第46卷 第7期 计 算 机 工 程 2020年7月
Vol.46 No.7 Computer Engineering July 2020

·人工智能与模式识别· 文章编号: 1000-3428(2020)07-0084-07 文献标志码: A 中图分类号: TP393

基于节点多属性相似性聚类的社团划分算法

邱少明[1],於 涛[1],杜秀丽[1],陈 波[2]

(1.大连大学 通信与网络重点实验室,辽宁 大连 116622; 2.岭南师范学院 信息工程学院,广东 湛江 524048)

摘 要: 针对当前社团划分算法存在划分方式单一和划分结果准确度低等问题,提出一种基于节点多属性相似性聚类的社团划分算法SM-CD。根据社会网络特性定义网络节点的结构属性与自身属性,通过调整两类属性在网络中所占的权重计算网络节点之间的相似度矩阵,并将网络节点按照相似度和模块度指标划分为不同的社团。在Zachary和Football真实网络数据集上的实验结果表明,SM-CD算法相比Newman、GN等算法具有更高的社团划分准确率。

关键词: 复杂网络; 社团划分; 节点属性; 相似度矩阵; 聚类

开放科学(资源服务)标志码(OSID):

中文引用格式: 邱少明,於涛,杜秀丽,等. 基于节点多属性相似性聚类的社团划分算法[J]. 计算机工程,2020,46(7):84-90,97.

英文引用格式: QIU Shaoming,YU Tao,DU Xiuli,et al. Community division algorithm based on similarity clustering of node multiple attribute[J]. Computer Engineering,2020,46(7):84-90,97.

Community Division Algorithm Based on Similarity Clustering of Node Multiple Attribute

QIU Shaoming[1],YU Tao[1],DU Xiuli[1],CHEN Bo[2]

(1.Key Laboratory of Communication and Network,Dalian University,Dalian,Liaoning 116622,China;
2.School of Information Engineering,Lingnan Normal University,Zhanjiang,Guangdong 524048,China)

【Abstract】 Existing community division algorithms lack diversity in the division method,and division results are not accurate. To address the problem,this paper proposes a community division algorithm,SM-CD,on the basis of similarity clustering of multiple attributes of nodes. The algorithm uses social network features to define the structure attributes of nodes and the attributes of oneself. By adjusting the weight of two kinds of attributes in network,the similarity matrix of network nodes is calculated. Then the nodes are divided into different communities according to similarity and modularity. Experimental results on the real network data from Zachary and Football show that SM-CD has a higher accuracy rate in community division than Newman,GN and other algorithms.

【Key words】 complex network; community division; node attribute; similarity matrix; clustering

DOI: 10.19678/j.issn.1000-3428.0055070

0 概述

网络是由节点和连线构成,表示不同对象之间的相互联系。在现实生活中,存在蛋白质网络[1]、神经网络[2]、社会网络[3]等各种类型的网络,不同的网络具有不同的结构和属性,而不同的属性也会表现出不一样的特征。随着对网络研究的进一步深入发现网络中存在一种特殊的拓扑结构,其主要特点是在该网络结构内的节点联系非常紧密,在网络结构之间的节点联系相对稀疏,研究人员将该结构特征初步定义为网络社团现象[4]。社会网络的社团宏观结构体现了网络成员之间的联系,反映了成员之间的关系属性,可以被定义为社会成员之间紧密相连的集合。

目前,社团划分方法主要包括基于模块度的社团划分方法、基于图划分的社团划分方法和基于层次聚类的社团划分方法。基于模块度的社团划分方法由NEWMAN于2004年提出[5],其核心是一种分

基金项目:装备发展部预研基金(6140002010101,6140001030111)。
作者简介:邱少明(1977—),男,副教授,主研方向为智能故障诊断、复杂网络;於 涛,硕士研究生;杜秀丽、陈 波,教授、博士。
收稿日期:2019-05-30 修回日期:2019-07-23 E-mail:17896015@qq.com

层贪婪算法,如 Fast-Unfolding 算法[6],可以快速实现对社团的划分,但容易陷入局部最优,不适用于大规模的网络社团划分。基于图划分的社团划分方法[7]如 Kernighan-Lin 算法[8]、谱平分算法等,均是经典的图谱理论划分算法,需要预知社团数目才可实现社团划分。基于层次聚类的社团划分方法[9]尽管具有较好的社团划分效果,但是聚类标准的选取对划分质量影响较大。

在此基础上,文献[10]提出局部社团检测算法,通过不同的隶属函数进行动态检测,但其是针对局部社团进行划分,忽略了网络全局特性,对网络整体结构的判断不够准确,容易陷入局部最优。通过进一步研究发现,社团划分与聚类问题非常相似,因此文献[11]提出基于密度的聚类方法,相对于已有方法聚类效果有所改进,但其聚类的初始条件选取存在随机性,会导致划分的社团不稳定。文献[12]提出相似性概念,但其在相似性网络定义上很少从自身属性进行考虑,并且社团划分方法仅采用单一指标,如聚集系数[13],不具说服力。文献[14]基于 RA 指标提出顶点相似性指标,并将其运用于推荐系统中,具有较好的推荐效果,但其仅针对小型网络,如果将其运用于具有较多属性的大型网络中,则聚类效果会迅速下降。文献[15]提出一种基于节点相异度的社团层次划分算法,结合核度和接近度评估得出节点相异度评价结果,从而计算节点间的相似度。虽然该算法能够针对性地选取初始节点,但相异度评估方式没有考虑到网络节点间的多属性关系,仅通过节点之间的最短距离衡量相异程度,缺乏一定的合理性。相似性指标主要分为局部相似性指标与全局相似性两类,常见的局部相似性指标主要有 Jaccard 指标[16]等,全局相似性指标主要有 SimRank 指标[17]等。文献[18]提出一种综合考虑用户行为与相似度的社区发现算法,将网络中用户间的多维关系抽象成相似度,并将相关因子作为模块度的目标函数进行社团划分,能够基本达到划分的目的,但是精确度较低。由于上述研究方法多数仅考虑了网络局部特性,未考虑网络全局特性,存在划分方法单一、划分精度不高等问题,因此本文提出基于节点多属性相似性聚类的社团划分算法 SM-CD。

1　节点多属性相似性度量

1.1　网络抽象

所有网络都可以被抽象成节点及其连边构成的结构,本文提出一种考虑节点多属性的度量方法,将网络中某节点与其邻居节点直接相连的网络节点之间的属性称为节点自身属性,将非直接相连的网络节点之间的属性称为结构属性,从节点自身属性和结构属性角度出发,根据相似性进行社团划分操作。

将具有多属性的复杂网络抽象成无向图 $G=<V,E,\boldsymbol{H}>$,其中,$V=\{v_1,v_2,\cdots,v_n\}$ 表示网络中所有节点的集合,$E=\{e_1,e_2,\cdots,e_m\}$ 表示连接 G 中边的集合,$n=|V|$ 表示节点个数,$m=|E|$ 表示边数,E 中的每条边都有 V 中两个不同节点形成的边与之对应,$\boldsymbol{H}^{\mathrm{T}}=\{h_i^1,h_i^2,\cdots,h_i^m\}$ 表示与 V 对应的描述顶点属性相关联的 m 个属性集,其中 h_i^m 表示节点 v_i 在第 m 个属性上的值。

1.2　节点相似度计算

1.2.1　节点属性选择

由于网络中的节点不是完全孤立存在的,而是与其他节点存在联系,在社会关系网络中其集聚性表现尤为明显,因此集聚性是重要的节点自身属性指标。

属性 1(聚集系数 C)　节点 i 的聚集系数计算如式(1)所示:

$$C(i)=\begin{cases}\dfrac{2|E(\varphi_i)|}{|\varphi_i|\times(|\varphi_i|-1)}, & |\varphi_i|>1\\ 0, & |\varphi_i|\leqslant 1\end{cases} \tag{1}$$

其中,φ_i 表示节点 i 的邻居节点集合,$|E(\varphi_i)|$ 表示节点 i 的邻居节点之间的实际连边数目。

在社会网络中,每个成员地位不同,有的成员与其他成员联系少,但是他可能是两个社团的中间联络人,如果忽视其作用,则可能导致两个社团中的联系出现中断或者难度加大。如果网络中的两个个体没有直接相连,且两者之间不存在间接相连,则两者之间存在阻碍,节点效率是用来描述该节点对网络中其他相关节点的影响程度,是重要的结构属性指标。

属性 2(节点效率 E_{ef})　设节点数量为 n,则节点 i 的效率计算如式(2)所示:

$$E_{\mathrm{ef}}(i)=\frac{\sum_j\left(1-\sum_l Y_{il}Y_{jl}\right)}{n} \tag{2}$$

其中,j 表示节点 i 的邻接节点,l 表示节点 i 和节点 j 的共同邻接节点,Y_{il} 和 Y_{jl} 分别表示节点 l 在节点 i 和节点 j 的邻居节点中所占的比例。

属性 3(核度)　节点的核度定义为节点在核中的深度,核度的最大值对应网络结构中最中心的位置。使用核度可以描述度分布所不能描述的网络特征,揭示源于系统特殊结构的层次性。核值的大小可以将网络中的节点按照影响力进行分类,在实际网络中,核度越大的节点,影响力也相对较大。

属性 4(共有节点引力) 在进行网络节点划分时,节点归属的强弱是节点划分的依据,如果两个节点之间的共有节点集合越多,说明两个节点之间的交互信息量越大,在考虑相似度时会在结构上产生更强的吸引力。因此,在定义共有节点引力时,共有节点占据相应节点的比例可以直观反映两节点在结构上的相似性。节点 i 和节点 j 的共有引力计算如式(3)所示:

$$\Gamma(i,j)=\frac{|\varphi_i,\varphi_j|}{|\varphi_i||\varphi_j|} \tag{3}$$

其中,$|\varphi_i,\varphi_j|$ 表示节点 i 和节点 j 的共同邻居数量,$|\varphi_i|$ 和 $|\varphi_j|$ 分别表示节点 i 和节点 j 的邻居节点数量,两者的比值越大,说明两类节点在结构上具有更强的关联性。

1.2.2 节点属性相似度矩阵求解

本文利用节点的多属性信息结合局部与全局信息对网络中的社团进行划分,具体的相似性指标定义为:

定义 1 邻接矩阵 $\boldsymbol{A}\in\mathbb{R}^{n\times n}$,其表示网络中节点的连接关系,$n$ 为节点总数,即如果节点 i 和 j 之间有连接关系,则 $a_{ij}=1$,如果两者之间没有连接关系,则 $a_{ij}=0$。

定义 2 中间矩阵 $\boldsymbol{M}^{m\times n}$,假设网络中有 n 个节点,衡量节点的属性有 m 个,属性向量表示为 $\boldsymbol{H}_i^{\mathrm{T}}=\{h_i^1,h_i^2,\cdots,h_i^m\}$,中间矩阵 $\boldsymbol{M}$ 表示如式(4)所示,每一列代表一个节点的 m 个属性,并同时包含结构属性与自身属性两类。

$$\boldsymbol{M}=[\boldsymbol{H}_1\ \boldsymbol{H}_2\cdots\boldsymbol{H}_n]=\begin{bmatrix}h_1^1 & h_2^1 & \cdots & h_n^1\\ h_1^2 & h_2^2 & \cdots & h_n^2\\ \vdots & \vdots & & \vdots\\ h_1^m & h_2^m & \cdots & h_n^m\end{bmatrix} \tag{4}$$

定义 3 相似度矩阵 $\boldsymbol{S}_{\mathrm{Sim}}\in\mathbb{R}^{n\times n}$,该相似度矩阵中的每一个值均刻画了该网络中的某个节点和其他节点的属性相似程度。

在获取中间矩阵后,需要进一步获取不同节点之间的相似度矩阵,相似度矩阵描述了网络中的各节点在各属性协同下,依据不同的权重值计算出的相似情况,从而求解相似度矩阵 $\boldsymbol{S}_{\mathrm{Sim}}$ 的每个值。例如,给定两个节点 v_i 和 v_j,其对应的属性向量为 $\boldsymbol{h}_i$ 和 $\boldsymbol{h}_j$,则两个矢量节点之间的相似度 S_{ij} 计算如式(5)所示:

$$S_{ij}=\begin{cases}\dfrac{\boldsymbol{g}_c(L^i)^{\mathrm{T}}\cdot\boldsymbol{g}_c(L^j)}{\sqrt{\boldsymbol{g}_c(L^i)^{\mathrm{T}}\cdot\boldsymbol{g}_c(L^i)}\cdot\sqrt{\boldsymbol{g}_c(L^j)^{\mathrm{T}}\cdot\boldsymbol{g}_c(L^j)}} & i,j\text{ 直接相连}\\ 0 & i,j\text{ 无法相连}\end{cases} \tag{5}$$

当 i、j 直接相连时,满足式(6):

$$\begin{cases}L^i=\alpha c_i^{\mathrm{Str}}\beta c_i^{\mathrm{Node}}\\ L^j=\alpha c_j^{\mathrm{Str}}\beta c_j^{\mathrm{Node}}\end{cases} \tag{6}$$

当 i、j 不直接相连时,满足式(7):

$$\begin{cases}L^i=\beta c_i^{\mathrm{Node}}\\ L^j=\beta c_j^{\mathrm{Node}}\end{cases} \tag{7}$$

其中,$\boldsymbol{g}_c(L^i)$ 表示一个 m 维的特征矢量,用来描述节点的多维属性,α 和 β 分别表示节点属性与结构属性的权重,c_i^{Node} 表示节点 i 的节点属性,c_i^{Str} 表示节点 i 的结构属性。在计算节点之间的相似度时,最终可得到 $n\times n$ 的对称相似度矩阵 $\boldsymbol{S}_{\mathrm{Sim}}$ 如式(8)所示:

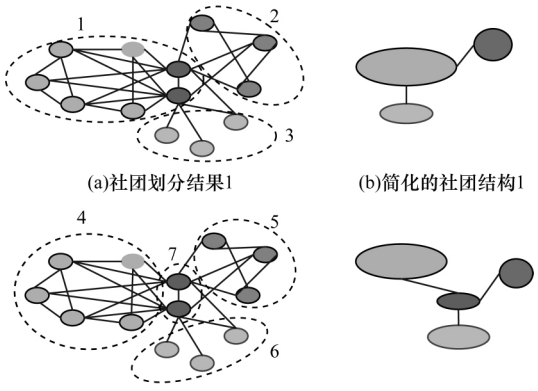
$$\boldsymbol{S}_{\mathrm{Sim}}=\begin{bmatrix}s_{11} & s_{12} & \cdots & s_{1n}\\ s_{21} & s_{22} & \cdots & s_{2n}\\ \vdots & \vdots & & \vdots\\ s_{n1} & s_{n2} & \cdots & s_{nn}\end{bmatrix} \tag{8}$$

2 基于节点多属性相似性聚类的社团划分

2.1 节点多属性相似性聚类

基于节点多属性的社团算法从不同的角度入手描述网络中的节点结构属性信息和节点自身属性信息,并采用分簇形式对社团进行高效划分。假设将一个网络拓扑图划分为 k 个不相交的图 $G_i(V_i,E_i,H)$,$V=U_{i=1}^k$ 和 $V_i\cap V_j=\varnothing$,对于任意 $i\neq j$ 均满足该情况,且划分后的社团满足以下要求:1)在同一个集群内的顶点在结构方面互相靠近,簇顶点之间相互分离;2)当被划分在同一个集群中的节点顶点有相似的属性值时,簇顶点可能会有不同的属性值。

图 1 描述了考虑节点不同属性的社团划分结果。在图 1(a)和图 1(c)中的簇节点划分情况在不断变化,因此在简化社团结构图 1(b)和图 1(d)中反映的结果也存在差异。在图 1(a)中,当不考虑簇 7 节点的结构属性(如共有节点引力及效率)而仅考虑节点的自身属性时,整个网络被划分成 3 个社团如图 1(b)中简化图所示,当考虑网络自身属性时,网络会分成 4 个社团如图 1(d)中简化图所示。



(a)社团划分结果1 (b)简化的社团结构1

(c)社团划分结果2 (d)简化的社团结构2

图 1 考虑节点不同属性的社团划分结果

Fig. 1 Community division results considering different attributes of nodes

根据节点多属性定义可以利用本文聚类算法将不同节点划分到不同社团结构中，因此，在进行社团聚类算法实现时，最重要的是如何使用聚类思想合理融合节点属性信息并对社团进行划分。在进行社团划分时，本文考虑到节点的不同属性，并在此基础上，结合节点相似度矩阵 S_{Sim}，提出一种基于节点多属性相似性聚类的社团划分算法。

传统聚类算法是对于给定的基本样本集，按照样本之间的距离，将样本集划分为 K 个簇，使簇内的点尽量紧密连接在一起，而簇间距离尽可能大。在本文算法中，将各节点之间的相似度作为样本集，即相似度矩阵中的各 s_{ij} 值，根据影响度函数选取初始质心节点，并采用聚类思想进行划分操作。

1) 社团 K 值的确定。依据当前研究发现，对于社团结构不是很明显的网络，无法在初期直接确定网络中的社团数目。所以，本文在理论范围内采用遍历方式寻找最佳 K 值，选出最优划分结构对应的 K 值即为最终输出的社团数目。

2) 聚类质心的选取。一个好的初始质点是均值聚类的必要条件，在选取聚类的初始质心时，本文从影响度的角度出发，而不是随机选取初始点。由分析可知，如果某个顶点 v_i 的周围节点连接比较稠密，则表明该节点在网络中具有较强的吸附能力，即更容易形成社团，则以这类节点作为质心节点相比于在网络中随机选取初始节点更具代表性。本文影响度函数的定义如式(9)所示:

$$f(v_i)=\sum_{v_j\in V}\left(1-\mathrm{e}^{-\frac{\sqrt{d(v_i,v_j)}}{2\partial^2}}\right),\ \partial=2^{-d(v_i,v_j)} \tag{9}$$

其中，影响度函数 $f(v_i)\in(0,1)$，用于衡量一个节点相对于其他节点的影响程度，$d(v_i,v_j)$ 表示节点 v_i 和 v_j 之间的欧几里得距离，考虑到网络节点之间的距离对影响度函数具有较大影响，所以本文引入 ∂ 调节因子，通过 ∂ 来调整不同距离之间的节点影响程度。在本文实验中发现，如果不引入调节因子会导致计算出的指标值相近，区分效果不明显。因此，本文依据影响度函数的大小，选取排名前 η 个的节点作为选取合理 K 值的依据。

2.2　算法描述

本文 SM-CD 算法的主要思想是同时考虑网络节点的自身属性与结构属性信息得到中间矩阵，并利用节点相似度计算方法，将计算得到的相似度值作为聚类样本集得到相似度矩阵，再结合影响度函数选取排名靠前的节点集合 η 作为遍历上限值，通过不断改变 α 与 β，选取合理的 K 值，直至模块度值达到最大，并将对应的 K 值作为最终社团划分数量，算法具体流程如图 2 所示。

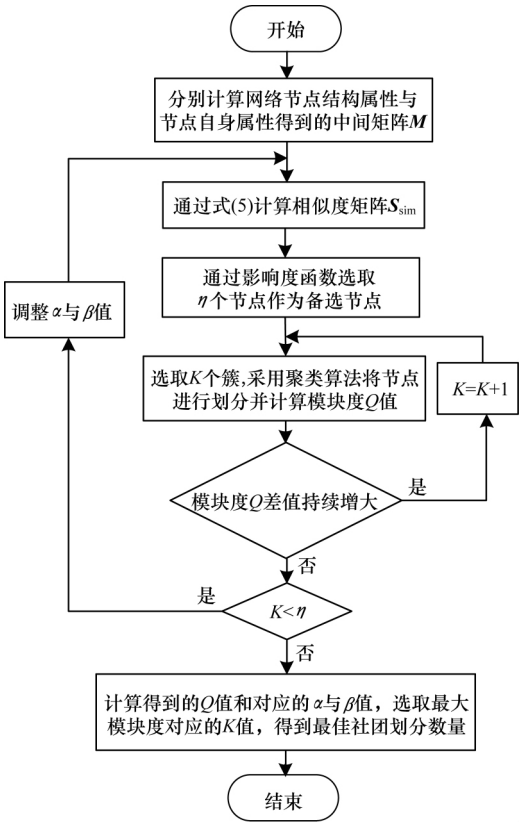


图 2　SM-CD 算法流程

Fig. 2　Procedure of SM-CD algorithm

SM-CD 算法实现步骤具体如下:

输入　简化了属性值的无向图 $G=(V,E)$

输出　通过算法得到的社团划分数目 K 与社团结构

步骤 1　输入网络的邻接矩阵 $\boldsymbol{A}$。

步骤 2　计算网络节点结构属性与自身属性得到中间矩阵 $\boldsymbol{M}$，其中属性 4 是网络中节点 i 与其他 $n-1$ 个节点之间的共有引力，作为结构属性，得到网络的相似度矩阵 $\boldsymbol{S}_{\mathrm{Sim}}$。

步骤 3　通过影响度函数选取前 η 个备选节点作为社团个数的有效集合，计算模块度 Q 值。

步骤 4　判断每次计算的模块度是否增加，如果当前计算得到的模块度差值持续增加，则调整 K 值继续计算当前 α 与 β 取值条件下模块度的变化情况，如果模块度值没有持续增加，则调整 α 与 β 的取值，跳转至步骤 2 重新计算相似度矩阵，从而得到不同权重下的模块度值。

步骤 5　统计在不同权重下的模块度值，选取最大模块度对应的 K 值，即社团划分的最佳数量。

3　实验设置

3.1　评价指标

3.1.1　模块度

由于在真实网络中无法预先明确网络中的社团结构，因此在衡量社团划分准确性时，本文采用模块

度 Q 作为评价指标。在每次进行节点聚类时，都会计算此时的模块度大小并观察其值的变化，从而决定是否将该节点划分到相应社团中。

模块度 Q 是常用的比较社团划分质量的评价指标，其中 $Q \in [0,1]$，社团划分的质量越高，其模块度的值越大，社团内的节点相似性越强，社团划分效果越好，Q 的具体定义如式(10)所示：

$$Q=\frac{1}{2m}\sum_{i,j}\left[A_{ij}-\frac{k_i k_j}{2m}\right]\delta(c_i,c_j) \tag{10}$$

其中：A_{ij} 表示节点 i 和 j 之间的连接关系，如果两者之间有连接，则 $A_{ij}=1$，否则为 0；δ 为隶属函数，如果节点 i 和 j 属于同一个社团，则只有当 $c_i=c_j$ 时，$\delta(c_i,c_j)=1$，否则为 0；$m=\frac{1}{2}\sum A_{ij}$ 为网络中边的数目。

3.1.2 划分准确率

模块度评价指标是聚类的内部评价指标，是一种无监督度量指标，虽然对于单个聚类有较好的评价效果，但是对于聚类最终结果与实际结果之间存在一定的误差。因此，需要引入一种有监督度量指标划分准确率 S，作为外部评价指标，其值一般为准确划分后的社团个数占全部节点的比例，如式(11)所示：

$$S=N_R/N \tag{11}$$

其中，N_R 表示准确划分的社团数目，N 表示总节点数。

3.2 真实网络实验

为验证本文算法的准确性，选取 2 个真实网络，分别为 Zachary 网络和 Football 网络(http://www-personal.umich.edu/~mejn/netdata/)，如表 1 所示。对这 2 个真实网络进行实验研究并将本文算法与其他算法进行性能比较，以验证本文算法的有效性。

表 1 真实网络拓扑结构

Table 1 Real network topology

网络名称	节点个数	边数	平均核度	平均聚集系数
Zachary	34	78	2.206	0.246
Football	115	613	5.330	0.202

3.2.1 Zachary 网络实验

Zachary 网络是美国某大学空手道俱乐部的关系网络，该网络包含 34 个节点及 78 条边，其中节点表示俱乐部成员，边表示成员之间存在的关系。节点 1 代表教练，节点 34 代表校长。由于教练和校长对于俱乐部收费问题存在分歧，因此导致该俱乐部分成以校长和教练为核心的小社团。本文算法对 Zachary 空手道俱乐部网络进行划分，在进行多次实验后发现，当 $K=2$ 时，社团模块度 $Q=0.427$，此时 $\beta=0.5$，调节因子 $\partial=0.4$，划分效果最优，社团划分结果如图 3 所示。

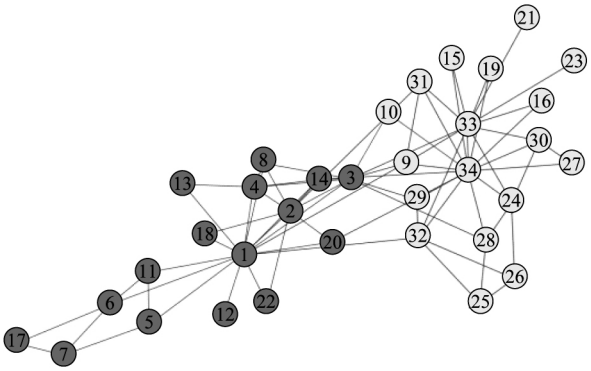


图 3 本文算法对 Zachary 网络的社团划分结果

Fig. 3 Community division results of the proposed algorithm for Zachary network

通过本文算法对 Zachary 网络结构进行社团划分，将网络划分成 2 个社团，社团 1 主要以节点 1 为中心，社团 2 主要以节点 34 为中心。在设定初始值时，将网络划分出的社团个数对应网络中的社团模块度，基于本文算法得到 Zachary 网络社团划分模块度与社团个数之间的关系如图 4 所示。

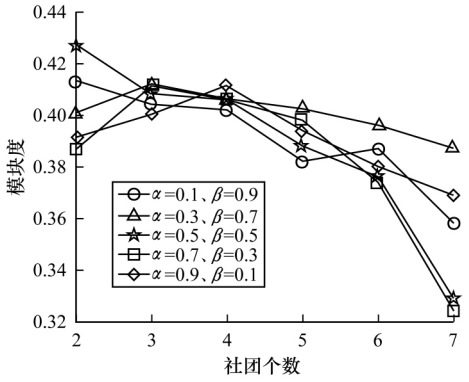


图 4 Zachary 网络社团个数与模块度的关系

Fig. 4 Relationship between the number of community and modularity in Zachary network

3.2.2 Football 网络实验

Football 数据集是一个经典的社团研究数据集，该网络由 115 个球队的 613 场比赛抽象而成，如何根据不同球队之间的实力合理划分球队，并合理安排相应的赛事是该实验关注的重点。因此，采用本文算法对该网络进行社团划分，初步结果如图 5 所示。将网络划分成 10 个社团，模块度 $Q=0.619\,6$ 达到最大值，此时 $\alpha=0.3$、$\beta=0.7$、$\partial=0.4$，划分效果最优。对网络社团进行反复划分实验，最终得到的仿真结果如图 6 所示，通过调节不同的 α 和 β 值，对不同社团个数计算网络模块度，可以看出当 $\alpha=0.3$、$\beta=0.7$ 时，模块度 $Q=0.619\,6$ 达到最大值，可以认为当划分得到 10 个社团时，网络划分结果最佳。

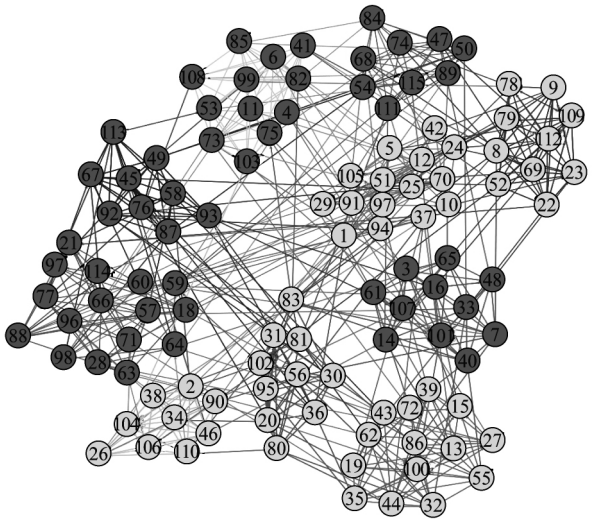


图 5　本文算法对 Football 网络的社团划分结果

Fig. 5　Community division results of the proposed algorithm for Football network

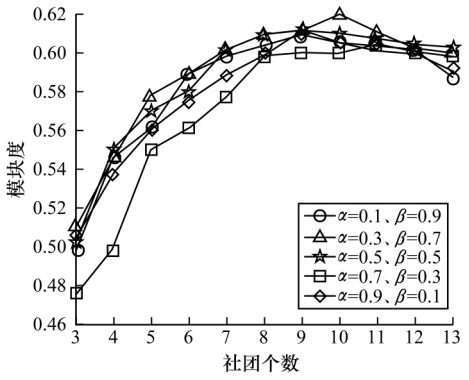


图 6　Football 网络社团个数与模块度的关系

Fig. 6　Relationship between the number of community and modularity in Football network

3.3　结果分析

针对相同网络数据集，将本文 SM-CD 算法社团划分结果与经典的 Newman 算法、GN 算法、NC 算法[15]、基于节点特征向量的复杂网络社团发现算法[16]、IJ-CD 算法[19]、基于节点内聚系数的局部社团发现算法[20]和 GD 算法[21]划分结果进行对比验证，结果如表 2、表 3 所示。

表 2　针对 Zachary 网络的算法划分结果对比

Table 2　Comparison of algorithm division results for Zachary network

算法	社团个数	模块度	时间复杂度
Newman 算法	3	0.371	$O(n^2)$
GN 算法	2	0.401	$O(n^3)$
NC 算法	4	0.402	$O((m+n)m)$
文献[16]算法	3	0.390	$O(n^2)$
IJ-CD 算法	2	0.371	$O(n^2)$
文献[20]算法	5	0.389	$O(n)$
GD 算法	4	0.419	$O(n^2)$
SM-CD 算法	2	0.427	$O(n^2)$

表 3　针对 Football 网络的算法划分结果对比

Table 3　Comparison of algorithm division results for Football network

算法	社团个数	模块度	时间复杂度
Newman 算法	12	0.546	$O(n^2)$
GN 算法	10	0.601	$O(n^3)$
NC 算法	7	0.552	$O((m+n)m)$
IJ-CD 算法	12	0.600	$O(n^2)$
文献[20]算法	17	0.422	$O(n)$
GD 算法	9	0.595	$O(n^2)$
SM-CD 算法	10	0.619	$O(n^2)$

Newman 算法基于贪心算法原则，选取模块度增长最大或者减小最少的社区，将其合并为一个新社区，不断迭代循环，将模块度最大值对应的网络社团作为最优社团划分结果。该算法在降低时间复杂度的同时，准确度也相应降低。从表 2、表 3 结果可以看出，本文算法相比 Newman 算法在模块度上分别提升了 15% 和 13%。

GN 算法的主要思想来源于聚类分裂法，原理是使用网络中的边介数作为相似度的度量。首先计算网络中所有边的介数，找到介数最高的边并将其从网络中移除，重新计算网络中剩余边的介数，不断重复该过程，直至网络中的任一顶点作为一个社区为止。虽然该算法准确度相对较高，但时间复杂度也较高。对比表 2 的 Zachary 网络与表 3 的 Football 网络时间复杂度指标，传统 GN 算法时间复杂度是 $O(n^3)$，本文算法的时间复杂度比 GN 算法低一个数量级。

文献[16]算法采用效能传递思想对社团进行聚类，与本文算法的区别是将网络节点之间的距离倒数作为信息传递效能指标并构建矩阵，从而计算模块度值作为划分依据，算法考虑相对单一。IJ-CD 算法通过改进 Jaccard 相似系数矩阵并选取部分特征值对应的特征向量作为聚类样本，虽然时间复杂度和本文算法相当，但在划分精度上本文算法相对更好。

文献[20]算法与 GD 算法思想类似，通过选取最大度节点作为起始社团，不断搜索其邻居节点，将强社团结构定义作为节点添加的约束条件最大度节点，不断添加至社团中形成新的社团，直至邻居节点集为空时停止。与本文算法相比，虽然文献[20]算法的时间复杂度比本文算法的时间复杂度低，但是其将网络划分为 5 个社团，使用本文算法将网络划分为 2 个社团，且模块度与本文算法相比相差较多，本文算法在划分精度方面优势明显。与此同时，文献[20]算法在划分精度上与其他算法相差很小。因此，本文算法划分精度比文献[20]算法高，更具实用价值。NC 算法虽然对于 Zachary 网络划分的模块度与本文算法很接近，但时间复杂度为 $O((m+n)m)$，其中 m 为边数，在仿真网络中 m 远比 n 要大得多，因此其时间复杂度大

于本文算法。在同等情况下,本文算法的划分精度相比其他算法略高,划分效果更好。

本文针对社团划分采用相同网络数据集,选取经典 Newman 算法、GN 算法和 IJ-CD 算法与本文算法在时间消耗和准确率方面进行比较,如图 7 和图 8 所示。从图 7 可以看出,与经典 Newman 算法、GN 算法相比,本文算法时间消耗更少,优势明显,与文献[20]算法的时间消耗相差较小,但从表 2、表 3 以及图 8 可以看出,本文算法在划分准确率上更具优势,划分效果更好。

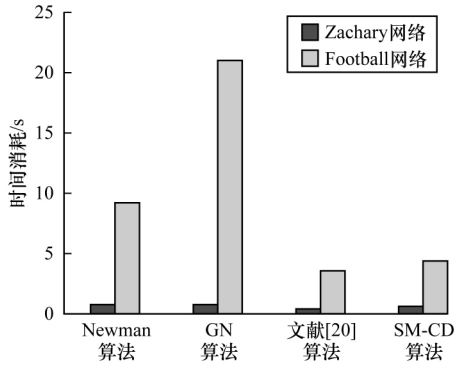


图 7 社团划分算法的时间消耗比较

Fig. 7 Comparison of the time consumption of community division algorithms

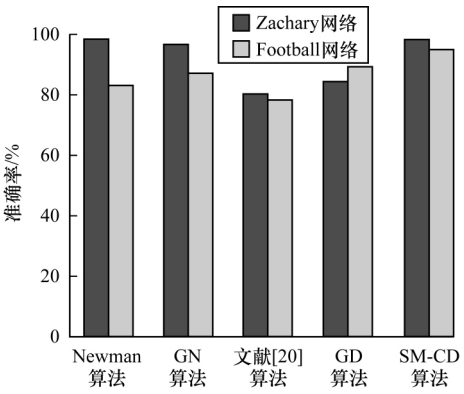


图 8 社团划分算法的准确率比较

Fig. 8 Comparison of the accuracy of community division algorithms

4 结束语

本文定义了网络中的节点自身属性和结构属性,提出一种基于节点多属性相似性聚类的社团划分算法。根据两类属性在网络上的权重不同,在实验中通过不断调整调节因子将网络划分为不同的社团结构并计算相应的模块度。实验结果表明,本文算法能有效提高社团划分的准确率。但由于本文中考虑的网络为无权无向网络,而在实际生活中网络节点之间的连接权重存在差异且可能具有方向性,因此下一步将对有权有向网络社团划分进行研究。

参考文献

[1] LI Peng, HE Tingting, HU Xiaohua. A novel protein complex identification algorithm based on Connected Affinity Clique Extension(CACE)[J]. IEEE Transactions on Nanobioscience, 2014, 13(2): 89-96.

[2] ANDREAS A, LOUKIANOS S, MARTIN-LOPEZ D, et al. Detection of interictal discharges with convolutional neural networks using discrete ordered multichannel intracranial EEG[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2013, 25(12): 2285-2294.

[3] JENNIFER G, HARISH C, ADAM W, et al. Node dominance: revealing community and core-periphery structure in social networks[J]. IEEE Transactions on Signal and Information Processing over Networks, 2016, 2(2): 186-199.

[4] WANG Xiaofan, LIU Yabing. Overview of algorithms for detecting community structure in complex networks[J]. Journal of University of Electronic Science and Technology of China, 2009, 38(5): 537-543. (in Chinese)
汪小帆,刘亚并. 复杂网络中的社团结构算法综述[J]. 电子科技大学学报, 2009, 38(5): 537-543.

[5] NEWMAN M E J. Analysis of weighted networks[J]. Physical Review E, 2004, 70(5): 53-131.

[6] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics Theory and Experiment, 2015, 30(2): 155-168.

[7] KERNIGHAN B W, LIN S. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal, 1970, 49(2): 291-307.

[8] CAPOCCI A, SERVEDIO V D P, CALDARELLI G, et al. Detecting communities in large networks[J]. Computer Science, 2004, 3243: 181-187.

[9] BORGATTI S P. Centrality and network flow[J]. Social Networks, 2005, 27(1): 55-71.

[10] LUO Wenjian, ZHANG Daofu, JIANG Hao, et al. Local community detection with the dynamic membership function[J]. IEEE Transactions on Fuzzy Systems, 2018, 26(5): 3136-3150.

[11] KRIEGEL H P, KRÖGER P, SANDER J, et al. Density-based clustering[J]. Data Mining and Knowledge Discovery, 2011, 1(3): 231-240.

[12] LEICHT E A, HOLME P, NEWMAN M E J. Vertex similarity in network[J]. Physical Review E, 2006, 73(2): 26-45.

[13] TAN Yuejin, WU Jun, DENG Hongzhong. Evaluation method for node importance based on node contraction in complex networks[J]. Systems Engineering—Theory & Practice, 2016, 26(11): 79-83. (in Chinese)
谭跃进,吴俊,邓宏钟. 复杂网络中节点重要度评估的节点收缩方法[J]. 系统工程理论与实践, 2016, 26(11): 79-83.

[14] CHEN Lingjiao, ZHANG Zike, LIU Jinhu, et al. A vertex similarity index for better personalized recommendation[J]. Physica A: Statistical Mechanics and Its Applications, 2017, 466: 607-615.

(下转第 97 页)

[6] CHEN Qian, ZHU Xiaodan, LING Zhenhua, et al. Enhanced LSTM for natural language inference [C] // Proceedings of IEEE Meeting on Association for Computational Linguistics. Washington D. C., USA: IEEE Press, 2017: 1657-1668.

[7] NIE Y, BANSAL M. Shortcut-stacked sentence encoders for multi-domain inference [C] //Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP. Washington D. C., USA: IEEE Press, 2017: 165-178.

[8] CONNEAU A, KIELA D, SCHWENK H, et al. Supervised learning of universal sentence representations from natural language inference data [EB/OL]. [2019-04-10]. https://arxiv.org/abs/1705.02364v5.

[9] TALMAN A, YLIJYRA A, TIEDEMANN J, et al. Natural language inference with hierarchical BiLSTM max pooling architecture [EB/OL]. [2019-04-10]. https://arxiv.org/abs/1808.08762v1.

[10] IM J, CHO S. Distance-based self-attention network for natural language inference [EB/OL]. [2019-04-10]. https://arxiv.org/abs/1712.02047v1.

[11] SHEN Tao, ZHOU Tianyi, LONG Guodong, et al. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling [EB/OL]. [2019-04-10]. https://arxiv.org/abs/1801.10296.

[12] CHENG J, DONG L, LAPATA M, et al. Long short-term memory-networks for machine reading [EB/OL]. [2019-04-10]. https://arxiv.org/abs/1601.06733.

[13] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [EB/OL]. [2019-04-10]. https://arxiv.org/abs/1412.3555.

[14] PARIKH A P, TACKSTROM O, DAS D, et al. A decomposable attention model for natural language inference [EB/OL]. [2019-04-10]. https://arxiv.org/abs/1606.01933v2.

[15] WANG Z, HAMZA W, FLORIAN R, et al. Bilateral multi-perspective matching for natural language sentences [C] // Proceedings of IEEE International Joint Conference on Artificial Intelligence. Washington D. C., USA: IEEE Press, 2017: 4144-4150.

[16] KIM S, KANG I, KWAK N. Semantic sentence matching with densely-connected recurrent and co-attentive information [EB/OL]. [2019-04-10]. https://arxiv.org/abs/1805.11360.

[17] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2019-04-10]. https://arxiv.org/abs/1409.0473.

[18] HE H, LIN J J. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement [C] // Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA: [s. n.], 2016: 937-948.

[19] WILLIAMS A, NANGIA N, BOEMAN S R, et al. A broad-coverage challenge corpus for sentence understanding through inference [EB/OL]. [2019-04-10]. https://arxiv.org/abs/1704.05426v2.

[20] PENNINGTON J, SOCHER R, MANNING C D, et al. GloVe: global vectors for word representation [EB/OL]. [2019-04-10]. https://www.aclweb.org/anthology/D14-1162.

编辑　金胡考

(上接第 90 页)

[15] LUO Mingwei, YAO Hongliang, LI Junzhao, et al. A hierarchical division algorithm for community based on node dissimilarity [J]. Computer Engineering, 2014, 40(1): 275-279. (in Chinese)
罗明伟,姚宏亮,李俊照,等. 一种基于节点相异度的社团层次划分算法[J]. 计算机工程,2014,40(1):275-279.

[16] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.

[17] 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(3): 36-76.

[18] WEI Qingjie, LI Jingteng, WANG Yu. Detection algorithm for microblogging networks community based on user closeness [J]. Computer Applications and Software, 2016, 33(9): 254-258. (in Chinese)
韦庆杰,李京腾,汪雨. 基于用户紧密度的微博网络社区发现算法[J]. 计算机应用与软件,2016,33(9):254-258.

[19] ZHANG Meng, LI Lingjuan. Community division algorithm based on improved Jaccard similarity coefficient matrix [J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science), 2018, 38(6): 96-102. (in Chinese)
张猛,李玲娟. 基于改进的 Jaccard 相似系数矩阵的社团划分算法[J]. 南京邮电大学学报(自然科学版),2018,38(6):96-102.

[20] ZHAO Wentao, ZHAO Haohao, MENG Lingjun. Local community detection algorithm based on node cohesive clustering coefficient [J]. Computer Applications and Software, 2016, 33(12): 270-274. (in Chinese)
赵文涛,赵好好,孟令军. 基于节点内聚系数的局部社团发现算法[J]. 计算机应用与软件,2016,33(12):270-274.

[21] CHEN Dongming, WANG Yunkai, HUANG Xinyu, et al. Community detection algorithm for complex networks based on group density [J]. Journal of Northeastern University (Natural Science), 2019, 40(2): 186-191. (in Chinese)
陈东明,王云开,黄新宇,等. 基于社团密合度的复杂网络社团发现算法[J]. 东北大学学报(自然科学版),2019,40(2):186-191.

编辑　陆燕菲