基于改进的ACCA的复杂网络社团结构发现

臧 丽,王 红,杨通辉
(山东师范大学 信息科学与工程学院,山东 济南 250014)

摘 要:复杂网络社团结构划分日益成为近年来复杂网络的研究热点,到目前为止,已经提出了很多分析复杂网络社团结构的算法。该文在聚类算法的基础上,提出了一种基于改进的ACCA的复杂网络社团结构发现方法。该文提出的方法的好处是社团数目不用事先被指定,并且此算法最大的优点就是能获取全局最优解。通过ZacharyKarate Club经典模型验证了该算法的可行性和有效性,实验结果表明,该算法能成功地发现各个社团,是一种行之有效的网络社团发现算法。

关键词:复杂网络;社团结构;聚类算法;改进的蚁群聚类算法
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2012)10-0129-04

Community Structure Detection in Complex Networks Based on Improved ACCA

ZANG LI, WANG Hong, YANG Tong-hui
(School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China)

Abstract: Community structure identification has been one of the most popular research areas in recent years and there has been many algorithm proposed so far to detect community structures in complex networks. In this paper, an algorithm for detecting community structures in complex network is presented, which is based on the improved ant colony clustering algorithm, based on the clustering algorithm. The benefit of the method proposed in this paper is number of community does not need to specified, and the biggest advantage of this algorithm is that it can obtain the global optimal solution. The feasibility and effectiveness of the algorithm have been validated through the ZacharyKarate Club classical model, experimental results show that the algorithm can successfully find each community, is a kind of effective algorithm to find network community.

Key words: complex networks; community structure; clustering algorithm; improved ant colony clustering algorithm

0 引 言

现实世界中的大多数复杂系统都可以用复杂网络[1]进行描述。复杂网络是由大量的节点和连接节点间的边构成的,其中复杂网络中的节点表示复杂系统中的不同个体,而节点之间的边则表示复杂系统中个体之间一种关系。现实世界中的许多系统都可以通过复杂网络进行处理,比如互联网、社会网以及生物网等等[2]。

大量的实证表明,大部分的实际网络是由许多类型的节点组合在一起的而并不是由许多性质相同的节点随机地连接在一起的。人们发现社团结构[3]是复杂网络的重要特性之一,是指在整个网络中,存在若干个社团,同社团内的节点之间连接紧密,不同社团之间的节点连接相对比较稀疏[3~5]。复杂网络的社团结构这一重要特性已被广泛关注。自动搜寻或发现大型复杂网络中的社团结构,对于更好地分析、利用以及挖掘这些网络具有重要的实用价值。

“物以类聚,人以群分”,聚类分析[6,7]是数据挖掘最常用的方法之一,是一种组合优化问题。聚类技术[8]通过分析节点间的相似性,将相似性较高的节点尽可能地划分在一起,形成一个特定的类,而将其它的节点尽量划分到不同的类中。由于社团和数据聚类(Data clustering)中的“簇”有相似之处,人们也将社团结构特性称为聚类特性。

蚁群聚类算法最早是由Deneubourg[9]提出的一种仿生聚类方法,是一种具有很强的通用性和鲁棒性的启发式智能优化算法,现在正被广泛地研究与应用在聚类分析中。该文对已有的一种蚁群算法进行了介绍并加以改进,提出了一种基于改进的ACCA的复杂网

收稿日期:2012-02-17;**修回日期**:2012-05-21
基金项目:国家自然科学基金资助项目(60970004);山东省研究生教育创新计划项目(SDYY10059);山东师范大学研究生重点课程项目
作者简介:臧 丽(1986-),女,山东潍坊人,硕士研究生,研究方向为复杂网络;王 红,教授,硕士生导师,研究方向为复杂网络、移动计算。

络社团结构发现方法。将每次 ACCA 的迭代过程中的一定数量的最优解予以保留并将其加入到下一次的循环过程中,这样就会使算法的性能得到提高。从而使得社团发现质量得到较大的提高。

1 社团结构

在复杂网络中,节点之间的关联是通过边来体现的。文献[10]给出了社团结构的定义。假设网络 G 的邻接矩阵 $\boldsymbol{W}$,w_{ij} 是 $\boldsymbol{W}$ 中的元素,节点 i 的度 $k_i = \sum_j^N w_{ij}$。考虑网络 G 的一个子网络 $G' \subset G$,且该子网络 G' 包含有节点 i。可以认为,节点 i 的度包含有 2 个分量,即:

$$k_i(G') = k_i^{in}(G') + k_i^{out}(G')$$

其中,$k_i^{in}(G') = \sum_{j \in G'} w_{ij}$ 表示节点 i 连接子网络 G' 中其他节点的边的条数;$k_i^{out}(G') = \sum_{j \in G'} w_{ij}$ 表示节点 i 连接子网络 G' 外其他节点的边的条数。

2 聚类算法

聚类算法是数据挖掘领域最为常见的技术之一,它依据“物以类聚”,分析数据对象中数据间的相似性,并根据相似性确定数据对象所属的类,使得同一类内的对象之间的相似性尽可能大,而不同类之间的对象相似度尽可能小。以下即为其一般的数学描述:

假设模式样本集 $X = \{X_i \mid X_i = (x_{i1}, x_{i2}, \cdots, x_{id}), i = 1,2,\cdots,n\}$ 有 n 个样本,其中样本 $X_i = \{X_i \mid X_i = (x_{i1}, x_{i2}, \cdots, x_{id})\}$ 为 d 维向量,样本集有 k 个模式分类。

聚类问题就是要找一个划分 $C = \{C_1, C_2, \cdots, C_k\}$,满足:$X = \bigcup_{i=1}^{k} C_i, C_i \neq \Phi, i = 1,2,\cdots,k; C_i \cap C_j = \Phi, i \neq j = 1,2,\cdots,k, i \neq j$;使类内离散度之和达到最小,即:

$$F = \min \sum_{j=1}^{k} \sum_{X_i \in C_j} d(X_i, m_j) \tag{1}$$

$d(X_i, m_j)$ 表示第 C_j 类中模式样本 X_i 到该类中心 m_j 的欧氏距离。其定义为:

$$d(X_i, m_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - m_{jk})^2} \tag{2}$$

其中 x_{ik}, m_{jk} 是样本 X_i 和类中心 m_j 在节点矩阵中的坐标,根据公式可知,如果两个节点之间的欧氏距离越小,那么表示它们越接近,否则的话就越远。

3 基于改进的 ACCA 的复杂网络社团结构发现

3.1 蚁群聚类算法

蚁群算法[11]是一种以仿生学为启发的群体随机搜索算法。蚁群算法的基本原理为:生物界的蚂蚁在觅食的过程中,会在其经过的路径上留下信息素来进行个体间的信息交互,能够使在一定范围内的其他蚂蚁觉察到并影响它们的行为,信息素自身会随着时间的流逝而渐渐挥发,路径上信息素的强度与蚂蚁选择该路径的几率这两者之间存在正比的关系。显然后面的蚂蚁会倾向于选择信息素强的那条路径,其留下的信息素随着路径上的蚂蚁数的增多而增多,导致后来的蚂蚁选择该路径的可能性也就越大,从而形成一种正反馈过程。整个过程分为适应和协作两块。在适应阶段,各个候选解会根据路径上残留的信息素强度对路径的选择进行不断的调整;在协作阶段,蚂蚁之间受到信息素的影响,会集中到信息素浓度较高的路径上去,这样就能得到更好的解。

基于上述聚类算法和基本蚁群算法,可以使用蚁群聚类算法 ACCA(Ant Colony Clustering Algorithm),来解决复杂网络社团划分问题。

3.2 ACCA 中相关参数的描述说明

在蚁群聚类算法 ACCA 中,用 $d(i,j)$ $(i = 1,2,\cdots,n; j = 1,2,\cdots,k)$ 来表示模式样本 X_i 和聚类中心 m_j 之间存在的欧氏距离,启发函数 $\eta(i,j)$ 与 $d(i,j)$ 成反比关系,将其定义为:$\eta(i,j) = 1/d(i,j)$ $(i = 1,2,\cdots,n; j = 1,2,\cdots,k)$。样本 X_i 到聚类中心 m_j 之间的信息素为 $Tau(i,j)$ $(i = 1,2,\cdots,n; j = 1,2,\cdots,k)$(第 i 个样本归属于第 j 个聚类的概率)。求出式(1)中 F 的最小值作为目标函数。按照式(3)计算蚂蚁搜索模式样本所归属的聚类中心的概率 $P_r(i,j)$,其中:参数 α, β 分别表示信息启发式因子,期望启发式因子。将随机选取的 k 个模式样本点作为初始的聚类中心。蚂蚁对整个空间进行搜索形成一个聚类,类 C_j 中模式样本各个属性的均值作为聚类中心 m_j 各个分量的值,其计算公式如式(4)。

$$P_r(i,j) = \frac{[Tau(i,j)]^{\alpha} \times [\eta(i,j)]^{\beta}}{\sum_{l=1}^{k} [Tau(i,l)]^{\alpha} \times [\eta(i,l)]^{\beta}} \tag{3}$$

$$m_j = \frac{1}{|C_j|} \sum_{X_i \in C_j} X_i \tag{4}$$

该文仅对一次迭代过程中的 m 只蚂蚁找到的最优聚类结果,进行信息素的修改增加,其它的进行减少。也就是采用全局信息素更新方式对信息素进行更新,即按式(5)进行更新。其中,用 $\Delta Tau(i,j)$ 来表示信息素的增量。在最优聚类结果中,各模式样本和其聚类中心之间的距离之和用 lmb 来表示,Q、$\rho(0 < \rho < 1)$ 为参数,信息素的挥发程度用 $1 - \rho(0 < \rho < 1)$ 来表示。

$$Tau(i,j) = \rho Tau(i,j) + \Delta Tau(i,j) \tag{5}$$

其中:

$$\Delta Tau(i,j) = \begin{cases} Q/lmb & \text{模式样本 } X_i \text{ 属于类 } C_j \\ 0 & \text{否则} \end{cases}$$

$$lmb = \sum_{j=1}^{k} \sum_{X_i \in C_j} d(X_i, m_j)$$

3.3　基于改进的 ACCA 的复杂网络社团结构发现

蚂蚁 a 随机选择一个模式样本 X_i 作为起点,蚂蚁按概率式(3)计算模式样本分配到各类的概率 $P_r(i,j)$,依据概率确定所属的类 b,$1 < b < k$,将样本 X_i 加入到类 C_b 中,然后对上述步骤重复进行循环操作,直到蚂蚁 a 将所有的样本空间搜索完毕,独立地构造一个解。但在进化初期,聚类中心可能并不是最优聚类中心,这样就会使更好的解不利于被发现,因此过早地收敛于非最优解,容易出现停滞现象。

为了搜索到最优解,向全局最优解进化,同时为了增加搜索的随机性,兼顾解空间的各种情况,将蚂蚁确定模式样本所属类的方法进行相应的改进:给定参数 $q_0(0 \leqslant q_0 < 1)$,生成随机数 $q(0 \leqslant q < 1)$。

(1) 当 $q \leqslant q_0$ 时,按式(6)计算,即蚂蚁按最近的原则将模式样本分配给 $[Tau(i,j)]^{\alpha} \times [\eta(i,j)]^{\beta}$ 达到最大的类。

$$\text{Max}([Tau(i,j)]^{\alpha} \times [\eta(i,j)]^{\beta}) \qquad (6)$$

(2) 当 $q > q_0$ 时,先将模式样本加入到各聚类中心的几率按式(3)进行计算得出,然后再按照轮盘赌的方法去确定其所属的类,这样在增加了随机搜索的可能性的同时又考虑到了几率的大小。

以下描述即为轮盘赌的伪代码:

a) 生成一随机数 z,$0 \leqslant z < 1$,$j = 0$,$s = p_j$;

b) 如果 $s \geqslant z$,则转 d;

c) $j = j + 1$,$s = s + p_j$,转 b;

d) 选取聚类 j 并将其输出,终止。

对参数 q_0 通过动态改变的方式进行调整。具体说就是:初期,基于强度较小的初始信息素引导作用不大的问题,赋予 q_0 一较大的值,这样就会使蚂蚁按最近的原则对模式样本进行分类,以便减少蚂蚁的搜索时间,同时更快地收敛于较好的解;末期,赋予 q_0 的值相对较小,通过采用轮盘赌的方法,使蚂蚁能够按照概率来对模式样本进行分类,通过信息素的影响,可以使随机搜索的可能性得到增加而且使停滞现象得以改善,最终达到最优解。

另外,基于每次迭代过程中最优适值对下一代影响的考虑,在迭代过程中对上一代的有效信息予以有效利用以便对下一代进行计算。详言之,就是在每次迭代过程中,将具有最优适值的固定数量的解予以保留,并使其加入到下一次的迭代循环中。这样,能够对算法的性能予以提高。

以下所述即为该文提出的算法的具体步骤:

A:对每个参数设定初始值:事先将各个参数 α、β、ρ、Q、q_0 的值进行初始化,设定蚂蚁的数目为 R,T 为最大迭代次数,L 和 E 分别表示局部搜索保留的数量和每次迭代保留的最优适值数量,用 $Tau(i,j)$ 来表示初始信息素的矩阵。

B:从数据集中任意选取 k 个模式样本点作为初始的聚类中心并对选取的 k 个聚类中心进行初始化。

C:按照公式(2)对各个聚类中心与模式样本之间的距离 $d(i,j)$ 进行计算得出,并根据其与启发函数 $\eta(i,j)$ 之间的反比关系得出 $\eta(i,j)$ 的函数值。

D:按照 3.3 节的方式,各个蚂蚁对模式样本进行分类并形成一个独立的解。对 q_0 进行动态调整,初期赋予其较大的值,末期对其赋值相对较小。

E:当数据集中的蚂蚁对模式样本进行分类后,再根据式(4)和新形成的解,对新的 k 个聚类中心和目标函数 F 进行计算。

F:在对 N 只蚂蚁所求出的目标函数值进行比较的基础上,通过将最优的 L 个解予以保留的方式,对其进行局部搜索。

G:对信息素矩阵及全局信息素进行更新(3.2 节中信息素的处理方式),对前最优的 l 个解予以保留并使其加入到下一次的迭代循环过程中。

H:如果达到了最大迭代次数,满足终止条件,蚂蚁停止移动;否则增加进化代数 $T = T + 1$,然后,转入步骤 3 继续移动。直到满足结束条件 $T > Nc$ 或当前解已稳定。

由于蚂蚁总是向着相似度高的节点方向移动,因此当网络上的蚂蚁集中在一些相似度高的节点上时,表明算法已经停止。根据社团的定义可以判定,这些有蚂蚁存在的节点,就组成了一个社团。

4　实验结果

Zachary Karate Club[12] 是用于复杂网络的社团结构分析的一个经典模型,通过对它进行验证来测试该文中所提出算法的可行性。

早在 20 世纪 70 年代初期, Zachary 观察发现了美国一所大学中的 karate club 内成员间的交互关系。他构造了这些成员之间的关系网,基于他们在整个 club 内及外部的社会关系。但是在观察过程中,针对是否应该抬高 club 收费的问题,该 club 的主管与校长之间产生了争执, 所以该 club 分裂成了两个分别以主管和校长为中心的小的 club。

通过该文的算法对 Zachary Karate Club 网络进行分析, 在多次的分析计算中,划分该复杂网络的社团结构的准确性大于 90% 。在个别不准确的情况中, 节

点 10 被错分到了别的社团中,见图 1。

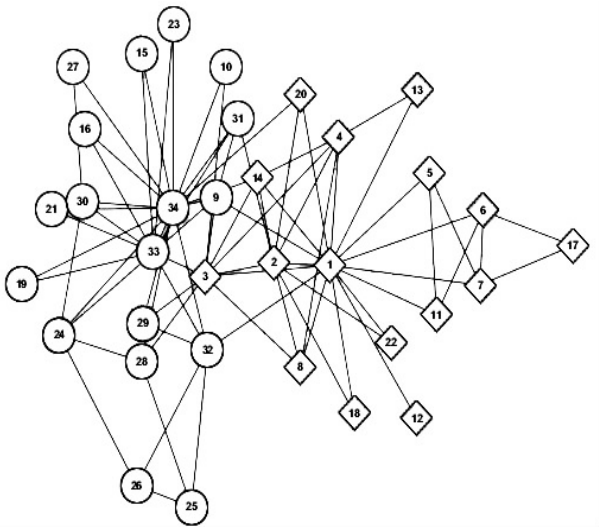

图 1 某次计算 Zachary club 网络的社团划分结果

5 结束语

复杂网络的研究已经备受关注,其社团结构发现已经成为一项很有挑战性的研究课题。该文用改进的 ACCA 来分析网络社团结构的划分,此方法的好处是社团数目不需要提前被指定,并且能够获取全局最优解是此算法最大的优点。实验表明用该算法分析复杂网络的社团结构是可行的。由于相关参数影响算法的性能,所以需进行深入研究。

参考文献:

[1] 汪小帆,李 翔,陈关荣. 复杂网络理论及应用[M]. 北京:清华大学出版社,2006.

[2] Hu Y Q,Li M H,Zhang P,et al. Community detection by signaling on complex networks[J]. Physical Review E,2008,78(1):016115.

[3] Girvan M,Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America,2002,99(12):7821-7826.

[4] Guimera` R,Amaral L A N. Functional cartography of complex metabolic networks[J]. Nature,2005,433(7028):895-900.

[5] Palla G,Derenyi I,Farkas I,et al. Uncovering the over-lapping community structure of complex networks in nature and society[J]. Nature,2005,435(7043):814-818.

[6] 毛国军,段立娟,王 石,等. 数据挖掘原理与算法[M]. 北京:清华大学出版社,2005:156-183.

[7] 张云涛,龚 玲. 数据挖掘原理与技术[M]. 北京:电子工业出版社,2004:49-57.

[8] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E,2004,69(6):066133.

[9] Deneubourg J L,Goss S,Franks N,et al. The Dynamics of Collectivesorting:Robot-like Ants and Ant-like Robots[C]// Proc of the 1st Int'l Conf on Simulation of Adaptive Haviour. [s. l.]:[s. n.],1991:356-365.

[10] Radicchi F,Castellano C,Cecconi F. Defining and Identifying Communities in Networks[J]. Proceedings of the National Academy of Sciences,2004,101(9):2658-2662.

[11] Dorigo M,Stutzle T. Ant Colony Optimization[M]. 张 军,胡晓敏,罗旭耀,等译. 北京:清华大学出版社,2006.

[12] Zachary W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research,1977,33:452-473.

(上接第 128 页)

能够客观、合理地确定信息能力各级指标间关于评估等级的联系度,是一种可行有效的信息能力的评估方法。

参考文献:

[1] 邹振宁. 基于信息系统的体系作战指挥信息能力研究[M]. 北京:海潮出版社,2011.

[2] 裘杭萍,王玲玲,权冀川,等. 基于信息质量的信息优势评估指标研究[J]. 计算机技术与发展,2010,20(5):128-131.

[3] 王新敏,赵洪利. C4ISR 系统信息能力研究[J]. 装备指挥技术学院学报,2005,16(5):6-9.

[4] 曹 蕾. 指挥信息系统[M]. 北京:国防工业出版社,2012.

[5] Alberts D S,Garstka T J,Richard E H,et al. Understanding Information Age Warfare[M]. [s. l.]:Ccrp Publication Series,2001.

[6] Evidence Based Research, Inc. Network Centric Operations Conceptual Framework Version 1. 0[R]. [s. l.]:Office of Force Transformation,2003.

[7] Saaty T L. The analytic hierarchy process[M]. New York:McGraw,1980.

[8] 许树柏. 层次分析法原理[M]. 天津:天津大学出版社,1998.

[9] 叶义成. 系统综合评价技术及其应用[M]. 北京:冶金工业出版社,2006.

[10] 庞彦军,刘开第. 模糊数学中"取大取小"运算引发的问题[J]. 系统工程理论与实践,2001,21(9):98-100.

[11] 赵克勤. 集对分析及其初步应用[M]. 杭州:浙江科学技术出版社,2000:15-17.

[12] 王万军. 多元联系数集对模型及其评价应用[J]. 甘肃联合大学学报:自然科学版,2007,21(4):76-78.

基于改进的ACCA的复杂网络社团结构发现

作者： 臧丽， 王红， 杨通辉
作者单位： 山东师范大学 信息科学与工程学院,山东 济南 250014
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2012(10)