文章编号：1006-4710(2013)04-0455-05

基于局部相似性的 K-means 谱聚类算法

王林[1]，高红艳[1,2]，王佰超[1]

(1. 西安理工大学 自动化与信息工程学院，陕西 西安 710048；
2. 宝鸡文理学院 物理与信息技术系，陕西 宝鸡 721016)

摘要：定义科学的局部相似性指数是基于局部相似性社团发现算法的关键，根据共有邻居信息定义的局部形似性指数对直接相连接点对的相似性数值存在低估倾向，本研究将节点对的关联信息加入到 *sφrensen* 局部相似性指数的定义中，结合 K-means 谱聚类算法对网络节点进行聚类。本研究定义的局部相似性指数克服了传统局部相似性指数的缺点，且保持了原有的计算复杂性。在计算机生成网络和实际网络上运行，并和经典算法做了比较，实验证明，所提算法能够较为有效、准确地检测网络的社团结构。

关键词：局部相似性；谱聚类；K-means 聚类

中图分类号：TP15　　**文献标志码**：A

Algorithm of K-means Spectral Clustering Based on Local Similarity

WANG Lin[1], GAO Hongyan[1,2], WANG Baichao[1]

(1. Faculty of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China;
2. Faculty of Physics and Information Engineering, Baoji University of Art and Science, Baoji 721016, China)

Abstract: The correlation information of node pairs isincoporated in the definition of *sφrensen* local similarity index, network nodes are clustered by this similarity measure combining with Kmeans spectral clustering. The similarity index proposed by the paper overcomes the shortcomings of traditional local similarity index, and maintains the original computational complexity . The proposed method is tested on both computer-generated and real-world networks, and is compared with the typical algorithms in community detection. Experimental results verify and confirm the feasibility and validity of the proposed method to monitor community structure of real internet accurately.

Key words: local similarity; spectral clustering; K-means clustering

社团结构是复杂网络的重要特性之一，研究社团结构有助于了解网络结构，分析网络特性，预测网络功能[1-3]，所以社团检测近年来成为学者们研究的热点问题[4]。

基于相似性的算法是一类重要的社团检测算法，它利用网络的全局或者局部特性计算节点或边的相似性，再结合传统聚类算法社团划分[5-6]。该算法的关键是定义合适的“相似性指数”。一类相似性指数是全局的，考虑了网络的全局信息，计算准确，但运算复杂度高，而且引进了额外参数，实用性不强。另一类相似性指数为局部指数[7-8]，采用节点的共有邻居数或简单变形来衡量节点相似性，无视节点是否直接相连，从而使直接相连的但无共同邻居的节点对具有零相似性，这显然是不合适的。因此，寻找一个计算简单，且能准确刻画网络结构的相似性指数仍然是人们迫切需要解决的问题。

基于上述问题，本研究提出了一种新的局部相似性指数，该指数以 *sφrensen* 指数[9]为原形，考虑了节点直接相连接对相似性的影响，采用 K-means 谱聚类法对网络节点进行聚类，与基于传统的局部相似性指数的社团发现算法和几种典型的社团发现算法做了比较，在计算机生成的网络和实际网络上验证，结果显示，所提的算法能够克服传统局部相似性的缺点，继承传统局部相似性计算量小的特点，能够准确，有效地检测人工合成网络和实际网络的社团结构。

收稿日期：2013-04-20

作者简介：王林，男，教授，研究方向为复杂网络、数据库等。E-mail：wanglin@xaut.edu.cn。

1 基于局部相似性的 K-means 谱聚类算法

1.1 相似性度量

求网络节点的相似性一般有两种方法,如果网络可以嵌入到 n 维欧几里得空间,每个节点对应欧几里得空间的一个点,就可以用每对节点之间的距离来表示节点之间的相似性。如果不能嵌入到欧几里得空间,则节点之间的相似性只能通过节点之间的邻接关系来推断,*sϕrensen* 指数就是一种通过节点的共有邻居数来描述节点之间相似性的局部相似性指数,定义为:

$$s_{ij} = \frac{2|N(i) \cap N(j)|}{k_i + k_j} \tag{1}$$

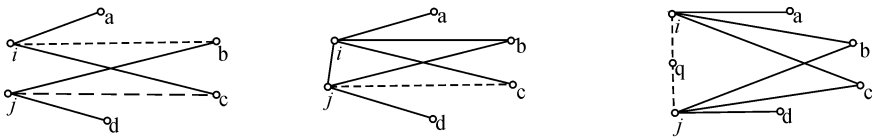
式中,$N(i)$ 表示节点 i 的邻居。分子表示节点 i 与节点 j 的共有邻居数,k_i,k_j 为节点 i,j 的度。

对于不直接相连的节点 *sϕrensen* 指数能够较为准确地评估它们的相似性。对于直接相连却无共同邻居的节点对,由于分子为零,所以 *sϕrensen* 赋予其0相似性。而在实际当中,同一社团内直接相连而无共同邻居的节点对是普遍存在的,这些节点对的相似性显然是不应该为0的。因此在定义节点相似性时,应考虑两种情况,即:

1)两节点之间无边相连,如图1(a)所示的节点 i,j;

2)两节点之间有边相连(这里考虑简单图,两节点之间只有一条边),如图1(b)所示节点 i,j。

为了能够准确地刻画所有节点对之间的相似性,在节点 i,j 之间添加一个辅助接点 q,如图1(c)所示,则节点 i 和 j 之间的边 (i,j) 可认为是由辅助节点 q 连接的边 (i,q) 和边 (j,q) 组成。这样,节点 i,j 之间的边断开,反映在邻接矩阵 $\boldsymbol{A}$ 中,就是 i 行 j 列和 j 行 i 列元素由1变为0,并且需要加入辅助节点所对应的一行和一列。假设图1(c)所对应的邻接矩阵为 $\boldsymbol{A}'$,那么新加的行和列仅有 $\boldsymbol{A}'(i,q)$,$\boldsymbol{A}'(j,q)$,$\boldsymbol{A}'(q,i)$,$\boldsymbol{A}'(q,j)$ 四个元素为1,其余为0。



(a)节点*i*与*j*之间无边　(b)节点*i*与*j*之间有边　(c)加辅助节点

图1　一个简单的例子

Fig. 1　a simple example

若网络的邻接矩阵为 $\boldsymbol{A}$,则(1)式等价为:

$$s_{ij} = \frac{2(\boldsymbol{A}^2)_{ij}}{k_i + k_j} \tag{2}$$

或为:

$$s_{ij} = \frac{2 \times \sum_{k=1}^{n} \boldsymbol{A}_{ik} \times \boldsymbol{A}_{kj}}{k_i + k_j} \tag{3}$$

添加辅助节点以后,将新的邻接矩阵带入(3)式得到:

$$s_{ij} = \frac{2 \times \sum_{k=1}^{n+1} \boldsymbol{A}'_{ik} \times \boldsymbol{A}'_{kj}}{k_i + k_j} = \frac{2 \times (1 + \sum_{k=1}^{n} (\boldsymbol{A}(i,k) \times \boldsymbol{A}(j,k)))}{k_i + k_j} \tag{4}$$

如此,第2)类情况可用表达式(4)进行计算。所以任意节点对的修正 *sϕrensen* 相似性指数可表示为:

$$s_{ij} = \frac{2 \times (\boldsymbol{A}(i,j) + \sum_{k=1}^{n} (\boldsymbol{A}(i,k) \times \boldsymbol{A}(j,k)))}{k_i + k_j} \tag{5}$$

由于在社团发现过程中主要考虑节点对之间相似性值的排序,故可以省去乘数因子,得到相似性指数为:

$$s_{ij} = \frac{\boldsymbol{A}(i,j) + \sum_{k=1}^{n} (\boldsymbol{A}(i,k) \times \boldsymbol{A}(j,k))}{k_i + k_j} \tag{6}$$

上式可以简写为:

$$s_{ij} = \frac{(\boldsymbol{A} + \boldsymbol{A}^2)_{ij}}{k_i + k_j} \tag{7}$$

修正 *sϕrensen* 考虑了节点直接相连对相似性的贡献,在计算复杂性上,修正 *sϕrensen* 指数与 *sϕrensen* 指数具有相同的时间复杂性,为 $o(n^2)$,所以它更适合检测网络的社团结构。

1.2 K-means 谱聚类算法

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。这里采用 K-means 谱聚类算法对网络节点进行聚类[10],算法执行过程为:

1)根据上节定义的相似性指数计算相似性矩阵 $\boldsymbol{S}$;

2)求相似性矩阵的特征值和相应的特征向量,将特征值按顺序排列,根据相邻特征值之间的最大距离估计网络的社团个数 C;

3)取特征值前 C 个最大值所对应的特征向量组成新的向量 $\boldsymbol{V}$;

4)以 $\boldsymbol{V}$ 为原始数据对其进行 K-means 聚类;

5)用模块度验证估算的社团个数 C 的合理性,计算社团划分的准确性。

该算法的时间复杂度由三部分组成,第一部分是计算相似性矩阵的复杂度为 $o(n^2)$,第二部分是谱聚类算法,计算特征向量时采用 Lanczos 方法快速计算前 k 个最大特征值对应的特征向量,该方法的复杂度约为 $o(km)$,其中 k 为社团个数,m 为网络节点数。第三部分是 K-means 聚类算法的复杂度 $o(ktn)$,t 为迭代次数,n 为网络节点数。因为 k 为常数,迭代次数 t 有限制总时间复杂度为 $o(n^2+m)$。

2 仿真

为了测试本研究所提算法的性能,笔者利用计算机合成的网络和真实世界网络进行了社团发现实验。

2.1 计算机生成网络

所用的基准网络[11]由 128 个节点组成,划分为 4 个社团,网络中节点的平均度为 16,其中社团之间的节点度期望值为 Z_{out},社团内节点度期望值为 Z_{in},让 Z_{out}从 1 到 8 连续增加,观察社团划分情况。

社团划分的准确性用归一化互信息(Normalized mutual information)来度量[12],定义为:

$$NMI(X,Y) = \frac{-2\sum_{i=1}^{C_X}\sum_{j=1}^{C_Y}N_{ij}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{C_X}N_i\log\left(\frac{N_{i.}}{N}\right)+\sum_{j=1}^{C_Y}N_j\log\left(\frac{N_{.j}}{N}\right)} \tag{8}$$

式中,X 为原始的社区划分,Y 为算法得出的社团划分,$NMI(X,Y)$表示划分 X 与划分 Y 之间的接近程度,取值范围为 0 ~ 1,其值越大算法所得的社团结构与真实社团越接近,当 $NMI=1$ 时表示所得结果与真实社团划分完全一致。当 $NMI=0$ 时表示算法得到的划分与真实划分完全不同。

社团划分的 NMI 曲线如图 1 所示,由于生成的基准网络具有随机性,对于每个 Z_{out}值笔者重复做了 100 次实验。图 2 中每个点对应 100 次实验的平均值。由图 2 可知,当 Z_{out}从零逐渐增加时,由于基准网络的社团结构由清晰变模糊,所以 NMI 曲线呈递减趋势;当 $Z_{\text{out}}=1\sim4$ 时该算法的正确划分率为 100%,$Z_{\text{out}}<6$ 时准确率在 0.9 以上。当 Z_{out}继续增加时,由于网络社团结构越来越模糊,社团检测准确率剧烈下降。可见当社团结构明显时,该算法的准确率是比较高的。

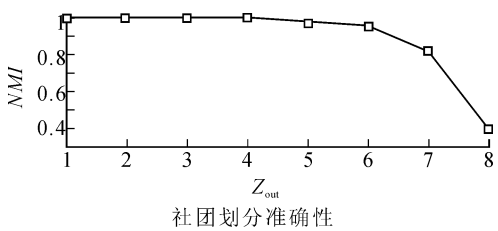


图 2 算法准确度

Fig. 2 The accuracy of the algorithm

为进一步验证算法性能本研究做了两组实验,首先将基于改进相似性指数的算法与基于原相似性指数的算法做了比较,结果如图 3 所示。由图 3 可知 $Z_{\text{out}}<4$ 时两种算法的准确率都比较高,但当 Z_{out}继续增加的时候,改进算法的准确率明显高于原算法。

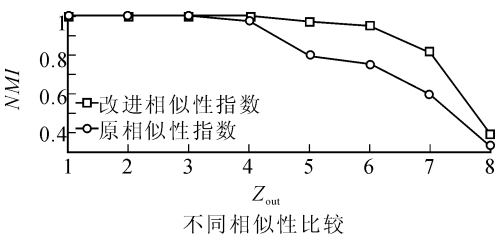


图 3 不同相似性比较

Fig. 3 The comparison of different similarity

其次,将本算法与 GN 算法(GN)和快速算法(fastgreedy)做了比较,结果如图 4 所示。

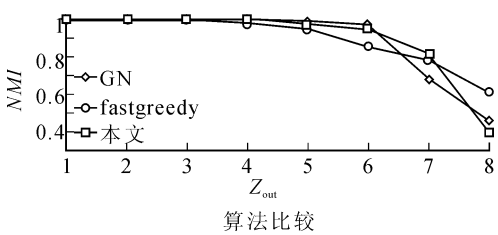


图 4 不同算法比较

Fig. 4 The comparison of different algorithms

由图 4 可见,当社团结构比较明显时($Z_{\text{out}}<7$),本研究算法性能高于 GN 算法和快速算法,当社团结构模糊时($Z_{\text{out}}>7$),由于本算法采用了节点的

局部信息,算法准确率低于 GN 算法和快速算法。从算法的复杂度来看,本算法时间复杂度大于 GN 算法,小于快速算法。

2.2 真实网络

研究社团发现算法的目的在于检测社团个数未知的真实网络的社团结构,所以还需要用真实网络检测算法的准确性和有效性,下面笔者将用常用实际基准网络 Karate 网和美国大学足球俱乐部网检测算法。

Zachary 网络是检测社团发现算法的基准网络之一[13],反映了美国一所大学中空手道俱乐部 34 名成员之间的相互社会关系。该俱乐部在被研究期间,由于内部争执而分裂成了 16 名成员和 18 名成员的小俱乐部。该俱乐部的自然划分如图 3 所示,图 3 中方形和圆形分别代表分裂后的小俱乐部中的各个成员。图 5 为本算法得到的相似矩阵特征序列图,图 6 为相应的模块度曲线。表 1 社团划分结果。

表 1 karate 网络社团划分结果

Tab. 1 The result of community detection of karate network

社团	所含节点
1	1,2,3,4,5,6,7,8,11,12,13,14,17,18,20,22
2	9,10,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34

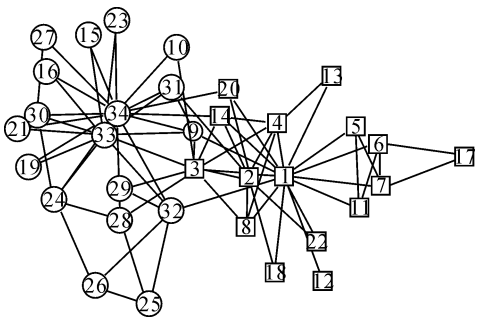


图 5 Zachary 空手道俱乐部成员划分结果

Fig. 5 The result of community detection of Zachary karate club

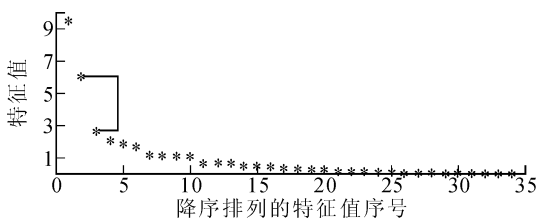


图 6 相似矩阵特征值序列图

Fig. 6 Sequence diagrams of similarity matrix

从图 6 可知,最大特征向量距离为 $\lambda_2 - \lambda_3$,所以预测社团个数为 2,由图 7 可知,当社团个数 $C=2$ 时,模块度最大,为 0.3715,所以预测时正确的。由表 1 可知网络划分为两个社团,第一个社团 16 个节点,第二个社团 18 个节点,划分准确率为 100%。

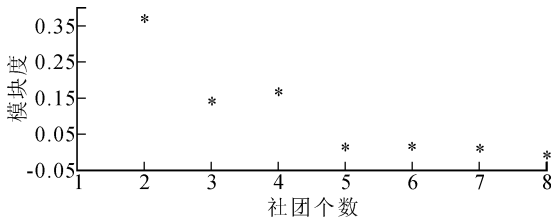


图 7 模块度值

Fig. 7 The values of modularity

美国大学足球联盟网络是根据 2000 年秋季常规赛季的比赛计划构建的[24]。网络中的节点代表球队,边代表两个球队之间常规赛季的比赛,它共包含 115 个节点和 616 条边。该网络通常由 8 ~ 12 个球队组成一个联盟(conference),同一个联盟球队之间的比赛次数多于不同联盟的球队间的比赛次数,每个联盟代表了一个真实社团。该网络自然划分为 12 个社团。采用本算法进行社团检测,社团划分结果如表 2 所示,当社团个数为 12 时,所得模块度最大为 0.6005,共有 11 个节点被错误划分,划分准确率为 90.43%。

表 2 football 网络社团划分结果

Tab. 2 The result of community detection of football network

社团	所含节点
1	1,5,10,17,24,42,94,105
2	2,26,34,38,46,90,104,106,110
3	3,7,14,16,33,40,48,61,65,101,107
4	4,6,11,41,53,73,75,82,85,99,103,108
5	8,9,22,23,52,69,78,79,109,112
6	12,25,29,51,70,91
7	13,15,19,27,32,35,39,43,44,55,62,72,86,100
8	18,21,28,57,63,66,71,77,88,96,97,114
9	20,30,31,36,56,80,81,83,95,102
10	37,59,60,64,98
11	45,49,58,67,76,87,92,93,113
12	47,50,54,68,74,84,89,111,115

3 结 论

(1)提出了一种新的网络节点相似性度量方法,该方法在检测网络社团方面比传统局部相似性指数更为准确,且计算复杂度不大;

(2)将该相似性和 K-means 谱聚类相结合,较为准确而有效地检测实际网络社团结构。文中定义

的相似性指数也可以和任何基于相似性的社团发现相结合以检测社团结构。

参考文献:

[1] Strogatz S H. Exploring complex networks[J]. Nature, 2001,410:268-276.

[2] Boccaletti S, Latora V, Moreno Y, et al. Complex networks: structure and dynamics[J]. Physics Reports, 2006, 424:175-308.

[3] 胡海波,王林. 关于因特网自治系统的连接率的幂律关系[J]. 西安理工大学学报,2005,21(2):204-206.
Hu Haibo, Wang Lin. Power law relationship of connection rates in ass of the internet[J]. Journal of Xi'an University of Technology, 2005,21(2):204-206.

[4] Santo Fortunato. Community detection in graphs[J]. Physics Reports, 2010,486:75-174.

[5] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal, 1970, 49:291-307.

[6] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69:026113.

[7] Zhou Tao, Lv Linyuan, Zhang Yicheng. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4): 623-630.

[8] Leicht E A, Holme P M E J. Newman, vertex similarity in networks[J]. Physical Review E, 2006, 73:026120.

[9] Sørensen T A. Method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons[J]. Biology Skr,1948,5(1):23-27.

[10] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):48-61.
Sun Jigui, Liu Jie, Zhao Lianyu. Clustering algorithms research[J]. Journal of Software,2008,19(1):48-61.

[11] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings National Academy Science, 2002, 99 (12) :7821-7826.

[12] Danon L, Diaz-Guilera A, Duch J. Comparing community structure identication[J]. Journal of Statistical Mechanics: Theory and Experiment, 2005,(9): 09008.

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

(责任编辑 李虹燕)