基于节点相似度的网络社团检测算法研究

姜雅文 贾彩燕 于 剑
(北京交通大学计算机与信息技术学院 北京 100044)

摘 要 社团结构是众多复杂网络的统计特性之一,挖掘网络中存在的社团结构日益受到人们的普遍关注。网络中的社团结构检测本质上类似于传统机器学习领域的聚类分析,其关键问题在于如何定义网络中节点间的相似度。首先提出了基于节点相似度的节点分裂算法SGN,相比传统的基于边界数(betweenness)的节点分裂算法GN,SGN在速度和精度上都有明显改善;接着,在利用各种节点相似度计算方法得到节点间的相似度之后,采用几种经典的聚类分析算法对网络进行社团划分,在模拟数据和真实数据上的实验表明:基于网络拓扑结构信息的signal和regular方法优于基于网络节点局部信息的Jaccard方法,而且对于复杂网络社团划分问题,如果选择好的网络节点相似度构造方法,已有的基于相似度矩阵的聚类分析算法都能快速有效地对网络社团进行划分。

关键词 复杂网络,社团结构,近邻传播,信号传递,节点相似度
中图法分类号 N94,TP393 **文献标识码** A

Community Detection in Complex Networks Based on Vertex Similarities

JIANG Ya-wen JIA Cai-yan YU Jian
(Institute of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract One of statistical characteristics in complex networks is a community structure. Detecting communities in networks has aroused great interest among researches in recent years. Actually, community detection is very similar to the classical cluster analysis in machine learning field. Thus, the key point is how to define vertex similarities in complex networks. We first proposed an algorithm named SGN based on vertex similarities. Compared with GN, SGN is much better and faster than GN. Secondly, we used four classical clustering algorithms to detect community structure in networks based on some existing vertex similarity measures. The results on artificial networks and real social networks show that the similarity measures based on signal propagation and regular equivalence theory by using the whole topology structure of networks are better than the methods of Jaccard based on local vertex information. Therefore, if vertex similarities are given well enough, proper clustering algorithms based on similarity matrices can be used to detect community structures fast and effectively in complex networks.

Keywords Complex network, Community structure, Affinity propagation, Signal propagation, Vertex similarity

1 引言

现实世界中的很多系统都可以抽象为网络,例如人际关系网[1]、疾病传播网、科学家合作关系网[2]、蛋白质相互作用网[3]等。这些网络都具有共同特点即复杂的内部结构,因此被称为复杂网络。通常,在复杂网络中每个节点表示一个个体,节点间的连边表示个体间的相互作用。对于疾病传播网来说,研究疾病如何进行传播以及网络结构本身具有的内在性质对于控制疾病的传播具有很重要的作用;对于蛋白质相互作用网络来说,研究网络中的簇结构可以预测未知蛋白质的功能。因此复杂网络研究具有很重要的现实意义。近几年来,已经涌现出了很多复杂网络的研究成果,例如小世界性[4]、无尺度性[5]、模块性[6]等。本文着重对网络的模块性进行研究,模块性是指网络自身都存在着簇结构,这种簇结构被称作"社团"(community),社团内部各节点之间连接紧密,而不同社团节点之间的连接稀疏。网络社团划分算法旨在挖掘网络中的社团结构。网络社团划分问题本质上是一个聚类问题,如果定义了网络中节点之间的相似度,那么就可以用任意一种聚类算法对网络中的节点进行聚类从而达到划分网络社团的目的。

目前,已经提出了一些划分网络社团的算法,大致分为基于谱思想的算法、层次聚类算法、基于优化思想的算法等。其中,基于谱思想的算法有谱平分法[7]和基于标准矩阵的谱方法[8];层次聚类算法最经典的是基于边介数(betweeness)节点分裂思想的GN算法[6]和基于凝聚思想的Newman贪婪算法[9];基于优化思想的算法有Kemighan-Lin算法[10]、极值优化算法[11]、基于模拟退火(simulated annealing)思想的GA算法[12]等。目前网络社团划分算法仍然存在很多问题,诸如算

到稿日期:2010-08-12 返修日期:2010-11-29 本文受国家自然科学基金(60905029,60875031)资助。
姜雅文(1985-),男,博士生,主要研究方向为机器学习,E-mail:09112077@bjtu.edu.cn;贾彩燕(1976-),女,博士,硕士生导师,主要研究方向为数据挖掘、生物信息学、复杂网络分析;于 剑(1969-),男,博士,教授,博士生导师,主要研究方向为机器学习。

法的效率问题(例如 GN 算法)、算法的参数选择问题(例如 Affinity Propagation[13]算法)以及节点相似度构造方法和选择问题等。

针对上述存在的问题,首先以提高传统 GN 算法效率为出发点,对各种网络节点相似度构造方法进行比较研究,提出了基于节点相似度的节点分裂式算法 SGN。实验表明,SGN 算法采用合适的相似度方法在效率上明显优于 GN 算法,但是由于其对相似度计算有较强的依赖性,SGN 算法相比其他经典的聚类算法不够稳定,算法的效率也不高。接着针对网络节点相似度选择问题进行研究,将各种相似度构造方法分别运用 AL(avarage linkage)[14]层次聚类算法、NJW[15]谱聚类算法和 AP(Affinity Propagation)算法。结果表明,基于网络全局拓扑结构信息的 signal[16]和 regular[17]方法优于基于节点邻居局部信息的 Jaccard 和 new-Jaccard[18]方法,而且能够得出结论:对于复杂网络社团划分问题,如果定义好网络节点的相似度,那么任何基于相似度矩阵的聚类算法都能快速有效地对网络中的节点进行聚类从而达到划分网络社团的目的。

本文第 2 节介绍几种目前网络节点的相似度构造方法;第 3 节首先介绍改进的基于节点相似度的节点分裂式 SGN 算法,然后分别介绍用于实验比对的 AL 算法、NJW 算法和 AP 算法;第 4 节在人工模拟数据集和真实数据集上对各种算法进行测试得到实验结果,然后进行分析得出结论;最后对本文进行总结和展望。

2 网络节点的相似度构造方法

本节介绍两类网络节点相似度构造方法,包括基于节点局部信息的 4 种方法和基于网络拓扑结构信息的两种方法。

2.1 基于节点局部信息构造网络节点相似度

该类方法考虑节点的邻居信息。一般来说,如果网络中的两个节点有着相同或者相近的邻居节点,那么这两个节点被认为是相似的,基于该思想的相似度构造方法大致有以下 3 种[18]:

假设 Γ_i 表示节点 i 的邻居集合,$|\Gamma_i|$ 表示该集合的势,$|\Gamma_i \cap \Gamma_j|$ 表示节点 i 和节点 j 共有的邻居个数。可以定义如下 3 种相似度:

$$S_{Jaccard}(i,j)=\frac{|\Gamma_i \cap \Gamma_j|}{|\Gamma_i \cup \Gamma_j|} \tag{1}$$

$$S_{cosine}(i,j)=\frac{|\Gamma_i \cap \Gamma_j|}{\sqrt{|\Gamma_i||\Gamma_j|}} \tag{2}$$

$$S_{\min}(i,j)=\frac{|\Gamma_i \cap \Gamma_j|}{\min(|\Gamma_i|,|\Gamma_j|)} \tag{3}$$

如果网络中两个节点之间不存在连边,则在 Jaccard 相似度构造的基础上减去一个惩罚项 Δ,由此可以构造一种新的相似度构造方法 new_Jaccard[18],即:

$$S_{new\_Jaccard}(i,j)=\begin{cases} S_{Jaccard}(i,j), & \text{if } (i,j)\in E \\ S_{Jaccard}(i,j)-\Delta, & \text{otherwise} \end{cases} \tag{4}$$

2.2 基于网络拓扑结构信息构造相似度矩阵

文献[16]基于信号传递思想将网络拓扑结构信息转化为空间向量信息,它的基本思想是将网络中的节点当作具有接收和发射信号的节点。首先从网络中任选一个节点 v,给 v 赋一个信号值,然后 v 向自己和自己的邻居节点发射该信号值,接到信号的节点记录并保存相应的信号值,同理,其它的节点也进行同样的接收和发送信号的过程。如此传递下去,经过 T 次传递之后,位于同一个社团里的节点对网络中其它节点传递的信号量是接近的。信号传递的过程可以用数学公式表示,即 $V=(I+A)^T$,I 表示单位矩阵,A 表示网络的邻接矩阵,T 表示信号传递的次数。文献[16]描述了在 5 个节点的网络上传递信号的过程,如图 1 所示。

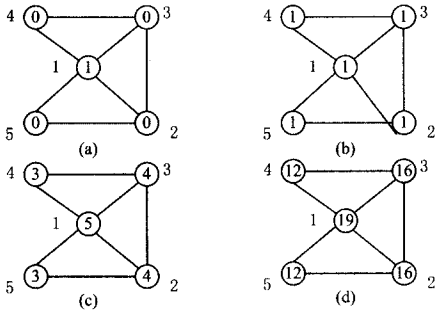

图 1 5 个节点构成的网络传递信号过程图[16]

从图 1 可以看出,信号传递 T 次后,包含 n 个节点的网络中的每一个节点的信号量是一个 n 维向量,它表示的是该节点对网络其它节点的影响程度。这样 n 个节点就有 n 个 n 维向量,从而将网络空间拓扑结构的信息转化为向量空间信息,在对其进行标准化之后就可以运用各种空间距离计算相应的相似度矩阵对其进行聚类。

如果一个节点相似于另一个节点的邻居节点,那么也认为这两个节点是相似的。基于这个思想,Leicht E. A. 等人提出了另一种基于网络拓扑结构的节点相似度构造方法,本文简称为 regular 方法[17]:

该方法采用一种迭代的思想,假设 A 为网络的邻接矩阵,$T(i)$表示中间矩阵,则

$$T(i+1)=\frac{\alpha}{\lambda_1} * A * T(i)+I_n \tag{5}$$

$$S_{regular}=D^{-1} * T(t) * D^{-1} \tag{6}$$

式中,I_n 表示 $n*n$ 的单位矩阵,a 是一个收敛参数,λ_1 是邻接矩阵的最大特征值,D 是对角矩阵,对角线的元素依次是每个节点的度,t 表示最大的迭代次数。

3 基于节点相似度的网络社团划分算法

定义了网络节点的相似度以后,网络社团划分问题就可以转化为聚类问题,可以用基于相似矩阵的聚类算法进行划分。本文分别采用层次聚类算法、谱聚类 NJW 算法和近邻传播算法(简称为 AP)对网络节点进行聚类,同时尝试用节点相似度取代边介数,对 GN 进行改进,提出另一种新的基于节点分裂思想的 SGN 算法。

3.1 基于节点相似度的 SGN 算法

GN 算法[6]是传统的基于边介数(betweenness)思想的一种层次分裂算法。该算法要计算每一条边的介数,每一条边的介数值表示所有的节点对之间通过该条边的最短路径的条数。由于计算每条边的边介数相当耗时,因此 GN 算法最大的缺点就是时间复杂度高,它的基本思想如下:

1. 计算网络中每一条边的 betweenness;
2. 选择具有最大 betweenness 值的边并删除;
3. 对剩下的边重新计算各自的 betweenness;
4. 重复步骤 2 的操作直到所有的边都去除为止。

针对 GN 算法复杂度高的不足,通过计算顶点间的相似

度代替 betweenness，从而达到改善算法时间复杂度的目的。如果网络中两个节点之间存在连边，那么用这两个节点间的相似度替代 betweenness 作为该条边的指标再对网络进行分裂得到 SGN 算法。SGN 算法的基本思想如下：

1. 计算网络每一条边两个端点之间的相似度；
2. 选择相似度最小的边删除；
3. 重新计算剩余边端点之间的相似度；
4. 重复步骤 2 的操作直到所有的边都去除为止。

3.2 其它基于相似矩阵的聚类算法

本节介绍其它基于相似矩阵的聚类算法，包括 Average Linkage(AL)层次聚类算法、NJW 算法和近邻传播算法(AP)。

3.2.1 Average Linkage 算法

AL(Average Linkage)算法[14]是一种基于凝聚思想的层次聚类算法，其基本步骤是：

1. 将每个数据对象看作一类，一共 N 类，每类仅包含一个对象。类间距离就是对象之间的距离；
2. 找到最接近的两个类并合并成一个新类；
3. 重新计算新类与所有旧类之间的距离；
4. 重复第 2 步和第 3 步，直到最后合并成一个类为止。

由于步骤 3 类间距离计算方式的不同，层次聚类算法又可分为 single-linkage、complete-linkage 以及 average-linkage，single-linkage 类间距离为两个类中对象之间的最小距离，complete-linkage 类间距离为两个类中对象之间的最大距离，而 average-linkage 类间距离为两个类中对象之间的平均距离，本文选用 average-linkage。

3.2.2 NJW 算法

NJW 算法[15]是一种规范并且常用的谱分割算法，该算法输入一个相似度矩阵 S，将相似度矩阵 S 的每行元素求和得到每个节点的度，以所有度值为对角线元素构成的对角矩阵称为对角矩阵 D，则该谱分割算法就将图划分问题转化为求解下述方程的最大特征值问题：$Nx=\lambda x$，其中，

$$N=D^{-\frac{1}{2}}(D-S)D^{-\frac{1}{2}} \tag{7}$$

算法首先将 N 矩阵的前 k 个最大特征值对应的特征向量合并成矩阵 V，将 V 按行矢量归一化得到 X，然后使用 k 均值方法对 X 的行矢量进行聚类得到聚类结果。

3.2.3 近邻传播算法(AP)

AP 算法[13]的核心思想是通过 3 个迭代公式不断地对两个消息 R 和 A 进行更新，因此 AP 算法又被称为基于消息传递的聚类算法。AP 算法将数据集中的所有样本点都视为候选的聚类中心，然后为每个样本点建立与其他样本点的吸引信息 $R(i,k)$和归属信息 $A(i,k)$。$R(i,k)$称为点 k 对点 i 的吸引度(responsibility)，它用来描述数据点 k 适合作为数据点 i 的类代表的程度；$A(i,k)$称为点 i 对点 k 的归属度(availability)，它用来描述数据点 i 选择数据点 k 作为其类代表的适合程度。二者之和越大，点 k 作为最终聚类中心的可能性就越大。

AP 算法步骤如下：

初始化：

$l=0, r^{(0)}(i,k)=0, a^{(0)}(i,k)=0, S(i,k)_{i\neq k}=s$

$S(k,k)=P, \lambda\in(0,1), iter$(默认迭代次数 1000)

Step 1 计算 responsibility

$$r^{(l+1)}(i,k)=\lambda r^{(l)}(i,k)+(1-\lambda)(S(i,k)-\max_{j\neq k}\{a^{(l)}(i,j)+S(i,j)\}) \tag{8}$$

Step 2 计算 availability

$$a^{(l+1)}(i,k)=\lambda a^{(l)}(i,k)+(1-\lambda)\min\{0, r^{(l+1)}(k,k)+\sum_{j\notin\{i,k\}}\max\{0, r^{(l+1)}(j,k)\}\}, i\neq k \tag{9}$$

$$a^{(l+1)}(k,k)=\lambda a^{(l)}(k,k)+(1-\lambda)\sum_{j\neq k}\max\{0, r^{(l+1)}(j,k)\} \tag{10}$$

Step 3 得到聚类结果

if $l<iter, l=l+1$，转步骤 1，否则：

$$c_i=\arg\max_k\{r^{(iter)}(i,k)+a^{(iter)}(i,k)\} \tag{11}$$

AP 算法有 3 个输入参数：相似度矩阵 $S(i,j)$，其对角线元素参数 P 以及震荡因子 λ。P 被称作偏向参数，$P<0$ 并且 P 值的大小决定最终的聚类数，P 值越大得到的聚类数越大。震荡因子是为了防止算法不收敛，一般取值在 0 和 1 之间，通常取值 0.9。

4 实验结果与分析

上节描述的网络社团划分算法均需要节点相似度作为输入参数，又由于 Jaccard，cosine 和 min 方法没有太大差别，因此我们分别采用 Jaccard，new-Jaccard，signal 和 regular 相似度构造方法与这 3 种算法自由组合，在人工数据集和真实数据集上进行测试对比，验证不同相似度构造方法和不同划分算法的有效性。

4.1 人工数据集

人工数据集[6]利用计算机程序生成，规则如下：每一个网络都由 128 个节点构成，分为 4 个团，每一个团 32 个节点；社团内部节点之间按照概率 P_{in} 随机添加，社团之间、节点之间按照概率 P_{out} 随机添加；两个概率的取值保证 $Z_{in}+Z_{out}=16$，Z_{in} 表示一个节点与社团内部节点连边的平均值，Z_{out} 表示一个节点与社团外部节点连边的平均值，因此 Z_{in} 越大，社团结构越明显，Z_{out} 越大，社团结构越模糊。程序生成已知的社团结构，实验以最终划分结果相比真实结果的准确率(accuracy)作为评价指标。

首先，将改进的 SGN 算法采用多种不同的相似度与 GN 算法在人工数据集上进行比对，每一次实验(例如 Z_{out} 等于 0)的结果均是计算 10 个该类人工网络求得的平均值，实验结果如图 2 所示。从图 2 可以看出，Jaccard 和 new-Jaccard 方法的结果相同，由于其不考虑网络的拓扑结构信息，结果不及 GN 算法；regular 方法考虑了网络的拓扑结构信息，在网络社团结构不明显的情况下结果好于 GN 算法，但是在其它情况下略差于 GN 算法；而基于信号传递思想的 SGN 算法在精度上明显优于 GN 算法。总的来说，SGN 分裂式算法对相似度计算有较强的依赖性，相似度选取的好坏直接影响到算法的精度。

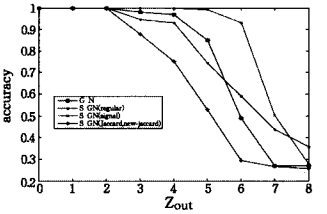

图 2 SGN 与 GN 在人工数据集上的结果比对

假设网络有 n 个节点，m 条边，算法迭代次数为 T，聚类数为 k。为了验证 SGN 算法的时间复杂度，又人工生成了两

个规模较大的网络分别运行比对几个经典算法并记录运行时间，这几个算法包括前面介绍的 GN 算法（$O(nm^2)$）以及基于相似度矩阵的 Avarage Linkage 算法（$O(n^2)$）、NJW 算法（$O(n^3)$）和 AP 算法（$O(n^2)$）；此外，FN（Newman 贪婪）[9] 算法（$O(mn)$）是对 GN 算法的改进，是经典的网络社团划分算法；MCL（马尔科夫聚类）算法（$O(n^3)$）[19] 尽管受到算法复杂度的影响，但是文献表明，该算法精度高，同样也是处理图划分问题的经典算法；k-means 算法（$O(nkT)$）是聚类分析中最为经典、高效的算法，由于可以利用 signal 方法将网络节点转化为空间向量中的点，因此可以利用 k-means 经典的算法对其进行社团划分。将以上 3 种经典算法也加入到比对实验中，其中 SGN，AL，NJW 和 AP 算法都选取基于信号传递思想的 signal 相似度构造方法。网络 1 是 512 个节点，16384 条边，4 个社团，节点平均度为 32，Z_{out} 和 Z_{in} 为 2－14；网络 2 是 1024 个节点，32768 条边，8 个社团，节点平均度为 32，Z_{out} 和 Z_{in} 为：2－14。实验结果如表 1 和表 2 所列。

表 1　SGN 与其它算法时间复杂度对比（网络 1）

算法	准确率	时间(s)
GN	1	2253.42
SGN	1	185.204
AL	1	97.187
NJW	1	2.515
AP	1	5.593
FN	1	7.422
MCL	1	6.594
k-means	1	0.609

表 2　SGN 与其它算法时间复杂度对比（网络 2）

算法	准确率	时间(s)
GN	1	22637.5
SGN	1	2974.53
AL	1	1725.11
NJW	1	23.406
AP	1	30.609
FN	0.999	35.422
MCL	1	66.406
k-means	1	7.11

由于 k-means 算法对初始值设置较为敏感，实验记录的是运行一次结果最优的 k-means 算法执行的时间。从上面的结果可以看出，所有算法在精度一致的情况下，GN 算法的时间复杂度最高，SGN 算法的运行时间相比 GN 算法有明显缩短，但是仍远高于 AP 算法、FN 算法、k-means 算法和 NJW 算法，谱聚类算法和 k-means 算法效率最高。各种算法的时间复杂度比较结果为：

k-means＜NJW＜AP＜FN＜MCL＜AL＜SGN＜GN

为了进一步验证基于信号传递的 signal 相似度方法优于其它 3 种方法，仍然采用上述 4 种不同相似度构造方法在人工数据集上运行 AL 算法，NJW 算法以及 AP 算法，实验结果如图 3、图 4 和图 5 所示（每一个实验结果均是由 10 个此类人工网络求得的平均值）。

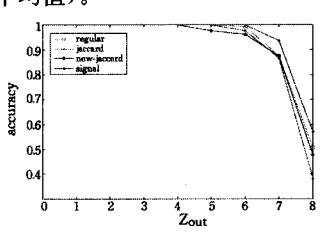

图 3　不同相似度方法用于 Avarage Linkage

由上述的结果可以看出，在所有的相似度构造方法中，Jaccard 和 new-Jaccard 在各种算法中的表现差别不大。但是由于这两种相似度构造方法仅仅考虑节点与节点之间的邻居信息，而并没有考虑整个网络的拓扑结构信息，因此效果不如 regular 和 signal 方法；由于考虑了网络的拓扑结构信息，regular 和 signal 相似度构造方法在 3 种聚类算法中均得到很好的划分效果，尤其基于信号传递思想的 signal 方法，在 3 种算法中均优于其它 3 种相似度构造方法。此外，相比 SGN 算法，谱聚类、层次聚类算法和 AP 算法对相似度依赖程度很低，而且效果较好。

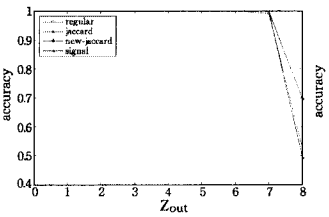

图 4　不同相似度方法用于 NJW 算法

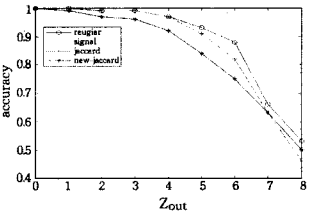

图 5　不同相似度方法用于 AP 算法

下面我们对 4 种不同相似度构造方法在这 3 种不同聚类算法下的速度进行对比，数据仍然采用上述规模较大的网络 1 和网络 2，表 3 和表 4 分别记录了各种算法运行的时间。

表 3　不同相似度方法的时间复杂度对比（网络 1）

网络 1(s)	AP	AL	NJW
Jaccard	7.078	101.031	4.188
new-Jaccard	7.359	103.703	4.281
signal	5.015	101.594	2.844
regular	4.656	98.484	1.516

表 4　不同相似度方法的时间复杂度对比（网络 2）

网络 2(s)	AP	AL	NJW
Jaccard	91.593	2037.39	24.047
new-Jaccard	33.734	1846.47	24.297
signal	30.281	1788.81	24
regular	**24.454**	2011.36	**9.703**

从表中的结果可以看出，3 种聚类算法中依然是谱聚类算法的时间复杂度最低，AP 算法略高于谱聚类算法，层次聚类算法最高；4 种相似度方法中在多数的情况下是 regular 方法最快，signal 方法略差，但两者都优于基于节点局部信息的 Jaccard 方法和 new-Jaccard 方法。综合两个方面的考虑，可以得出结论，采用 regular 相似度构造方法的谱聚类算法运行效率最高。

为了比对不同算法对于网络社团划分的效果，选取准确率最高的基于信号传递的 signal 相似度构造方法，将上述聚类算法仍旧与经典的社团划分算法 GN，FN，MCL 以及经典的 k-means 算法进行比对，如图 6 所示，其中 k-means 选取执行 100 次算法结果中最优的结果。

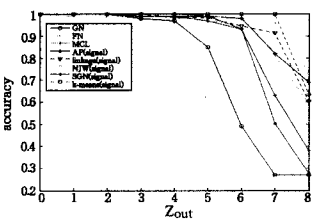

图 6　不同网络社团划分算法比对

从图中可以看出，各种算法在准确率上大致呈现如下结果：k-means＞NJW＞AL＞AP＞FN＞MCL＞SGN＞GN，在所有网络社团划分算法中，基于信号传递相似度方法的 k-

means 算法、层次聚类算法、谱聚类算法和 AP 算法最优，其次是 Newman 贪婪算法；MCL 算法和 SGN 算法效果相当；GN 算法效果最差。因此，可以得出结论，对于复杂网络社团划分问题，如果选择一个好的网络节点相似度构造方法，那么一般的基于相似度矩阵的聚类算法，例如 k-means 算法、层次聚类、谱聚类和近邻传播聚类算法都能快速有效地得到聚类结果。

4.2 真实数据集

真实数据采用了经典的 Zachary 空手道俱乐部网络 katate[20]、美国大学足球赛网络 football[6]、海豚关系网 dolphins[21]和美国政论著作网络 books on politics（由 Valdis Krebs 收集，未公开发表，参见 http://www.orgnet.com/）。其中空手道俱乐部网络由 34 个节点组成，每一个节点表示一个成员，边表示成员之间的社会交往关系，由于俱乐部主管和校长之间发生矛盾，俱乐部成员就被分成了分别以主管和校长为中心的两个社团；美国大学足球赛网路由 115 个节点组成，每一个节点代表一支队伍，边代表端点的两支队伍进行过常规赛，全部队伍一共被分为 12 个联盟，与联盟内球队进行比赛的概率大于与联盟外球队比赛的概率；海豚关系网是由 2 个群体共 62 个海豚构成，节点表示海豚的个体，连边表示两个海豚之间接触频繁；政论著作网络由 105 个节点构成，每一个节点表示在线售书商亚马逊网站出售的每一本关于美国政论的著作，节点间的连边表明顾客同时购买了该两本著作，根据政见的不同，所有节点被分成了“自由派”、“中立派”和“保守派”三类著作。4 个真实网络都有清晰的社团结构，因此仍然以社团划分的准确率作为评价指标。

仍然采用前述 4 种相似度构造方法分别用于 SGN 算法、AL 算法、NJW 算法以及 AP 算法，并与 GN 算法、FN 算法、MCL 算法、k-means 算法进行比较，其中 k-means 算法仍然运行多次取最好的结果。实验发现，除了 AL 算法，其它算法中仍然是基于信号传递思想的 signal 方法效果最优，如表 5 所列。其中采用信号传递相似度构造方法的 AP 算法和谱聚类算法的准确率最高，FN 最差，其余算法效果相当。表 5 括号里列出的是结果最好的相似度构造方法。

表 5 不同社团划分算法用于真实数据

	katate	football	dolphins	polbooks
SGN(signal)	1	0.8522	1	0.8381
AL(Jaccard)	1	0.9130	0.9677	0.8476
NJW(signal)	1	0.9217	1	0.8667
AP(signal)	1	0.9652	1	0.8476
GN	0.9706	0.9043	1	0.8381
FN	0.7353	0.5739	0.6935	0.8095
MCL	1	0.9043	1	0.8381
k-means	1	0.9043	1	0.8476

结束语 首先基于目前提出的几种相似度构造方法，提出一种改进的节点分裂式算法 SGN。如果采用基于信号传递思想的 signal 相似度构造方法，SGN 算法相比传统的 GN 算法速度和精度有显著改善，但是由于其对相似度计算有较强的依赖性，所以相比其他经典聚类算法例如 k-means 算法、层次聚类、AP 和谱聚类算法不够稳定；其次通过采用多种不同的算法对各种相似度构造方法进行比较研究发现，基于 Jaccard 和 new-Jaccard 的方法效果相当，两者都没有考虑网络的拓扑结构，效果不好；而基于信号传递和 regular 方法的相似度构造方法考虑了网络的拓扑结构，因此结果最优，算法也最稳定。其中基于信号传递思想的相似度构造方法，由于它将网络中的节点转化为向量空间中的点，这样，基于该方法的经典聚类算法例如 k-means 算法、层次聚类、谱聚类和近邻传播聚类算法都能对网络社团进行有效划分，因此，signal 方法在准确率上更优于 regular 方法；而通过时间复杂度的对比实验发现，regular 方法尤其在谱聚类算法下的运行速度要快于 signal 方法，因此，从时间复杂度上看，regular 方法要优于 signal 方法。总之，对于复杂网络社团划分问题，如果选择一个好的网络节点相似度构造方法，那么任何基于相似度矩阵的聚类算法都能快速有效地对网络社团进行划分。

参 考 文 献

[1] Amaral L A N, Scala A, Barthelemy M, et al. Classes of small-world networks [J]. Proc. Natl Acad Sci USA, 2000, 97(21): 11149-11152

[2] Redner S. How popular is your paper? An empirical study of the citation distribution [J]. Eur Phys J B, 1998, 4: 131-134

[3] Drewes G, Bouwmeester T. Global approaches to protein-protein interactions[J]. Curr Opin Cell Biol, 2003, 15(2): 199-205

[4] Watts D J, Strogatz S H. Collective Dynamics of Small-World Networks[J]. Nature, 1998, 393(6648): 440-442

[5] Barabasi A L, Albert R. Emergence of Scaling in Random Networks[J]. Science, 1999, 286(5439): 509-512

[6] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proc. Natl Acad Sci USA, 2002, 99: 7821-7826

[7] Pothen A, Simon H, Liou K-P. Partitioning sparse matrices with eigenvectors of graphs[J]. SIAM J Matrix AnalApp l, 1990, 11(3): 430-452

[8] Capocci A, Servedio V D P, Caldarelli G, et al. Detecting communities in large networks [J]. Physica A, 2005, 352: 669-676

[9] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Phys Rev E, 2004, 69(6): 066133

[10] Kernighan B W, Lin S. A efficient heuristic procedure for partitioning graphs [J]. Bell System Technical Journal, 1970, 49(2): 291-307

[11] Duch J, Arenas A. Community detection in complex networks using extremal optimization[J]. Phys Rev E, 2005, 72(2): 027104

[12] Guimera R, Amaral I LAN. Functional canography of complex metabolic networks[J]. Nature, 2005, 433(7028): 895-900

[13] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315: 972-976

[14] Scott J. Social Network Analysis: A Handbook (2nd edition) [M]. Sage, London, 2000

[15] Ng A, Jordan M, Weiss Y. On spectral clustering: Analysis and an algorithm[M]. Advances in Neural Information Processing Systems 14, Cambridge, MA, MIT Press, 2002

[16] Hu Yanqing, Li Menghui, Zhang Peng, et al. Community detection by signaling on complex networks[J]. Phys Rev E, 2008, 78

[17] Leicht E A, Petter H, Newman M E J. Vertex similarity in networks[J]. Physical Review E, 2006, 73(2): 026120

[18] Liu Zhi-yuan, Li Peng, Zheng Ya-bin, et al. Community detection by affinity propagation[R]. 2008

[19] Dongen S V. Graph clustering by flow simulation[D]. Utrecht: Centers for mathematics and computer science (CWI), University of Utrecht, 2000

[20] Zachary W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 33(4): 452-473

[21] Lusseau D, Boisseau S K, et al. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations[J]. Behavioral Ecology and Sociobiology, 2003, 54: 396-405