证据理论还应用于处理传感器信息融合系统中的不确定性,Ye 等[25]提出了一种基于灰色关联和 DS 证据理论的不确定性融合算法,解决了传感器之间的不一致性和监测环境的复杂性带来的不确定性问题。Jiang 等[26]将 Z-number 模型与 DS 证据理论进行结合,对传感器数据融合系统中的不确定性进行建模和处理,提高了故障检测的可靠性。此外,DS 证据理论还用于解决服务器集群负载不均衡的问题[27]。为了更好地解释 DS 证据理论,本文接下来介绍一些相关概念。

定义 1　识别框架(Frame Of Discernment, FOD)。给定一组基本的命题 $E_1, E_2, \cdots, E_i, \cdots, E_n$,命题 E_i 是 Φ 的基本元素,表示如下:

$$\Phi = \{E_1, E_2, \cdots, E_i, \cdots, E_n\} \tag{7}$$

要求 Φ 中的元素是相互排斥的并且是完备的。在 DS 理论中,Φ 被就称为识别框架。符号 2^{Φ} 表示 Φ 的幂集:

$$2^{\Phi} = \{\varnothing, \{E_1\}, \{E_2\}, \cdots, \{E_n\}, \{E_1, E_2\}, \cdots, \Phi\} \quad (8)$$

其中 $\varnothing$ 表示空集。

定义 2 基本概率分配函数。在识别框架 Φ 上的基本概率分配函数是一个从 2^{Φ} 到[0,1]的映射函数,用于给各命题分配信任程度,记作 m:

$$m: 2^{\Phi} \rightarrow [0,1] \quad (9)$$

此函数满足如下性质:

$$m(\varnothing) = 0 \text{ 并且 } \sum_{A \in 2^{\Phi}} m(A) = 1, 0 \leqslant m(A) \leqslant 1$$

其中 $m(A)$ 反映对命题 A 的信任程度大小。

定义 3 Dempster 合成规则。给定两个独立的基本概率分配函数 m_1 和 m_2,Dempster 合成规则根据 m_1、m_2 产生一个新的基本概率分配函数,新的基本概率分配函数表示为 $m = m_1 \oplus m_2$,具体公式如下:

$$m(A) = \frac{1}{k} \sum_{B \cap C = A} m_1(B) m_2(C); A \neq \varnothing \quad (10)$$

$$k = \sum_{B \cap C \neq \varnothing} m_1(B) m_2(C) \quad (11)$$

Dempster 合成规则既满足结合律,又满足交换律:

$$m_1 \oplus m_2 = m_2 \oplus m_1 \quad (12)$$

$$(m_1 \oplus m_2) \oplus m_3 = m_1 \oplus (m_2 \oplus m_3) \quad (13)$$

3 本文方法

在链接预测中,节点聚集系数和非对称链接聚集系数分别从不同的角度定义了共同邻居对两个节点之间是否存在链接的评估。本文利用 DS 证据理论将两者进行融合得到一个综合性度量指标,利用该指标去评估节点间存在链接的概率。最后将融合后的度量指标与 IMP 模型相结合,设计了一个新的链接预测方法,记为 IMP_DS。接下来,首先对两种聚集系数进行介绍,然后给出 IMP_DS 方法的流程,并通过例子演示了 IMP_DS 方法的计算过程。

3.1 节点聚集系数

聚集系数用于衡量网络中节点的聚集程度,其定义建立在网络中的“三角形”结构之上。节点的聚集系数定义为该节点与其邻居之间组成的三角形的个数与所有可能的三角形个数之比。给定节点 z,其聚集系数的计算如式(12)所示:

$$C_z = \frac{2N_{\triangle}}{k_z(k_z - 1)} \quad (14)$$

其中:$N_{\triangle}$ 表示节点 z 与其邻居之间的三角形个数;k_z 表示节点 z 的度,$k_z(k_z - 1)/2$ 表示最大可能的三角形个数。

3.2 非对称性链接聚集系数

非对称链接聚集系数[20]的定义原理与节点的聚集系数相似,其定义为通过一条链接的三角形个数除以可能的最大三角形个数。这里,最大三角形个数只与节点对中的某一点相关,这个点为共同邻居节点。给定节点 x 与 y,z 是它们的一个共同邻居,链接(x,z) 的非对称聚集系数定义为:

$$LC_{x,z} = \frac{|O_{xz}|}{k_z - 1} \quad (15)$$

其中:O_{xz} 表示节点 x 和 z 的共同邻居集合;$|O_{xz}|$ 表示集合 O_{xz} 中元素数量。

式(13)表明,$LC_{x,z}$ 是非对称的,只在节点 x 与节点 z 的度相同时 $LC_{x,z}$ 与 $LC_{z,x}$ 才相等。本文使用的链接聚集系数分别为 $LC_{x,z}$ 与 $LC_{y,z}$。

3.3 IMP_DS 方法

给定节点 x 与 y,z 为它们的一个共同邻居。C_z、$LC_{x,z}$ 和 $LC_{y,z}$ 从不同的角度度量了 x、y 之间存在链接的概率。本文将三种聚集系数进行融合得到一个新的度量指标,然后将融合后的指标引入 IMP 模型中,设计一个新的链接预测方法。本文方法包含三步,具体的过程介绍如下。

1) 首先,将邻居 z 的两个非对称链接聚集系数相结合,得到 z 的平均链接聚集系数 LC_z,定义如下:

$$LC_z = \frac{LC_{x,z} + LC_{y,z}}{2} \quad (16)$$

2) 然后,利用 DS 证据理论融合 C_z 与 LC_z。节点 x 与 y 存在两个假设,即节点 x、y 之间存在链接(l_z^{xy})与节点 x、y 之间不存在链接($\overline{l_z^{xy}}$)。根据 DS 证据理论得到识别框架 $FOD = \{l_z^{xy}, \overline{l_z^{xy}}\}$。这里,利用邻居 z 的节点聚集系数 C_z 与链接聚集系数 LC_z 为该识别框架定义两个基本概率分配函数,分别表示为 m_c 和 m_{lc},具体定义如下:

$$m_c(\{l_z^{xy}\}) = C_z \quad (17)$$

$$m_c(\{\overline{l_z^{xy}}\}) = 1 - C_z \quad (18)$$

以及

$$m_{lc}(\{l_z^{xy}\}) = LC_z \quad (19)$$

$$m_{lc}(\{\overline{l_z^{xy}}\}) = 1 - LC_z \quad (20)$$

定义 m_f 为融合后的基本概率分配函数,则:

$$m_f(\{l_z^{xy}\}) = m_c(\{l_z^{xy}\}) \oplus m_{lc}(\{l_z^{xy}\}) \quad (21)$$

$$m_f(\{\overline{l_z^{xy}}\}) = m_c(\{\overline{l_z^{xy}}\}) \oplus m_{lc}(\{\overline{l_z^{xy}}\}) \quad (22)$$

$m_f(\{l_z^{xy}\})$ 与 $m_f(\{\overline{l_z^{xy}}\})$ 可以根据式(10)进行计算。实际上,可以得出 $m_f(\{l_z^{xy}\}) = 1 - m_f(\{\overline{l_z^{xy}}\})$。

3)最后,结合基本概率分配函数 m_f 和 IMP 模型设计新的链接预测方法 IMP_DS。与 IMP_CN 一样,IMP_DS 采用节点间的共同邻居作为特征集(见式(2))。在 IMP_DS 中,使用 $m_f(\{l_z^{xy}\})$ 代替式(2)中的 $p(l_z^{xy})$ 得到新的计算公式,定义如下:

$$S_{xy}^{\text{IMP\_DS}} = 1 - \prod_{z \in O_{xy}} (1 - m_f(\{l_z^{xy}\})) \quad (23)$$

接下来通过一个例子描述 IMP_DS 方法的计算过程。

例 1 利用 IMP_DS 算法计算节点的相似性。在图 1 所示的网络中,节点对(x,y) 有 4 个共同邻居,分别是 z_1,z_2,z_3,z_4。使用 IMP_DS 算法评估 x、y 之间的相似性,首先计算 4 个邻居的节点聚集系数和链接聚集系数,结果如下:

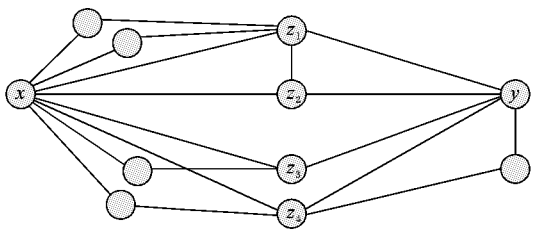

图 1 描述 IMP_DS 计算过程的示意网络

Fig. 1 Schematic network used to show computation process of IMP_DS

$$C_{z_1} = \frac{2}{5}, C_{z_2} = \frac{2}{3}, C_{z_3} = \frac{1}{3}, C_{z_4} = \frac{1}{3}$$

$$LC_{x,z_1}=\frac{3}{4},LC_{y,z_1}=\frac{1}{4},LC_{z_1}=\frac{1}{2}$$

$$LC_{x,z_2}=\frac{1}{2},LC_{y,z_2}=\frac{1}{2},LC_{z_2}=\frac{1}{2}$$

$$LC_{x,z_3}=\frac{1}{2},LC_{y,z_3}=0,LC_{z_3}=\frac{1}{4}$$

$$LC_{x,z_4}=\frac{1}{3},LC_{y,z_4}=\frac{1}{3},LC_{z_4}=\frac{1}{3}$$

之后对每一个共同邻居的节点聚集系数与链接聚集系数进行融合,得到相应的融合概率,结果如下:

$$m_f(\{l_{z_1}^{xy}\})=\frac{2}{5},m_f(\{l_{z_2}^{xy}\})=\frac{2}{3}$$

$$m_f(\{l_{z_3}^{xy}\})=\frac{1}{7},m_f(\{l_{z_4}^{xy}\})=\frac{1}{5}$$

将融合后的概率代入 IMP_DS 的计算公式中,得到节点对(x,y)相似性分数为:

$$S_{xy}^{\text{IMP\_DS}}=0.8629$$

4　实验数据集与评价指标

4.1　数据集

本文选取了 9 个真实网络进行实验及分析,网络的简单介绍如下。

1) Florida[28]:弗洛里达海湾雨季的食物链网络。

2) Word[29]:小说《大卫·科波菲尔》中常见形容词和名词的邻接网络。

3) Karate[30]:70 年代美国一所大学空手道俱乐部 34 名成员之间的友谊网络。

4) Cypwet[31]:赛普拉斯海湾雨季食物链网络。

5) Jazz[32]:爵士乐音乐家之间的协作网络。

6) Celegansneural(CE)[33]:线虫 Caenorhabditis elegans 的神经网络。

7) Polblogs(PB)[34]:政治博客网络。

8) Yeast[29]:酵母蛋白质相互作用网络。

9) Lesmis[35]:小说《悲惨世界》中人物的同时出现的网络。

表 1 展示了 9 个网络的基本拓扑结构,其中:$|V|$ 表示网络的节点数量,$|E|$ 表示网络中链接数量,C 表示网络的平均聚集系数,r 表示网络的同配系数,k 表示节点的平均度,d 表示平均最短距离,N_d 表示网络密度。

表 1　9 个网络的基本拓扑特征

Tab. 1　Basic topological features of 9 networks

网络	$\|V\|$	$\|E\|$	C	r	k	d	N_d
Florida	128	2 061	0.333	−0.114	32.422	1.776	0.253 6
Word	112	425	0.173	−0.129	7.589	2.536	0.068 4
Karate	34	78	0.571	−0.476	4.588	2.408	0.139 0
Cypwet	71	612	0.500	−0.277	17.239	1.785	0.246 3
Jazz	198	2 742	0.617	0.020	27.697	2.235	0.140 6
CE	297	2 148	0.292	−0.163	14.465	2.455	0.048 9
PB	1 222	16 714	0.320	−0.221	27.355	2.738	0.022 4
Yeast	2 224	6 609	0.138	−0.105	5.943	4.376	0.002 7
Lesmis	77	254	0.573	−0.165	6.597	2.641	0.086 8

4.2　评价指标

本文采用受试者工作特征(Receiver Operating Characteristic, ROC)曲线下方面积(Area Under the ROC Curve, AUC)[36]与精度值(Precision)[37]两种指标衡量链接预测算法的性能。实验中,将网络中的链接集合随机划分为训练集 E_{tr} 与测试集 E_{ts},其满足:

$$E_{\text{tr}}\cup E_{\text{ts}}=E \tag{24}$$

$$E_{\text{tr}}\cap E_{\text{ts}}=\varnothing \tag{25}$$

AUC 是一种依靠整体排名结果的度量,类似于概率。具体定义如下:进行 n 次独立比较,每次独立比较都从测试集和不存在的链接中分别取一条链接,链接预测算法根据训练集信息分别对两条链接进行评分。如果一个算法有较好的预测性能,测试集中链接的对应指标分数应该比不存在的链接的分数要高。因此,假设在 n 次独立比较中,测试集链接比不存在链接拥有更高分数 n' 次,两者拥有相同分数 n'' 次,则对应 AUC 的计算公式如下:

$$AUC=\frac{n'+0.5n''}{n} \tag{26}$$

AUC 值越高,链接预测算法的预测准确度越高。随机预测的 AUC 值约等于 0.5,因此 AUC 大于 0.5 的程度表明了相应算法在多大程度上比随机预测的方法更精确。

Precision 定义为将训练集与网络中所有不存在链接按照相似性分数进行降序排列,计算前 L 个链接中属于训练集的链接所占比例。如果排名前 L 的链接中有 l 个属于测试集,则 *Precision* 计算公式为:

$$Precision=\frac{l}{L} \tag{27}$$

5　实验结果及分析

以 AUC 与 Precision 为衡量指标,在 9 个真实网络中测试 IMP_DS 算法的预测效果,具体结果分为两部分:一是在不同网络中 IMP_DS 与其他相似性指标的对比结果分析;二是 IMP_DS 性能提升的原因分析。

5.1　与其他相似性指标对比

表 2 给出了 CN、AA、RA、IMP_CN 以及 IMP_DS 5 种算法在各个网络上的 *AUC* 与 *Precision* 的实验结果,表中加粗字体表明效果最好。两个表中的结果均为 50 次独立实验的平均值,每次实验中,原始网络被随机地划分为一个训练集和一个测试集,其中训练集占 90% 的链接,测试集占 10% 的链接。表 2 中的 *Precision* 是取 $L=10$ 时的实验结果。

从表 2 中可以看出,IMP_DS 算法在 Florida、Word、Cypwet、CE 和 PB 5 个网络上取得最好的 *AUC* 结果。在 Karate 上,RA 的 *AUC* 值最高,IMP_DS 第二。在其他 3 个网络上,IMP_DS 的性能与 IMP_CN 非常接近。结果表明融合两种聚集系数的方法在 IMP 模型上是可行的,并且比单一的节点聚集系数的效果更好。特别地,在 Florida 和 Cypwet 两个网络上,与其他算法相比,IMP_DS 的 *AUC* 结果提升非常明显。从表 1 中可以看到:Florida 和 Cypwet 两个网络的密度非常高,是两个非常稠密的网络,因此,两个网络的节点聚集系数和链接聚集系数都非常高,通过融合节点聚集系数和链接聚集系数能够显著提高链接预测的性能。相反地,在 Yeast 这个特别稀疏的网络上,IMP_DS 以及 IMP_CN 两个方法的 *AUC* 值均低于 CN、AA 和 RA 三个方法。这是因为稀疏网络上的节点间的共同邻居数据很少,并且节点的聚集系数和链接聚集系数的值也变得非常低,降低了 IMP 模型的性能[19]。

表 2 中的 Precision 结果再次证明 IMP_DS 方法的预测精度高于对比算法。例如,在 Florida 网络上,IMP_DS 方法的预测精度相比 CN、AA、RA 和 IMP_CN 算法分别提高了 130.9%、139.5%、169.4% 和 106.4%,因此本文认为融合共同邻居的节点聚集系数与非对称链接聚集系数能够明显提高 IMP 模型的预测精度。

表 2 9 个网络上的 *AUC* 和 *Precision* 结果

Tab. 2 *AUC* and *Precision* values on 9 networks

网络	*AUC*					*Precision*				
	CN	AA	RA	IMP_CN	IMP_DS	CN	AA	RA	IMP_CN	IMP_DS
Florida	0.6099	0.6116	0.6147	0.6437	**0.7313**	0.1680	0.1620	0.1440	0.1880	**0.3880**
Word	0.6693	0.6697	0.6677	0.6753	**0.6915**	0.1000	0.0960	0.0880	**0.1180**	0.1160
Karate	0.6601	0.6973	**0.7085**	0.6765	0.7057	0.0900	0.0880	0.0860	**0.0940**	0.0780
Cypwet	0.7389	0.7510	0.7596	0.7630	**0.8082**	0.1920	0.1740	0.1600	0.1820	**0.3200**
Jazz	0.9540	0.9611	**0.9703**	0.9599	0.9562	0.9100	0.9140	0.9720	**0.9780**	0.9660
CE	0.8451	0.8606	0.8652	0.8618	**0.8692**	**0.1300**	0.0950	0.0850	0.1100	0.1150
PB	0.9175	0.9203	0.9221	0.9202	**0.9247**	0.6200	0.3950	0.2150	0.6700	**0.7450**
Yeast	0.7044	**0.7048**	0.7046	0.6984	0.6931	0.2300	0.2250	0.1500	0.4300	**0.4600**
Lesmis	0.8861	0.8943	**0.8947**	0.8841	0.8789	0.5250	0.6100	**0.7850**	0.6000	0.6250

接下来,在 9 个网络上选取不同比例的训练集进行实验,观察 *AUC* 的结果与变化趋势。本实验的结果也是 50 次独立实验的平均值。图 2 描述了从 E 中选取不同比例训练集 E_{tr}(从 0.7 到 0.9)时各预测方法 *AUC* 的变化情况。从图 2 中可以看出,在不同比例训练集的情况下,IMP_DS在超过一半的网络上都获得了较高的 *AUC* 值。观察 *AUC* 的变化趋势发现,当训练集的比例从 0.7 上升到 0.9 时,*AUC* 值呈明显上升趋势。这是因为,训练集 E_{tr} 的比例越大,为训练提供的信息越多,预测越准确;相反,低比例的 E_{tr} 会增加链接预测的难度[38]。

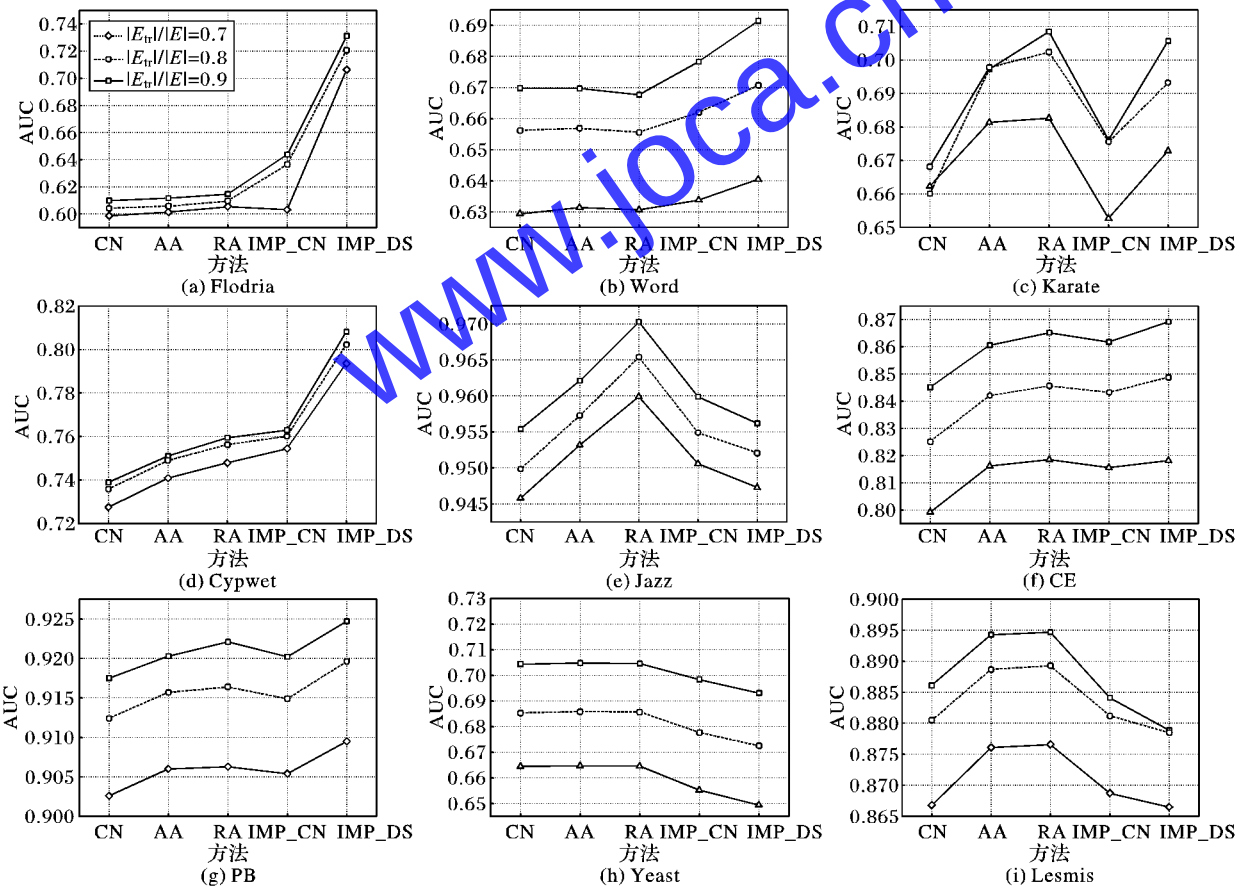

图 2 不同比例训练集时的 *AUC* 结果

Fig. 2 *AUC* values under different proportions of training set

图 3 描述训练集 E_{tr} 的比例从 0.7 增长到 0.9 时 *Precision* 的变化趋势,L 的值同样设置为 10。

从图 3 中可以看出,与 *AUC* 相比,*Precision* 随着训练集比例变化呈现相反的变化趋势,即当比例从 0.7 上升到 0.9 时,*Precision* 值呈现下降趋势。这是因为训练集 E_{tr} 的减少会导致 *AUC* 定义中 n' 与 n'' 变小,从而降低 *AUC* 的值[39],但是,随着测试集 E_{ts} 的提高(训练集 E_{tr} 减小),获得相关信息的可能性增加,使得发现缺失链接更容易[39]。比较各个方法的 *Precision* 值,整体而言,IMP_DS 在不同比例训练集上的性能均优于对比方法。

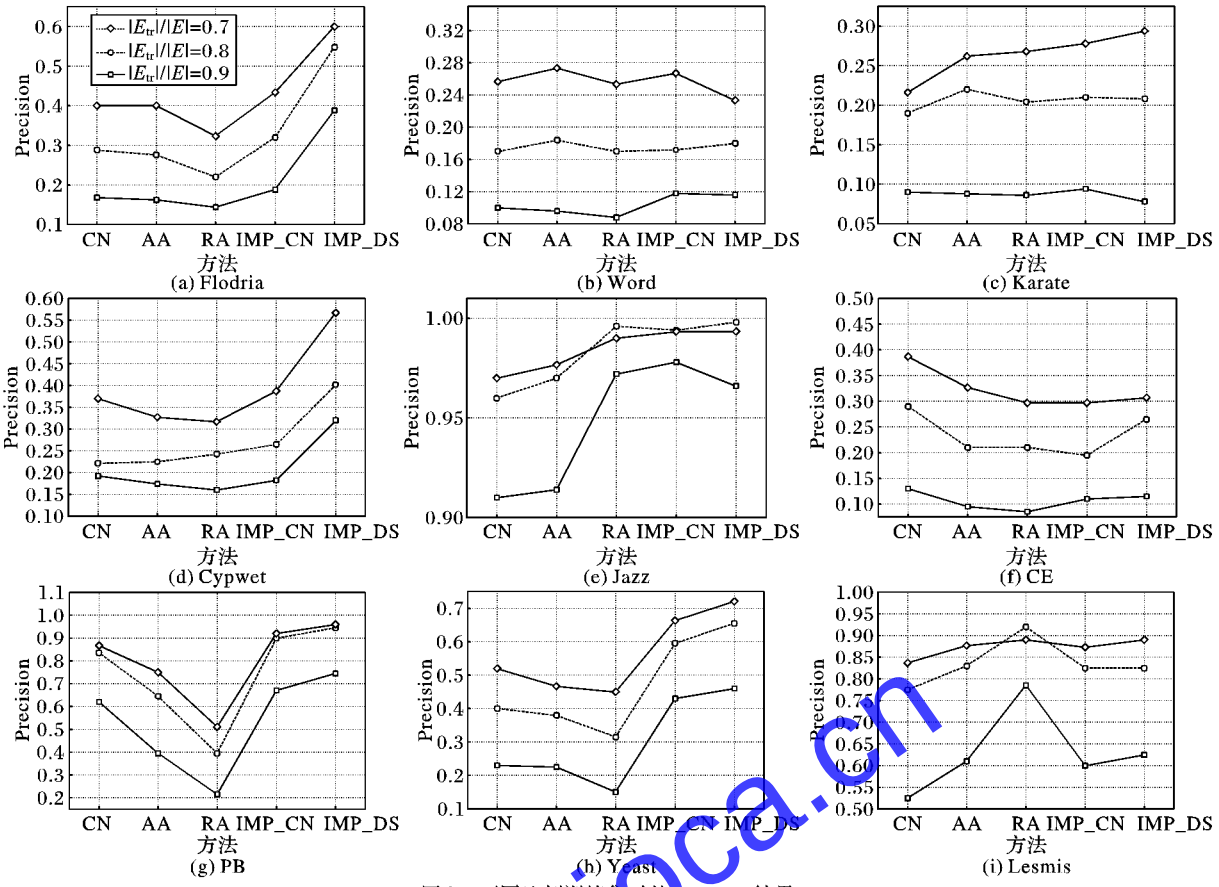

图 3　不同比例训练集时的 *Precision* 结果

Fig. 3　*Precision* values under different proportions of training set

图 4 显示了在取不同 L 值时每个方法的 *Precision* 值及变化趋势。图 4 中,训练集与测试集的比例为 9:1,结果仍然是 50 次独立实验的平均值。这里,只给出了在 6 个较大网络上的实验结果。从图 4 中可以看出,在这 6 个网络上,IMP_DS 的性能具有明显的优势,尤其是在 Flodria、Cypwet 和 PB 3 个网络中,其 *Precision* 值显著高于对比方法。在不同的网络上,其他方法的排序随着 L 取值改变有较大变化。例如,在 PB 网络上,随着 L 的改变各种方法的排序基本不变,但是在 Jazz 与 CE 网络上,4 种对比方法的排序有很大波动。另外,在大多数网络中,随着 L 值的增大,*Precision* 呈现逐渐下降的趋势。这是因为 L 的增加,使得发现丢失链接的概率降低,从而导致精度值降低[38]。

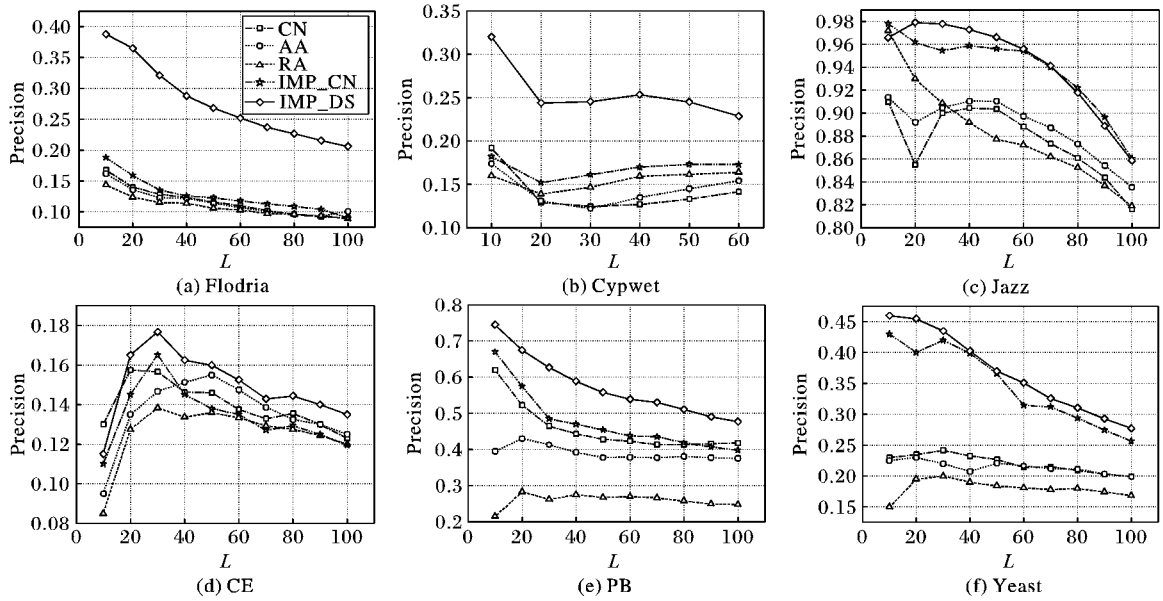

图 4　6 个网络上选取不同 L 时的 *Precision* 结果

Fig. 4　*Precision* values under different L on 6 networks

5.2 IMP_DS 性能提升原因分析

最后,通过实例分析的方式进一步研究 IMP_DS 性能提升的原因。参考文献[19]中的分析方法,图 5 选取了四个对比算法预测的前 100 条链接,并将这 100 条链接在不同算法的排名进行对比。本文实验中,将 PB 随机划分成一个训练集和一个测试集,训练集和测试集的比例是 9∶1。图 5 中,使用半对数坐标绘制了每一对算法预测的前 100 条链接的相对排序。以图 5(c)(d)子图为例,(c)子图表示 AA 预测的前 100 条链接在 IMP_DS 结果中的排序,(d)子图表示 IMP_DS 预测的前 100 条链接在 AA 结果中的排序。观察图中的结果可以发现,AA 预测的前 100 条链接中,41 条是正确的,59 条是错误的,而 IMP_DS 将这些错误结果中的大部分排在了 100~1 000。另一方面,IMP_DS 预测的前 100 条链接中,52 条是正确的,48 条是错误的,而 AA 将正确预测结果中的 20 条排在了 100 以外,因此,IMP_DS 能够取得比 AA 更高的预测精度。其他三个子图上的结果也与此类似。

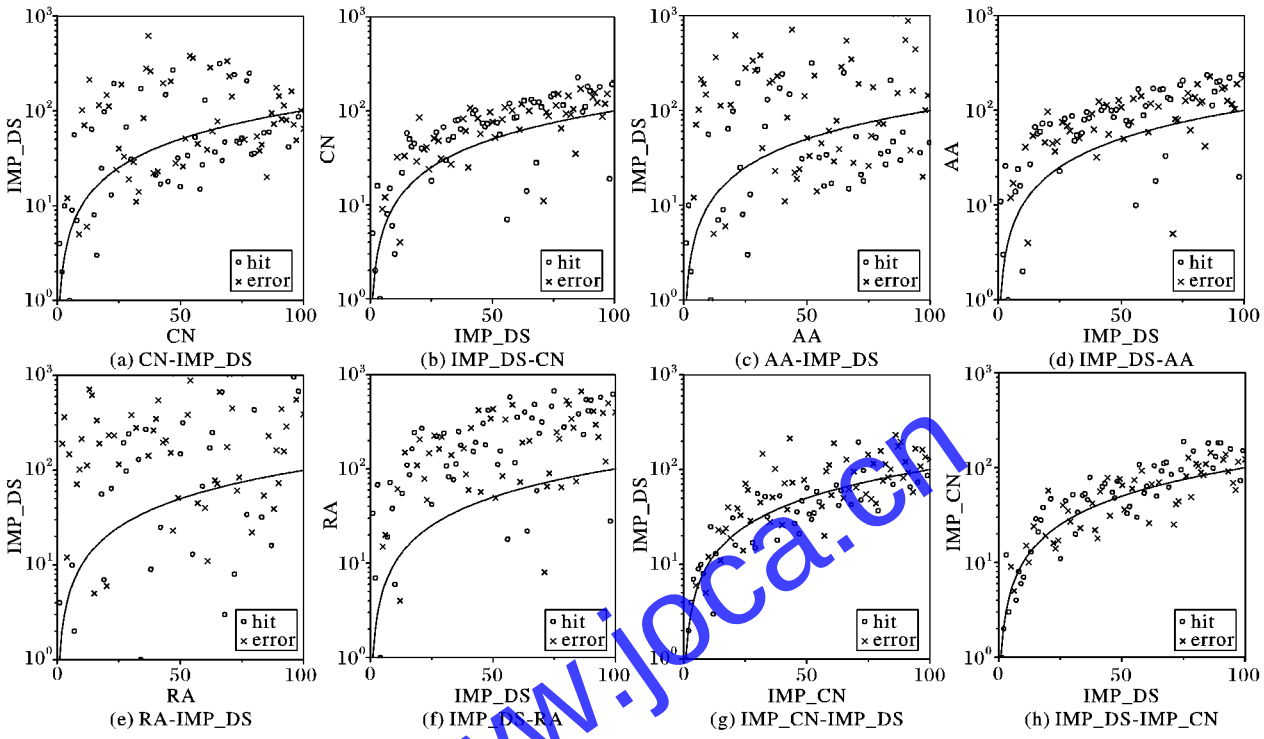

图 5 各算法在 PB 网络上预测的前 100 条链接的对比

Fig. 5 Comparison of top-100 predicted links of different algorithms on PB network

6 结语

针对许多基于网络结构信息的链接预测算法只考虑节点的聚集系数,而忽略了预测节点与共同邻居节点之间链接的聚集系数对链接预测影响的问题,本文提出了一种基于 Dempster-Shafer 证据理论,融合节点聚集系数和非对称链接聚集系数的链接预测算法。首先,针对每个共同邻居节点计算出对应的聚集系数和平均链接聚集系数;然后,将两种聚集系数进行融合得到一个综合性度量指标;最后将这个综合性度量指标应用于中间概率模型,得到一个新的节点间相似性指标。在 9 个真实网络数据上的实验结果表明,IMP_DS 方法具有较高的 AUC 与 Precision 值,可以用于复杂网络链接预测。尽管本文设计的融合两种聚集系数的链接预测算法取得了优秀的预测效果,但仍有许多问题待解决,例如可以进一步研究不同特征的融合以及具体融合过程对链接预测效果的影响。

参考文献(References)

[1] LÜ L, ZHOU T. Link prediction in complex networks: a survey[J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170.

[2] KOVÁCS I A, LUCK K, SPIROHN K, et al. Network-based prediction of protein interactions[J]. Nature Communications, 2019, 10(1): 1240.

[3] 丁超,赵海,司帅宗,等. 正常衰老的人脑功能网络演化模型[J]. 计算机应用, 2019, 39(4): 963-971. (DING C, ZHAO H, SI S Z, et al. Evolution model of normal aging human brain functional network[J]. Journal of Computer Applications, 2019, 39(4): 963-971.)

[4] LI D, ZHANG Y, XU Z, et al. Exploiting information diffusion feature for link prediction in Sina Weibo[J]. Scientific Reports, 2016, 6: No. 20058.

[5] CRONE S F, SOOPRAMANIEN D. Predicting customer online shopping adoption - an evaluation of data mining and market modelling approaches[EB/OL]. [2019-01-11]. https://pdfs.semanticscholar.org/fd1d/f142629dd35ade4260f3d456847330e86016.pdf.

[6] WANG J, MO H, WANG F, et al. Exploring the network structure and nodal centrality of China's air transport network: a complex network approach[J]. Journal of Transport Geography, 2011, 19(4): 712-721.

[7] GRABOWSKI A, KRUSZEWSKA N, KOSINSKI R A. Dynamic phenomena and human activity in an artificial society[J]. Physical Review E, 2008, 78(6 Pt 2): No. 066110.

[8] LIBEN-NOWELL D, KLEINBERG J. The link-prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031.

[9] ZHOU T, LÜ L, ZHANG Y. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4): 623-630.

[10] ADAMIC L A, ADAR E. Friends and neighbors on the Web[J]. Social Networks, 2003, 25(3): 211 – 230.
[11] KATZ L. A new status index derived from sociometric analysis[J]. Psychometrika, 1953, 18(1): 39 – 43.
[12] KLEIN D J, RANDIC M. Resistance distance[J]. Journal of Mathematical Chemistry, 1993, 12(1): 81 – 95.
[13] FOUSS F, PIROTTE A, RENDERS J M. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 355 – 369.
[14] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1/2/3/4/5/6/7): 107 – 117.
[15] JEH G, WIDOM J. SimRank: a measure of structural-context similarity[C]// Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002: 538 – 543.
[16] LIU W, LÜ L. Link prediction based on local random walk[J]. EPL (Europhysics Letters), 2010, 89(5): 58007.
[17] WU Z, LIN Y, WANG J, et al. Link prediction with node clustering coefficient[J]. Physica A: Statistical Mechanics and its Applications, 2016, 452: 1 – 8.
[18] LIU Z, ZHANG Q, LÜ L, et al. Link prediction in complex networks: A local naive Bayes model[J]. EPL (Europhysics Letters), 2011, 96(4): No. 48007.
[19] ZHANG X, PANG W, XIA Y. An intermediary probability model for link prediction[J]. Physica A: Statistical Mechanics and its Applications, 2018, 512: 902 – 912.
[20] WU Z, LIN Y, ZHAO Y, et al. Improving local clustering based top-L link prediction methods via asymmetric link clustering information[J]. Physica A: Statistical Mechanics and its Applications, 2018, 492: 1859 – 1874.
[21] WANG J, LI M, WANG H, et al. Identification of essential proteins based on edge clustering coefficient[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(4): 1070 – 1080.
[22] SHAFER G. A Mathematical Theory of Evidence[M]. Princeton: Princeton University Press, 1976: 1 – 32.
[23] KANG B, CHHIPI-SHRESTHA G, DENG Y, et al. Development of a predictive model for Clostridium difficile infection incidence in hospitals using Gaussian mixture model and Dempster-Shafer theory [J]. Stochastic Environmental Research and Risk Assessment, 2018, 32(6): 1743 – 1758.
[24] LIU T, DENG Y, CHAN F. Evidential supplier selection based on DEMATEL and game theory[J]. International Journal of Fuzzy Systems, 2018, 20(4): 1321 – 1333.
[25] YE F, CHEN J, LI Y, et al. Decision-making algorithm for multisensor fusion based on grey relation and DS evidence theory[J]. Journal of Sensors, 2016, 2016: No. 3954573.
[26] JIANG W, XIE C, ZHUANG M, et al. Sensor data fusion with Z-numbers and its application in fault diagnosis[J]. Sensors, 2016, 16(9): No. 1509.
[27] 邰滢滢, 庞影, 段苛苛, 等. 基于改进权重的 D-S 证据理论的动态负载平衡算法[J]. 计算机应用, 2018, 38(10): 2976 – 2981, 2989. (TAI Y Y, PANG Y, DUAN K K, et al. Dynamic algorithm of load balancing based on D-S evidence theory with improved weight[J]. Journal of Computer Applications, 2018, 38(10): 2976 – 2981, 2989.)
[28] ULANOWICZ R E, DEANGELIS D L. Network analysis of trophic dynamics in south Florida ecosystems[J]. US Geological Survey Program on the South Florida Ecosystem, 2005, 114: 45.
[29] NEWMAN M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Physical Review E, 2006, 74(3 Pt 2): No. 036104.
[30] ZACHARY W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 33(4): 452 – 473.
[31] ULANOWICZ R E, DEANGELIS D L. Network analysis of trophic dynamics in south Florida ecosystems[C]//Proceedings of South Florida Restoration Science Forum. Boca Raton: [s. n.], 1999: 114 – 115.
[32] GLEISER P M, DANON L. Community structure in jazz[J]. Advances in Complex Systems, 2003, 6(4): 565 – 573.
[33] WATTS D J, STROGATZ S H. Collective dynamics of 'small-world' networks[J]. Nature, 1998, 393(6684): 440 – 442.
[34] ISELLA L, STEHLÉ J, BARRAT A, et al. What's in a crowd? Analysis of face-to-face behavioral networks[J]. Journal of Theoretical Biology, 2011, 271(1): 166 – 180.
[35] GUIMERA R, DANON L, DIAZ-GUILERA A, et al. Self-similar community structure in a network of human interactions[J]. Physical Review E, 2003, 68(6 Pt 2): No. 065103.
[36] FAWCETT T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8): 861 – 874.
[37] HERLOCKER J L, KONSTAN J A, TERVEEN L G, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems, 2004, 22(1): 5 – 53.
[38] LI L, BAI S, LENG M, et al. Finding missing links in complex networks: a multiple-attribute decision-making method[J]. Complexity, 2018, 2018: No. 3579758.
[39] YANG J, ZHANG X. Predicting missing links in complex networks based on common neighbors and distance[J]. Scientific Reports, 2016, 6: No. 38208.

This work is partially supported by the Youth Program of National Natural Science Foundation of China (61602225), the Fundamental Research Funds for the Central Universities (lzujbky-2019-90).

LIU Yuyang, born in 1996, M. S. candidate. His research interests include network information mining, link prediction.

LI Longjie, born in 1981, Ph. D., engineer. His research interests include data mining, machine learning, network information mining.

SHAN Na, born in 1996, M. S. candidate. Her research interests include network information mining, link prediction.

CHEN Xiaoyun, born in 1954, M. S., professor. Her research interests include data mining, big data analysis, artificial intelligence, network information mining.