

## 基于 FCM 的复杂网络重叠社团结构发现算法

潘惠勇<sup>1</sup>, 王 鹏<sup>2</sup>, 张慧乐<sup>3</sup>

(1 中原工学院 计算机学院, 河南 郑州 450007; 2 复旦大学 信息科学与工程学院, 上海 200433;

3 河南联通漯河分公司, 河南 漯河 462000)

**摘 要:** 复杂网络中的社团结构发现是对网络数据集进行数据挖掘的普遍性问题. 针对网络中大量存在的重叠社团现象, 提出了基于 FCM 的发现重叠社团结构算法, 并进一步在 NG 模块度的基础上, 给出了评价重叠社团结构的模块度函数. 算法首先将网络的节点映射成欧氏空间的节点, 再以此做模糊聚类得到各重叠社团结构, 根据模块度函数选择最佳重叠社团结构. 最后, 在经典网络上的实验结果表明, 算法能够得到满意度高的重叠社团结构, 而且时间复杂度较低.

**关键词:** 复杂网络; 社团发现; FCM 聚类; 模块度

**中图分类号:** N94

**文献标识码:** A

**文章编号:** 1000-7180(2011)12-0111-04

## Fast Partitioning Algorithm for Detecting Communities in Complex Networks

PAN Hui-yong<sup>1</sup>, WANG Peng<sup>2</sup>, ZHANG Hui-le<sup>3</sup>

(1 School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China;

2 School of Information Science and Technology, Fudan University, Shanghai 200433;

3 Luohe branch Company of China Unicom, Luohe 462000)

**Abstract:** Detection of overlapping communities in a complex network is a general problem in data mining of network data sets. In view of the phenomenon of a lot of overlapping communities in real networks, we devise the novel algorithm of detection of overlapping community based on Fuzzy c-means clustering. Further more, we propose the improved modularity to evaluate the overlapping community structure based on NG modularity. Firstly, network nodes were mapped into the nodes in Euclidean space. Secondly, the nodes were clustered by Fuzzy c-means algorithm. Thirdly, the best overlapping community was selected by computing the function of improved modularity. Finally, the developed algorithm has been tested on two common real-world networks. Computational results demonstrate that the improved modularity and the algorithm have better partitioning ability and lower time complexity.

**Key words:** complex networks; detecting communities; FCM clustering; modularity

### 1 引言

互联网、神经网络、食物链网、新陈代谢网、社会网络、物流与供应链网络以及通信与分布式网络都是对现实世界中的复杂系统抽象而形成的网络<sup>[1-2]</sup>. 这些网络几乎都存在一定的结构, 这种结构

被认为是复杂网络的一个共同的性质——社团结构. 社团结构指整个网络由许多个社团或组构成, 社团内部的连接比较紧密, 而社团之间的连接比较松散<sup>[3]</sup>.

社团结构对于研究复杂网络的特征、功能和结构有重要意义, 在自然科学与社会科学各相关研究

领域都有着广泛的应用<sup>[3-6]</sup>. 谱平分法是近年来在该领域主要的研究方法<sup>[1-3]</sup>. 谱分割法主要包括了基于 Laplace 矩阵、基于 Normal 矩阵的方法. 网络的 Laplace 矩阵和 Normal 矩阵是谱分割法的基础, 网络的 Laplace 矩阵  $L = K - A$ , 每个节点的度值构成了对角矩阵  $K$ , 邻接矩阵  $A$  表示网络的连接关系. Laplace 矩阵第二小特征值对应的特征向量可作为网络划分的支撑, 可依据其中的正负元素将网络一分为二. 此算法仅适应于二分网络的情形. 将网络划分为多个社团则需利用网络的 Normal 矩阵, 若社团有  $g$  个社团, 那么 Normal 矩阵  $N = K^{-1}A$  存在  $g-1$  个靠近 1 的特征值, 与其它特征值有显著区别, 此  $g-1$  个特征值对应的特征向量表现出一个显著特征: 以特征向量作为节点的坐标, 社团内部节点之间的距离明显小于不同社团节点之间的距离, 那么社团划分也就轻而易举了. 多个社团的情况明显适合于此算法, 算法的时间复杂度主要在于特征值和特征向量的计算.

本文在分析上述算法的基础上, 提出一种基于 Fuzzy c-means (FCM) 的重叠社团结构发现算法. 本文算法借鉴谱平分法思想, 通过谱方法将网络节点映射到欧氏空间上之后, 用模糊聚类得到重叠社团结构.

## 2 基于 FCM 的重叠社团结构发现算法

### 2.1 模块度函数

至今已经有若干种算法可用于复杂网络的社团结构划分, 那么就需要一种评价机制, 来分辨各种算法得到社团结构的优劣. Newman 和 Girvan 提出了一种评价网络社团结构划分的函数——模块度<sup>[7]</sup> (Modularity). 模块度函数表示为下式:

$$Q(P_t) = \sum_{c=1}^t (e_c - a_c^2) = \text{Tr}(e) - \|e^2\| \quad (1)$$

式中,  $P_t$  表示网络划分的一种结果, 此划分包含了  $t$  个社团, 对称矩阵  $e$  为  $t$  阶, 其中的元素  $e_{ij}$  表示社团  $i$  与社团  $j$  之间的连接边数占网络总连接边数的比例,  $a_i = \sum_{c=1}^t e_{ci}$ ,  $\text{Tr}(e) = \sum_i e_{ii}$  是矩阵  $e$  的迹,  $\|e^2\|$  表示矩阵  $e^2$  的和范数. 模块度函数  $Q$  的衡量的标准是: 复杂网络中社团内部的连边所占之比例与随机连接情况下社团内部连边所占比例之期望值的差值. 差值的大小, 对应于网络社团结构的显著与否. 实际复杂网络中,  $Q$  的最大值的范围一般在 0.3 ~ 0.7, 更大的值较少出现. 式(1)也可表述为式(2), 二式是等价的.

$$Q(P_t) = \frac{1}{2m} \sum_{v,w} [A_{vw} - \frac{k_v k_w}{2m}] \delta(c_v, c_w) \quad (2)$$

式中,  $m$  为网络的总边数;  $A$  为网络的连接矩阵,  $A_{vw} = 1$  代表节点与节点  $w$  之间有连边,  $A_{vw} = 0$  则表示无连边;  $k_v$  为节点  $v$  的度;  $c_v$  为节点  $v$  所在的社团; 当节点  $v$  和节点  $w$  在同一个社团中时,  $\delta(c_v, c_w) = 1$ , 否则  $\delta(c_v, c_w) = 0$ .

NG 模块度广泛应用于社团结构的评价中, 但是对于重叠社团结构却并不适用, 因重叠社团中的节点或边可能同时属于若干个社团, 这样就使 NG 模块度不能准确的衡量社团结构的优劣. 本文在 NG 模块度的基础上提出了改进, 函数如下式:

$$Q'(P_t) = \frac{1}{2m} \sum_{v,w} [A_{vw} - \frac{k_v k_w}{2m}] \frac{1}{S_v S_w} \delta(c_v, c_w) \quad (3)$$

式中,  $S_v$  代表节点  $v$  所属的社团个数. 如果节点  $v$  是重叠节点, 它属于若干个社团; 节点  $v$  和  $w$  之间有连边, 且同属于某一个社团. 那么边  $(v, w)$  对社团的贡献实际上是较小的, 而且节点  $v$  所属的社团数越多, 贡献越小. 上式对 NG 模块度加入  $1/S_v S_w$ , 恰好使社团中重叠节点连接边的贡献减小, 符合衡量重叠社团结构的要求.

### 2.2 FCM 聚类

FCM 是用隶属度确定每个数据点属于某个聚类的程度的一种聚类算法. FCM 把  $n$  个向量  $x_i$  ( $i = 1, 2, \dots, n$ ) 分为  $c$  个模糊组, 并求每组的聚类中心, 使得非相似性指标的价值函数达到最小. 与引入模糊划分相适应, 隶属矩阵  $U$  允许有取值在  $[0, 1]$  间的元素. 不过, 加上归一化规定, 一个数据集的隶属度的和总等于 1.

$$\sum_{i=1}^c u_{ij} = 1, \forall j = 1, \dots, n \quad (4)$$

那么, FCM 的目标函数就是式(5)的一般化形式:

$$J(U, c_1, \dots, c_c) = \sum_{i=1}^c J_i = \sum_{i=1}^c \sum_j^n u_{ij}^m d_{ij}^2 \quad (5)$$

这里  $u_{ij}$  介于  $[0, 1]$  间;  $c_i$  为模糊组  $I$  的聚类中心,  $d_{ij} = \|c_i - x_j\|$  表示聚类  $I$  的几何中心与数据点  $j$  之间的欧氏距离; 其中加权指数  $m \in [1, \infty)$ . 通过在循环迭代过程中不断更新隶属度  $u_{ij}$  和聚类中心  $c_i$ , FCM 在优化目标函数的同时是实现了聚类. 此迭代过程最终收敛于局部最小值或鞍点  $J_m$ .

### 2.3 算法描述

本文算法是一种基于谱映射和 FCM 的重叠社

团结划分算法,最佳社团结构的评价标准采用改进的模块度函数,即最佳的社团数目和社团结构通过模块度函数来确定.算法可表述为如下步骤:

Input: 网络的邻接矩阵  $A$ , 社团个数上界  $T$  和选取重叠节点的阈值  $\lambda$ ;

Output: 社团划分结果  $P_{\max}$ ;

Step1 谱映射:

Step1.1 计算对角矩阵  $K=(k_{ii})$ , 其中是  $k_{ii}$  节点  $i$  的度;

Step1.2 求一般线性系统  $Ax=\lambda Kx$  的前  $T$  个特征向量, 得到特征矩阵  $E_T=[e_1, e_2, \dots, e_T]$ ;

Step2 FCM 算法( $2\leq t\leq T$ ):

Step2.1 从矩阵  $E_T$  中得到矩阵  $E_t=[e_1, e_2, \dots, e_t]$ ;

Step2.2 用欧氏距离规格化  $E_t$  中的各行向量;

Step2.3 以  $E_t$  中每个行向量作为数据单位, 用 FCM 算法对  $n$  个数据进行聚类, 得到软分配矩阵  $U_t$ ;

Step3 最佳社团划分:

Step3.1 将软分配矩阵  $U_t$  中大于阈值  $\lambda$  的元素更新为 1, 小于  $\lambda$  的元素更新为 0, 得到实际分配矩阵, 即重叠社团结构划分;

Step3.2 对各重叠社团结构计算式(3), 选择模块度最大值所对应的划分为最佳重叠社团结构.

2.4 时间复杂度分析

当面对大规模复杂网络时, 计算复杂性是必须要考虑的关键问题. 已有的一些方法能够快速地计算矩阵的特征系统和 FCM 聚类. 如果采用 IRLM 算法计算矩阵的  $T-1$  个特征向量, 文献[8]中的高效方法实现 FCM 聚类, 可以得到在最差情况下时间复杂度分别为  $O(mTh+nT^2h+T^3h)$  和  $O(nT^2)$ , 其中  $m$  是网络的所有边数,  $h$  是收敛所需要是次数. 那么对于大规模稀疏网络有  $m\sim n$  和  $T\ll n$ , 则在此情形下总的计算复杂性可大致认为是  $n$  的线性函数.

3 实验

本文算法以 Matlab 为实验平台, 所用数据为空手道俱乐部 Zachary 网络和大学足球联盟 Football 网络.

3.1 空手道俱乐部 Zachary 网络

社会网络中最经典案例就是空手道俱乐部网络. 由于该俱乐部主管与校长的争执, 而使该俱乐部

分裂成了两个小俱乐部. 该俱乐部主管和校长分别由图 1 中的节点 1 和节点 33 代表, 分裂后的两小俱乐部分别由图中左侧的 18 个节点和右侧的 16 个节点代表. 在社团个数上界  $T=10$  和阈值  $\lambda=0.25$  时, 运行本文算法后得到的最佳重叠社团结构如图 1 所示, 社团数目为 4, 其中节点 1、节点 9、节点 10 和节点 24 为重叠节点. 而用 K-means 算法得到的最佳非重叠社团结构如图 2 所示, 同样划分为 4 个社团, 与图 1 的不同仅在于节点 1、9、10 和 24 只属于 1 个社团. 图 3 是 NG 模块度和改进模块度在不同社团数目时的取值, 从中可看出两种模块度的变化趋势是一致的, 重叠社团结构的模块度值比非重叠社团结构的值略高, 这是由于包含在社团中边数增加导致的.

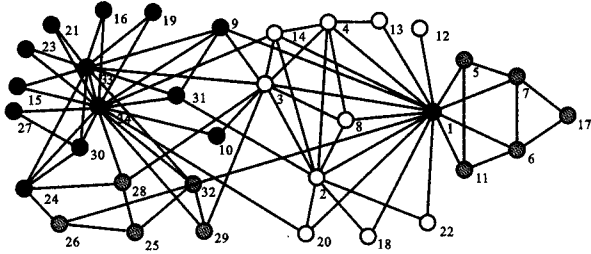


图 1 Zachary 网络划分社团的最佳重叠社团结构

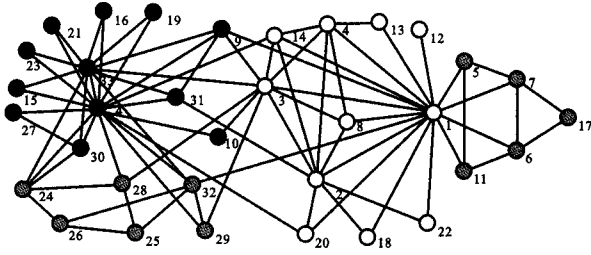


图 2 Zachary 网络划分社团的最佳非重叠社团结构

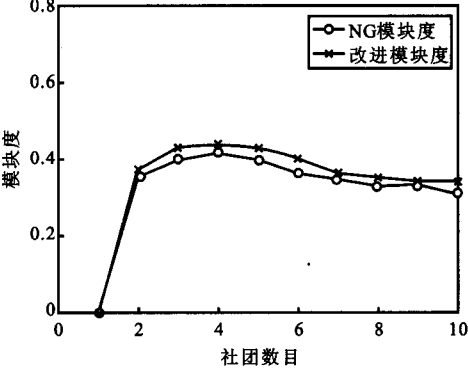


图 3 Zachary 网络两种社团结构对应的两种模块度值变化图

3.2 美国大学足球联盟 Football 网络

本文第二个实验研究的网络是美国大学足球联盟, 115 支球队被抽象为网络中的 115 个节点, 而 613 场赛事就被表示为 613 条边. 整个联盟按照它

们之间的赛事关系被划分成若干分区联盟,每个分区联盟由 8~12 支球队构成,分区联盟内的球队的比赛比分分区联盟之间球队的比赛多. 参数  $T$  为 24,  $\lambda$  为 0.10, 运行本文算法得到最大模块度时, Football 网络的重叠社团结构如图 4 所示, 被划分为 10 个社团, 其中节点 12、25、29、37、51、59、64、70、91 和 98 共 10 个节点为重叠节点. 图 5 中给出了社团数目与两种模块度值的相对关系, 从图中可知, 当 Football 网络所划分的社团数目不断增加时, 模块度值也同时变大, 但达到峰值( $t=10$ )后, 改进模块度值开始变小. 非重叠社团结构在划分为 11 个社团时 NG 模块度值达到最大值, 其中节点 12、25、29、51、70 和 91 为第 11 个社团, 而这 6 个节点在重叠社团结构中是重叠节点.

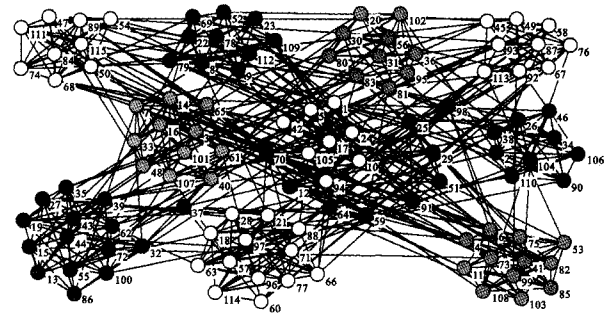


图 4 美国大学足球联盟网络的最佳重叠社团结构( $t=10$ )

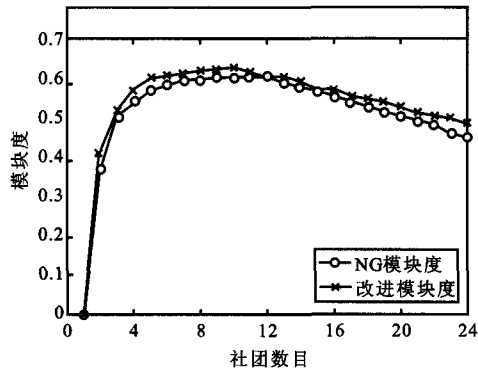


图 5 Football 网络两种社团结构对应的两种模块度值变化图

4 结束语

复杂网络中的社团结构发现是对当今研究复杂系统领域的热点. 针对网络中大量存在的重叠社团现象, 本文借鉴谱平分算法的思想, 提出了基于 FCM

的发现重叠社团结构算法, 算法首先将网络的拓扑结构转化为欧氏空间聚类问题, 再以此做模糊聚类得到各重叠社团结构; 进一步在 NG 模块度的基础上, 给出了改进模块度函数用于评价复杂网络的重叠社团结构; 文章分析了算法的时间复杂度, 说明了对于大规模稀疏网络, 本文算法同样能快速得到重叠社团结构, 体现出算法具有较高的效率; 最后, 在经典网络上的实验结果表明, 算法能够得到满意度高的重叠社团结构.

参考文献:

[1] 汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用[M]. 北京: 清华大学出版社, 2006: 162-191.

[2] Newman M E J. Modularity and community structure in networks[J]. Proc Natl Acad Sci USA, 2006, 103(23): 8577-8582.

[3] 李晓佳, 张鹏, 狄增如, 等. 复杂网络中的社团结构[J]. 复杂系统与复杂性科学, 2008, 5(3): 19-42.

[4] 王存睿, 段晓东, 刘向东, 等. 一种解决网络社区划分物理方法[J]. 微电子学与计算机, 2010, 27(9): 33-36.

[5] 刘鑫炎, 黄美发, 刘夫云, 等. 基于复杂网络的产品族结构树的研究与实现[J]. 微电子学与计算机, 2009, 26(11): 177-180.

[6] 林煜熙, 梁颖红, 韩艳, 等. 基于加权复杂网络的中文文档关键词抽取[J]. 微电子学与计算机, 2009, 26(10): 65-68.

[7] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Phys Rev E, 2004, 69(2): 26-113.

[8] Kelen J F, Hutcheson. T. Reducing the time complexity of the fuzzy c-means algorithm[J]. IEEE Transaction on Fuzzy Systems, 2002, 10(2): 263 - 267.

作者简介:

潘惠勇 男, (1977-), 硕士, 讲师. 研究方向为复杂网络与智能计算.

王 鹏 男, (1981-), 博士研究生. 研究方向为复杂网络分析与控制.

张慧乐 女, (1973-), 本科, 工程师. 研究方向为计算机通讯网络.