计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.181202433

基于特征向量局部相似性的社区检测算法

杨旭华　沈　敏
浙江工业大学计算机科学与技术学院　杭州 310023

摘　要　社区的发现和分析是复杂网络结构和功能研究中的一个热点。目前广泛应用的社区划分算法存在时间复杂度过高、社区核心数量无法准确量化、划分精度不高等问题。文中提出了一种基于特征向量局部相似性的社区检测算法 ELSC。该算法首先计算网络中每个节点的特征向量中心性,在此基础上提出了特征向量局部相似性(ELS)和特征向量吸引性(EA)指标。ELS 指标表示节点之间的相似性,用来形成初始社区,在同一个社区内部节点之间的相似性较高,在不同社区节点之间的相似性较低;EA 指标同时考虑了局部相似性和特征向量中心性的占比,表示节点之间的吸引性,用来优化初始社区,并在此基础上完成网络的社区划分。该算法由最值确定节点,避免了节点数量阈值不确定的问题。在 7 个真实网络上将所提算法与 6 种知名算法的模块度和标准化互信息两个指标进行综合比较,结果表明,该算法具有良好的准确性,并且具有较低的时间复杂度。

关键词:社区检测;特征向量中心性;特征向量局部相似性;特征向量吸引性
中图法分类号　TP391

Community Detection Algorithm Based on Local Similarity of Feature Vectors

YANG Xu-hua and SHEN Min
College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China

Abstract　Community discovery and analysis is a hot topic in the study of complex network structures and functions. At present, the widely used algorithm for community partitioning has some problems, such as high time complexity, inaccurate quantification of the number of community cores, and low partitioning accuracy. Therefore, this paper proposed a community detection algorithm ELSC based on local similarity of feature vectors. The algorithm first calculates the eigenvector centrality of each node in the network. On this basis, the eigenvector local similarity (ELS) and eigenvector attractiveness (EA) indicators were proposed. The ELS index indicates the similarity between nodes. To form the initial community, the similarity between the nodes within the same community is higher, and the similarity between different community nodes is lower. The EA index considers the local similarity and the eigenvector centrality ratio, indicating the node. The attraction is used to optimize the initial community and complete the community division of the network. The algorithm determines the node by the most value, avoiding the problem that the threshold number of nodes is uncertain. The modularity and standardized mutual information between the proposed algorithm and six well-known algorithms were compared on seven real networks. Numerical simulation results show that the algorithm has high accuracy and low time complexity.

Keywords　Community detection, Eigenvector centrality, Eigenvector local similarity, Eigenvector attractiveness

1　引言

近年来,复杂网络的研究已经在社会学、计算机科学、数学、生物学等[1-2]许多领域引起了广泛的关注。复杂网络中的社区结构可以揭示节点之间的潜在关系[3-4],同一个社区内部的节点相似度较高,不同社区的节点之间的相似度较低[5],关于社区结构划分的算法已经成为复杂网络领域的研究热点[6]。

目前,学者们在社区检测领域已经展开了广泛的研究。Newman 等提出用模块化函数 Q 来评估社区检测的效果[7-8];快速贪婪算法[9]用当前最优解对社区进行划分;层次聚类算法[10]在欧氏空间下研究基于接近度指标的聚类方法;标签传播算法[11-13]用已标记节点的标签信息预测未标记节点的标签信息,找出相似度更高的节点划分到一起,更接近正确分类;基于密度的算法[14-15]将密度大于阈值的点加到相邻类别以检测任意形状的聚类;基于随机游走的算法[16-17]从未标记顶点开始随机漫步,首次到达各类标记顶点的概率代表了未标记点归属于标记类的可能性,把最大的概率所在类的标签赋给未标记顶点,完成社团划分。上述算法一般需要迭代计算网络的某些数据,计算复杂度较高,因此寻求更低计算复杂度的社区检测算法是当前的一个重要研究方向。

CNM 算法基于节点重要性来检测社区[7-8]。CNM 算法用节点度数表示节点重要性,度数越高的节点越有可能被其

到稿日期:2018-12-27　返修日期:2019-04-24　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61773348);浙江省自然科学基金(LY17F030016)
This work was supported by the National Natural Science Foundation of China (61773348) and Zhejiang Natural Science Foundation (LY17F030016).
通信作者:杨旭华(xhyang@zjut.edu.cn)

他节点连接起来,因此度数被认为是中心性的标准之一[3,7]。其他一些计算节点中心性的方法也已经被广泛使用[29-30],例如页面排序、特征向量中心性、信息熵等。在基于节点重要性划分社区的算法中,一般首先确定中心节点即社区核心,然后通过模块性、标签等方法将其余节点划分到中心节点所在的社区,通过多次迭代完成社区划分。然而如何准确地确定中心节点是目前研究算法中的一个难点[18-20,23]。

基于局部相似性指标的社区检测算法在现实世界网络划分中的性能较好[22-25]。两个节点之间的相似性程度越高,它们属于同一个社区的可能性就越大[31-32]。相似性指标中有基于局部信息的相似性指标,包括基于共同邻居的相似性指标、偏好连接相似性指标、角色函数相似性指标;有基于路径的相似性指标,包括局部路径指标、Katz 指标、LHN-II 指标;还有基于随机游走[16,24]的相似性指标,包括全局随机游走指标、局部随机游走指标等。这些相似性指标可以结合标签传播算法[12,23],每个节点在迭代过程中获得一个受邻居影响的唯一标签,最终联系紧密的节点相互吸引,聚集成社区。然而这些方法有着明显的不足,如没有合理地思考节点本身在社团形成过程中不可或缺的局域影响力等。

为改善上述方法存在的问题,提升社团划分算法的性能,本文提出了一种基于特征向量局部相似性的社区检测算法(ELSC)。节点在网络中的重要性由特征向量中心性表示,该算法基于特征向量的局部相似性把网络节点汇聚成初始社区,然后基于特征向量吸引性优化社区。局部相似性的方法在计算复杂度上优于聚类算法。本文在 7 个真实网络上将所提算法与现有的 6 种知名算法进行对比,结果表明该方法在计算速度和精度上都有很好的表现。

2 相关工作

2.1 特征向量中心性 EC

对网络中某节点的所有邻居节点赋予一个"中心性值",但是并非所有邻居节点的值都是相同的,很多情况下,一个节点会由于连接到一些本身很重要的节点,从而使自身的重要性得到提升。这就是特征向量中心性的本质,计算网络中节点的特征向量中心性值时不是简单地为每个邻居节点赋一个值,而是根据该邻居节点的特征向量中心性值之和叠加计算该邻居节点的"中心性值"。节点 i 的特征向量中心性定义为:

$$EC(i)=a_1^{-1}\sum_j A_{ij}EC(j) \tag{1}$$

其中,$\boldsymbol{A}$ 为网络的邻接矩阵,A_{ij} 表示节点 i 和 j 之间的连接情况,若 i 和 j 相连,则 $A_{ij}=1$,否则 $A_{ij}=0$。a_1 表示矩阵 $\boldsymbol{A}$ 的特征值的最大值。对于所有的节点 i,设其特征向量中心性的初始值 $EC(i)=1$,将 $EC(i)$的初始向量值不断重复地与邻接矩阵 $\boldsymbol{A}$ 相乘,直至向量趋于稳定,从而得到 $EC(i)$的最终值。

2.2 模块度 Q

模块度也称模块化度量值,是目前常用的一种衡量网络社区结构强度的方法,最早由 Mark 提出[6]。模块度值的大小主要取决于网络中节点的社区分配 C,其值越接近 1,表示网络划分的社区结构的强度越高,对应的分区越精确。模块度的定义为:

$$Q=\frac{1}{2}\sum_{ij}\left[A_{ij}-\frac{k_ik_j}{2m}\right]\delta(C_i,C_j) \tag{2}$$

其中,$\boldsymbol{A}$ 为网络的邻接矩阵;m 是网络中所有边的数量;k_i 是节点 i 的度;$\delta(C_i,C_j)$表示 i 和 j 是否在同一社区,如果是则等于 1,否则等于 0。

2.3 标准化互信息 NMI

标准化互信息(NMI)[29]被广泛用于量化社区检测算法的准确性,其考虑信息理论概念,仅需要少量的附加信息就可以推断一个集群与另一个集群的相似性。令 $A=\{A_1,\cdots,A_a\}$,$B=\{B_1,\cdots,B_b\}$是给定的两个集群,并定义 H 是一个混淆矩阵,其中 H_{xy} 表示出现在生成社区 B_y 中的现实社区 A_x 的节点数量。标准化互信息 NMI 的定义为:

$$NMI(A,B)=\frac{-2\sum_{x=1}^{H_A}\sum_{y=1}^{H_B}H_{xy}\log(H_{xy}N/H_{x.}H_{.y})}{\sum_{x=1}^{H_A}H_{x.}\log(H_{x.}/N)+\sum_{y=1}^{H_B}H_{.y}\log(H_{.y}/N)} \tag{3}$$

其中,$H_A(H_B)$是分区 $A(B)$中的社区数,$H_{x.}(H_{.y})$是矩阵 H 中第 x 行(y 列)元素的总和,N 是网络节点的数量。如果 $A=B$,则 $NMI(A,B)=1$;如果 A 和 B 完全不同,则 $NMI(A,B)=0$。NMI 的值越大表示生成社区的结果越好。

3 基于特征向量局部相似性的社区检测算法

令 $G=(V,E)$为一个无向无权的网络,网络中有 n 个节点和 m 条边,其中 $V=\{v_1,v_2,\cdots,v_n\}$表示节点的集合,$E=\{e_1,e_2,\cdots,e_m\}$表示节点之间连边的集合。$\Gamma(v_i)$是节点 v_i 的邻居节点集合,包括节点 v_i。本文首先基于特征向量的局部相似性划分初始社区;然后基于特征向量吸引性优化初始社区,从而得到二次社区;最后将节点数量小于阈值的小社区合并到相应的邻近的较大社区,从而得到最终的社区结构。

3.1 特征向量的局部相似性(ELS)

本文定义任意相连节点 i 和 j 之间的邻居特征向量的局部相似性为:

$$ELS(i,j)=\frac{\sum_{u\in\Gamma_i\cap\Gamma_j}EC(u)}{\sum_{v\in\Gamma_i\cup\Gamma_j}EC(v)},j\in\Gamma_i,j\neq i \tag{4}$$

其中,$\Gamma_i\cap\Gamma_j$ 表示节点 i 和 j 的共同邻居集合,$\Gamma_i\cup\Gamma_j$ 表示节点 i 和 j 的联合邻居集合。节点 i 的最相似节点为连边(i,j)的 ELS 值取最大值 $\max\{ELS(i,j)\}$且 $j\in\Gamma_i$ 时的节点 j。本文把任一节点和它的最相似节点划分到同一个初始社区。

3.2 特征向量吸引性(EA)

本文综合考虑特征向量局部相似性和特征向量中心性的占比,定义任意节点 i 的任意邻居节点 j 的特征向量吸引性为:

$$EA(i,j)=\frac{\sum_{u\in\Gamma_i\cap\Gamma_j}EC(u)}{\sum_{v\in\Gamma_i\cap\Gamma_j}EC(v)}+\frac{EC(j)}{\sum_{v\in\Gamma_i\cap\Gamma_j}EC(v)},j\in\Gamma_i,j\neq i \tag{5}$$

其中,节点 j 的特征向量中心性值必须大于节点 i,否则 EA 的值为 0。节点 i 的最吸引节点为连边(i,j)的 EA 值取最大值 $\max\{EA(i,j)\}$且 $j\in\Gamma_i$ 时的节点 j。本文把任一节点和它的最吸引节点划分到同一个社区。

3.3 ELSC 算法

ELSC 算法的步骤如下。

(1)数据预处理。1)遍历网络中的每个节点 i,计算其度值和特征向量中心性值;2)遍历网络中的每一个相连节点对(i,j),计算其共同邻居集合 $\Gamma_i\cap\Gamma_j$ 和联合邻居集合 $\Gamma_i\cup\Gamma_j$。

(2)基于特征向量局部相似性获取初始社区。根据式(4)计算所有节点对(i,j)的特征向量局部相似性,得到最相似节点对集合P_1,本文把每一个最相似节点对划分到同一个社区;最相似节点之间有连边,把具有共同节点的连边划分到同一个社区,这样就形成了初始社区。

$$P_1=\begin{pmatrix} N_{s1} & N_{s2} & \cdots & N_{sw} \\ N_{s1c} & N_{s2c} & \cdots & N_{swc} \end{pmatrix}$$

其中,集合的第一行N_{sw}表示节点标号,N_{sw}小于或等于网络的节点数量n;集合的第二行N_{swc}表示第一行所示节点对应的特征向量局部相似性取最大值时的连边节点标号。

(3)基于特征向量吸引性优化初始社区,从而得到二次社区。根据式(5)计算所有节点对(i,j)的特征向量吸引性,得到最吸引节点对集合。对集合中节点对的EA值求平均,将低于平均值的节点对直接删除以避免过度整合社区,从而得到集合P_2,基于此进一步得到集合$P=P_1\cup P_2$。本文把每个节点对划分到同一个社区,把具有共同节点的连边划分到同一个社区,完成对初始社区的优化,从而得到二次社区。

$$P_2=\begin{pmatrix} N_{t1} & N_{t2} & \cdots & N_{tx} \\ N_{t1c} & N_{t2c} & \cdots & N_{txc} \end{pmatrix}$$

其中,集合的第一行N_{tx}表示节点标号,N_{tx}小于或等于网络的节点数量n;集合的第二行N_{txc}表示第一行所示节点对应的特征向量吸引性取最大值时的连边节点标号。

(4)合并小社区。找到节点数量小于阈值的小社区结构,删除小社区内的连边,计算小社区内所有节点的特征向量局部相似性,并找到其中每个节点的最相似点,把新得到的最相似节点对加入已有的P集合,P集合的第一行表示节点标号,第二行表示第一行所示节点对应的连边节点标号,这样会使小社区合并到相应的邻近大社区中,从而得到最终的社区结构。

ELSC算法是一种基于节点重要性的相似性算法。在数据预处理阶段,需要遍历网络中的所有节点对,其计算复杂度为$O(n^2)$,其中n是网络中的节点数。特征向量局部相似性用来获得初始社区,需要遍历并计算网络中每个节点对的相似性值,时间复杂度为$O(n^2\log m)$,其中m是边数,查找最相似性得到的节点对集合的计算复杂度小于$O(n)$。特征向量吸引性用来获得二次社区,需要遍历并计算网络中每个节点对的吸引性值,时间复杂度为$O(n^2\log m)$,找到每个节点的非零吸引性最大值,计算平均值并删掉低于平均值的节点对,其时间复杂度为$O(n^2)$。最后,将数量小于阈值λ的社区删边重连,计算复杂度也远小于$O(n)$。因此,算法的复杂度可以表示为$O(n^2)$。

4 数值仿真和结果分析

本文在不同类型的网络中进行实验来验证所提算法的性能。首先,对一些知名的真实网络进行仿真,包括Zachary空手道俱乐部网络、宽吻海豚网络、悲惨世界网络、美国政治书籍网络、足球网络、爵士乐网络和电子邮件网络等。同时,在7个网络中,将ELSC算法与现有的经典算法(CNM算法、FN算法、GN算法、K-means算法、Infomap算法、Walktrap算法)的Q值以及NMI值进行比较,以验证算法的性能。

4.1 网络数据描述

(1)Zachary空手道俱乐部网络[2]。该网络包含34个节点和78条边,其中一个节点表示一个成员,一条边表示任何两个成员之间是朋友关系。

(2)Dolphins宽吻海豚网络[3]。该网络包括62个节点和159条边,其中一个节点表示一只海豚,一条边表示任何两个海豚之间可以相互联系。

(3)Les Mis悲惨世界网络[5]。该网络由Knuth根据Victor Hugo的小说*Les Miserables*中的主要角色之间的互动来编辑。该网络包含77个节点和508条边,其中一个节点表示一个角色,一条边表示一个或多个场景中相应角色的共现。

(4)Polbooks美国政治书籍网络[8]。该网络包含105个节点和441条边,其中一个节点表示在亚马逊在线书店销售的关于美国政治的书籍,并且任意两个节点之间的一条边表示这两个书籍都是由同一个人购买的[6]。

(5)Football美国大学足球俱乐部网络[11]。该网络是2000年秋季常规赛期间IA大学之间的美式足球比赛网络,包含115个节点和613条边,每个节点代表了参加美国2000年橄榄球赛季的高校代表队,连接两个节点之间的边则表示相应的两支球队之间至少曾有过一场比赛。

(6)Jazz爵士乐网络[26]。爵士乐网络描述舞者跳爵士乐的情况。该网络包含198个节点和2742条边。其中一个节点表示一个舞者,一条边表示舞者之间有过至少一次舞蹈。

(7)Email电子邮件通信网络[27]。该网络包含1133个节点和10903条边,其中一个节点表示一个电子邮件地址,一条边表示地址之间有过至少一次收发电子邮件。

以上7个现实网络的具体信息如表1所列。

表1 7个现实世界网络的信息

Table 1 Information of seven real world networks

No.	Networks	Nodes	Edges	Communities
1	Zachary	34	78	2
2	Dolphins	62	159	4
3	Les Mis	77	508	11
4	Polbooks	105	441	3
5	Football	115	613	12
6	Jazz	198	2742	4
7	Email	1133	10903	10

注:No.表示标号,Networks表示网络,Nodes表示对应网络中节点的数量,Edges表示对应网络中连边的数量,Communities表示对应网络应划分的社区数量

4.2 数值仿真

4.2.1 Zachary空手道俱乐部网络

在空手道俱乐部网络中,首先进行数据预处理,计算网络中每个节点的特征向量中心性值以及节点对邻居集合。其次,利用特征向量局部相似性算法得到空手道网络的部分最相似节点,如表2所列,并得到最相似节点对集合P_1。可以看出,节点1的最相似节点是节点3,节点3的最相似节点是节点4,把每一个最相似节点对划分到同一个社区即节点1和节点3在同一个社区,把具有共同节点的连边划分到同一个社区即节点4和节点1、节点3在同一个社区,这样就形成了网络的初始划分。然后,计算每个节点对的特征向量吸引性,部分最吸引节点如表3所列,找出每行的最大值并求它们的平均值,把低于平均值的节点对删除,将剩下的节点对组成集合P_2,在此基础上得到集合$P=P_1\cup P_2$,从而完成对初始社区的优化。最后,找出节点数量小于或等于阈值$\lambda(\lambda=5)$

的小社区，包括(5，6，7，11，17)，(24，27，28，30)和(25，26)。删除找出的这 3 个小社区内的节点的连边，断开连边后重新计算得到的节点之间的特征向量局部相似性，这些节点对应的特征向量局部相似性最大值的节点对如表 4 所列，将新得到的节点对集合加入 P 集合，这样就将 3 个小社区合并到相应的大社区，对于不同的 λ 值，社区检测情况如表 5 所列，可以看出 $\lambda=5$ 时社团划分结果最好。最终社区检测结果如图 1 所示，网络被分为绿色和黄色两个社区，Q 值为 0.37，仅次于 FN 算法的 0.38，NMI 值为 1，比除 k-means 外的 5 种经典算法都高，由此可见 ELSC 算法在空手道俱乐部网络中的效果较佳。

表 2 Zachary 网络中部分节点的特征向量局部相似性

Table 2 Eigenvector local similarity of part of nodes in Zachary network

	1	2	3	4	5	6	7	8	9	10
1	0	0.899	**0.910**	0.908	0.536	0.540	0.554	0.904	0.688	0
2	0.899	0	0.907	**0.980**	0	0	0	0.976	0	0
3	0.910	0.907	0	**0.924**	0	0	0	0.920	0.707	0.153
4	0.908	0.980	0.924	0	0	0	0	**0.996**	0	0
5	0.536	0	0	0	0	0	**0.967**	0	0	0
6	0.540	0	0	0	0	0	**0.974**	0	0	0
7	0.554	0	0	0	0.969	**0.974**	0	0	0	0
8	0.904	0.976	0.92	**0.996**	0	0	0	0	0	0
9	0.688	0	**0.707**	0	0	0	0	0	0	0
10	0	0	**0.153**	0	0	0	0	0	0	0

表 3 Zachary 网络中部分节点的特征向量吸引性

Table 3 Feature vector attraction of part of nodes in Zachary network

	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	0
2	**1.410**	0	0	0	0	0	0	0	0	0
3	**1.409**	1.128	0	0	0	0	0	0	0	0
4	**1.426**	1.219	1.073	0	0	0	0	0	0	0
5	**1.054**	0	0	0	0	0	0	0	0	0
6	**1.058**	0	0	0	0	0	0	0	0	0
7	**1.072**	0	0	0	0.993	1.007	0	0	0	0
8	**1.197**	0	0.854	0	0	0	0	0	0	0
9	**1.197**	0	0.854	0	0	0	0	0	0	0
10	0	0	**0.303**	0	0	0	0	0	0	0

表 4 Zachary 网络中删边节点的特征向量局部相似性

Table 4 Eigenvector local similarity of edge-deleted-nodes in Zachary network

Node	Node	ELS	Node	Node	ELS	Node	Node	ELS
5	1	0.536	24	26	0.853	25	28	0.273
6	1	0.54	27	34	0.089	26	24	0.835
7	1	0.554	28	25	0.273			
11	1	0.554	30	34	0.062			

表 5 λ 取不同值时 Zachary 网络的社区检测情况

Table 5 Community detection with different λ values in Zachary network

λ	*communities*	Q	NMI
2	4	0.37	0.54
3	4	0.37	0.54
4	3	0.35	0.71
5	2	0.37	1
6	2	0.37	1

图 1 Zachary 网络中的社区

Fig. 1 Communities in Zachary Network

4.2.2 Dolphins 宽吻海豚网络

在宽吻海豚网络中，首先进行数据预处理，计算网络中每个节点的特征向量中心性值以及节点对邻居集合。其次，利用特征向量局部相似性算法得到海豚网络的部分最相似节点如表 6 所列，并得到最相似节点对集合 P_1。可以看出，节点 33 的最相似节点是节点 37，把节点 33 和节点 37 划分到同一个社区，从而形成网络的初始划分。然后，计算每个节点对的特征向量吸引性，部分最相吸点如表 7 所列，找出每行的最大值并求它们的平均值，把低于平均值的节点对删除，将剩下的节点对组成集合 P_2，基于此进一步得到集合 $P=P_1\cup P_2$，完成对初始社区的优化。最后，找出节点数量小于或等于阈值 $\lambda(\lambda=6)$的社区，包括(2，26，27，28)，(4，9，60)，(5，12，22，24，46，52)，(8，20)，(42，55)和(47，50)。删除社区内的连边，重新计算特征向量局部相似性，将最大值的节点对集合加入 P 集合，这样就将 6 个小社区连入相应的大社区，对于不同的 λ 值，社区检测情况如表 8 所列。可以看出，$\lambda=6$ 时社团划分结果最好。最终社区的检测结果如图 2 所示，网络被分成 4 个社区，Q 值为 0.51，NMI 值为 0.83，都比 CNM 算法在内的 6 种经典算法的值高，由此可见 ELSC 算法在宽吻海豚网络中的效果较佳。

表 6 Dolphins 中部分节点的特征向量局部相似性

Table 6 Eigenvector local similarity of part of nodes in Dolphins network

	33	34	35	36	37	38	39	40
33	0	0.428	0	0	**0.628**	0.432	0	0.374
34	**0.429**	0	0	0	0.397	0	0	0
35	0	0	0	0	0	0	0	0
36	0	0	0	0	0.22	0	0.144	**0.288**
37	**0.628**	0.397	0	0.22	0	0	0	0.421
38	**0.432**	0	0	0	0	0	0	0
39	0	0	0	0.144	0	0	0	0
40	0.374	0	0	0.288	**0.421**	0	0	0

表 7 Dolphins 网络中部分节点的特征向量吸引性

Table 7 Feature vector attraction of part of nodes in Dolphins network

	33	34	35	36	37	38	39	40
33	0	0	0	0	**0.730**	0	0	0
34	**0.544**	0	0	0	0.514	0	0	0
35	0	0	0	0	0	0	0	0
36	0	0	0	0	0.319	0	0	**0.380**
37	0	0	0	0	0	0	0	0
38	**0.534**	0	0	0	0	0	0	0
39	0	0	0	**0.257**	0	0	0	0
40	0.468	0	0	0	**0.521**	0	0	0

表 8 λ 取不同值时 Dolphins 网络的社区检测情况

Table 8 Community detection with different λ values in Dolphins network

λ	communities	Q	NMI
2	7	0.45	0.62
3	6	0.44	0.68
4	5	0.46	0.75
5	5	0.46	0.75
6	4	0.51	0.83
7	4	0.51	0.83

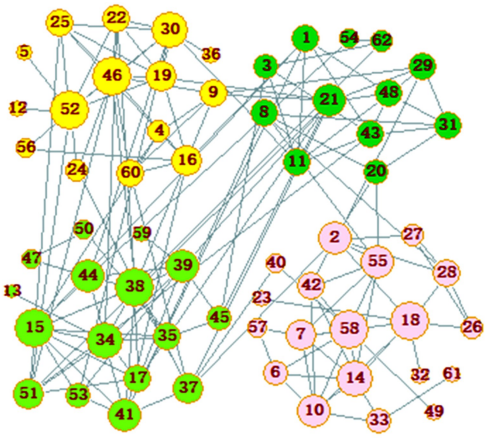

图 2 Dolphins 网络中的社区

Fig. 2 Communities in Dolphins Network

4.2.3 Les Mis 悲惨世界网络

在悲惨世界网络中，首先预处理数据，计算网络中每个节点的特征向量中心性值以及节点对邻居集合。其次，利用特征向量局部相似性算法得到悲惨世界网络的部分最相似节点，由最相似节点对集合 P_1 得到网络初始划分。然后，计算每个节点对的特征向量吸引性，找出每行的最大值并求它们的平均值，把低于平均值的节点对删除，将剩下的最相吸节点对组成集合 P_2，基于此进一步得到集合 $P=P_1 \cup P_2$，完成对初始社区的优化。最后，找出节点数量小于或等于阈值 $\lambda(\lambda=5)$ 的社区，删除社区内的连边，重新计算最相似性得到节点对集合并加入 P 集合，将小社区连入相应的大社区。最终的社区检测结果如图 3 所示，网络被分成 7 个社区，Q 值为 0.32，略低于其他算法，NMI 值为 0.71，略低于 Walktrap 算法的 0.80，由此可见 ELSC 算法在悲惨世界网络中的效果良好。

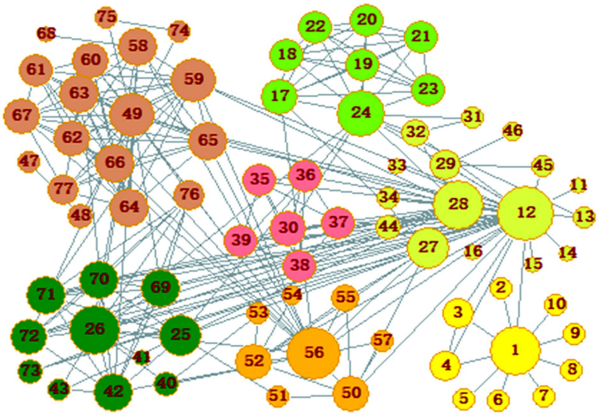

图 3 Les Mis 网络中的社区

Fig. 3 Communities in Les Mis Network

4.2.4 Football 美国大学足球俱乐部网络

在足球俱乐部网络中，首先进行数据预处理，计算网络中每个节点的特征向量中心性值以及节点对邻居集合。其次，利用特征向量局部相似性算法得到足球网络的部分最相似节点，由最相似节点对集合 P_1 得到网络初始划分。然后，计算每个节点对的特征向量吸引性，找出每行的最大值并求它们的平均值，把低于平均值的节点对删除，将剩下的最相吸节点对组成集合 P_2，基于此进一步得到集合 $P=P_1 \cup P_2$，完成对初始社区的优化。最后，找出节点数量小于或等于阈值 $\lambda(\lambda=4)$ 的社区，删除社区内的连边，重新计算最相似性得到节点对集合并加入 P 集合，将小社区连入相应的大社区。最终，社区检测结果如图 4 所示，网络被分成 11 个社区，Q 值为 0.60，NMI 值为 0.91，都比 CNM 算法在内的 6 种经典算法的值高，由此可见 ELSC 算法在足球俱乐部网络中的效果较佳。

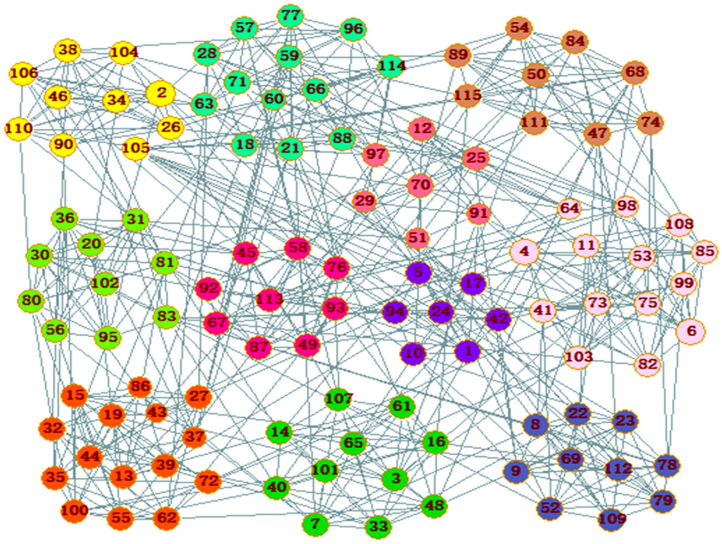

图 4 Football 网络中的社区

Fig. 4 Communities in Football Network

4.3 不同算法在7个真实网络上的性能比较

在现实世界的网络中，模块性 Q 值和标准化互信息 NMI 值分别由 ELSC 算法、CNM 算法、FN 算法、GN 算法、K-means 算法、Infomap 算法、Walktrap 算法基于上述 7 个真实世界网络计算得到，这 7 种算法在不同真实世界网络中的对比分析如下。

模块化 Q 度量用于验证复杂网络中社区检测的准确性。普遍认为 Q 值越高，社区检测的结果越好。从表 9 和图 5 可以看到，ELSC 算法在 Dolphins，Polbooks，Football 和 Email 这 4 个网络中的 Q 值最高；FN 算法在 Zachary，Les Mis，Polbooks，Jazz 和 Email 这 5 个网络中的 Q 值最高，ELSC 算法在这 5 个网络中比 FN 算法的 Q 值稍小，在 Zachary 和 Jazz 网络中 Q 值仅差 0.01，在 Les Mis 网络中 Q 值相差 0.02，大致相当，这表明 ELSC 算法和 FN 算法在这 7 个真实世界网络上的划分结果接近，性能相当，都表现良好。

经过多次实验，ELSC 算法在表 9 中的 7 个数据集中选用的最佳阈值 λ 分别为 5，6，5，6，4，3，5。对比的经典算法中，CNM 算法、FN 算法、GN 算法、Infomap 算法和 Walktrap 算法通过反复迭代得到最终社区划分，不需要设置参数，而 K-means 算法所需参数 k 在本文中经过反复调参，选取最优值作为实验结果，然后将这些结果与本文算法的结果做对比分析。

表 9 各算法模块度值的比较

Table 9 Comparison of module degree

Networks	Nodes	Edges	ELSC	CNM	FN	GN	K-means	Infomap	Walktrap
Zachary	34	78	0.37	0.24	**0.38**	0.36	0.23	0.24	0.22
Dolphins	62	159	**0.51**	0.40	0.49	0.45	0.01	0.42	0.40
Les Mis	77	254	0.32	0.45	**0.50**	0.45	0.1	0.40	0.41
Polbooks	105	441	**0.48**	0.09	**0.48**	0.44	0.08	0.45	0.45
Football	115	613	**0.60**	0.53	0.55	0.59	0.41	0.48	0.27
Jazz	198	2,742	0.40	0.28	**0.41**	0.38	−0.04	0.19	0.27
Email	1133	10,903	**0.48**	0.49	**0.48**	0.44	0.21	0.47	0.42

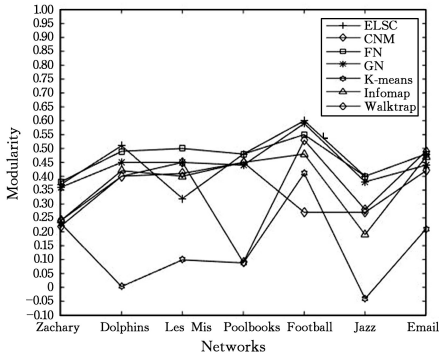

图 5 各算法的模块度值的比较

Fig. 5 Comparison of modular value of algorithms

NMI 用于验证复杂网络中社区检测的有效性。人们普遍认为 NMI 值越高，社区检测结果就越好。从表 10 和图 6 可以看出，ELSC 算法在 Zachary，Dolphins 和 Football 这 3 个网络中的 NMI 值最高，在 Les Mis 网络中的 NMI 值稍小于 Walktrap 算法，这些算法的参数设置同计算模块度的参数。可见，ELSC 算法在这 4 个真实世界网络上的表现良好。

通过 ELSC 算法与 CNM 算法、FN 算法、GN 算法、K-means 算法、Infomap 算法、Walktrap 算法在空手道俱乐部网络、宽吻海豚网络、悲惨世界网络、爵士乐队网络、美国大学橄榄球网络、美国政治书籍网络和电子邮件网络这 7 个真实网络数据集上的比较，根据表 9 和图 5 的模块度值以及对比表 10 和图 6 的标准化互信息值后，综合考虑模块性和标准化互信息两个指标可以看出，ELSC 算法在 7 个真实世界网络中的表现良好，大部分指标值超过其他 6 种经典算法，少数指标值与表现最好的经典算法的值相当，因此 ELSC 算法的综合性能最佳。

表 10 各算法的 NMI 值比较

Table 10 Comparison of NMI values of algorithms

Networks	Nodes	Edges	ELSC	CNM	FN	GN	K-means	Infomap	Walktrap
Zachary	34	78	**1.00**	0.69	0.57	0.83	**1.00**	0.69	0.56
Dolphins	62	159	**0.83**	0.57	0.67	0.67	0.40	0.59	0.55
Les Mis	77	254	0.71	0.66	0.62	0.67	0.60	0.79	**0.80**
Football	115	613	**0.91**	0.71	0.61	0.90	0.87	0.24	0.24

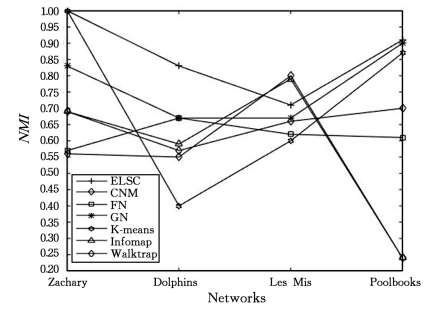

图 6 各算法的标准化互信息值的比较

Fig. 6 Comparison of NMI values of algorithms

结束语 本文提出了一种基于特征向量局部相似性和特征向量吸引性的社区检测算法。首先计算网络中每个节点的特征向量中心性，在此基础上提出特征向量局部相似性指标(ELS)和特征向量吸引性指标(EA)，基于 ELS 指标划分网络的初始社区，基于 EA 指标优化社区结构。在 7 个真实网络上将所提算法与 6 种知名算法进行数值仿真比较，结果表明该算法具有良好的检测精度和较低的计算复杂度。

参 考 文 献

[1] NEWMAN M E J，CLAUSET A. Structure and inference in annotated networks [J]. Nature Communications，2016，7：11863.

[2] HU F, WANG M Z, WANG Y R, et al. An algorithm J-SC of detecting communities in complex networks [J]. Physics Letters A, 2017, 381(42): 3604-3612.
[3] HU F, LIU Y H. A new algorithm CNM-Centrality of detecting communities based on node centrality [J]. Physics A, 2016, 446: 138-151.
[4] WANG T, YIUN L Y, WANG X X. A community detection method based on local similarity and degree clustering information [J]. Physics A, 2018, 490: 1344-1354.
[5] CHEN L G, WANG Y R, HUANG X M, et al. SA-SOM algorithm for detecting communities in complex networks [J]. Physics Letters B, 2017, 31(29): 1750262.
[6] CLAUSET A, NEWMAN M E J, MOORE C. Finding community structure in very large networks [J]. Physical Review E, 2004, 70: 066111.
[7] ZHANG X K, REN J, SONG C, et al. Label propagation algorithm for community detection based on node importance and label influence [J]. Physics Letters A, 2017, 381: 2691-2698.
[8] LI Y F, JIA C Y, YU J. A parameter-free community detection method based on centrality and dispersion of nodes in complex networks [J]. Physics A, 2015, 438: 321-334.
[9] YUTAKA I. SUEMATSU L, YUTA K. A Framework for Fast Community Extraction of Large-Scale Networks [J]. ACM, 2008, 978: 1215-1216.
[10] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics, 2008, 2008: p10008.
[11] BARBE R M J, CLARK J W. Detecting network communities by propagating labels under constraints [J]. Physical Review E, 2009, 80: 026129.
[12] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks [J]. Physical Review E, 2007, 76: 036106.
[13] ŠUBELJ L, BAJEC M. Ubiquitousness of link-density and link-pattern communities in real-world networks [J]. The European Physical Journal B, 2012, 85: 1-11.
[14] JIN H, WANG S, LI C. Community detection in complex networks by density-based clustering [J]. Physics A, 2013, 392(19): 4606-4618.
[15] GONG M, LIU J. Novel heuristic density-based method for community detection in networks [J]. Physics A, 2014, 403: 71-84.
[16] ZHOU H. Distance, dissimilarity index, and network community structure [J]. Physical Review E, 2003, 67(6): 061901.
[17] ROSVALL M, BERGSTROM C T. Maps of random walks on complex networks reveal community structure [J]. PNAS, 2008, 105(4): 1118-1123.
[18] MACQUEEN J B. Some methods for classification and analysis of multivariate observations in: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability [J]. American Mathematical Society, 1967(66): 281-297.
[19] JIANG Y W, JIA C V, YU J. An efficient community detection method based on rank centrality [J]. Physics A, 2013, 392: 2182-2194.
[20] XU H T, WU H, FANG X J, et al. Finding Key Stations of Hangzhou Public Bicycle System by an Improved K-Means Algorithm [J]. Journal of Mechanics of Materials and Structures, 2012, 209: 925-929.
[21] FENDER A, EMAD N, PETITION S. Parallel Jaccard and Related Graph Clustering Techniques [J]. ACM ISBN, 2017, 978: 4503-5125.
[22] EUSTACE J. Community detection using local neighborhood in complex networks [J]. Physics A, 2015, 436: 665-677.
[23] BAGROW J P, BOLLT E M. Local method for detecting communities [J]. Physical Review E, 2005, 72 (4): 046108.
[24] CLAUSET A. Finding local community structure in networks [J]. Physical Review E, 2005, 72(2): 026132.
[25] HUANG J, SUN H, LIU Y, et al. Towards online multiresolution community detection in large-scale networks [J]. PLOS ONE, 2011, 6(8): e23829.
[26] GLEISER P M, DANON L. Community structure in Jazz[J]. Advances in Complex Systems, 2003, 6: 565-573.
[27] LANCICHINETTI A, FORTUNATO S, RADIHHI F. Benchmark graphs for testing community detection algorithms [J]. Physical Review E, 2008, 78: 046110.
[28] HU Q C, ZHANG Y, XING C X. Research on Maximization Method of Social Network Influence Based on Overlapping Community Division [J]. Computer Science, 2018, 45(6): 32-35.
[29] REN X L, LV L Y. Review of ranking nodes in complex networks [J]. Chinese Science Bulletin, 2014, 9(13): 1175-1197.
[30] ZHANG D, LI X H, LIU H, et al. Service-Oriented Network Node Importance Ranking Method in SDN Network[J]. Chinese Journal of Computers, 2018, 41(11): 2624-2636.
[31] QIAO S J, HAN N, ZHANG K F, et al. Overlapping community detection algorithm in complex network big data[J]. Journal of Software, 2017, 28(3): 631-647.
[32] DAI D B, TANG C L, XIONG W. Sequence clustering algorithm based on global and local similarity [J]. Journal of Software, 2010, 21(4): 702-717.

YANG Xu-hua, born in 1971, Ph.D, professor, Ph.D supervisor, is member of China Computer Federation. His main research interests include machine learning, complex networks and intelligent transportation systems.