基于谱聚类的二分网络社区发现算法

张晓琴[1] 安晓丹[1] 曹付元[2]
(山西大学数学科学学院 太原 030006)[1] (山西大学计算机与信息技术学院 太原 030006)[2]

摘 要 二分网络是一类特殊的网络,在探索网络深层结构上具有重要作用。针对二分网络社区划分方法仍存在划分精度不高的问题,应用标准化谱聚类,提出了二分网络社区发现算法——谱聚类交互算法(SPCI)。首先,根据二分网络中两类节点之间的连边关系,构建相似性矩阵;然后,利用谱聚类算法将其中一类节点聚类;最后,利用交互度指标实现二分网络的社区划分。在人工数据和真实数据上的验证表明,SPCI不仅拥有比资源分布矩阵算法、边集聚系数算法和联合谱聚类算法更高的准确性和模块度,而且可以较为准确地确定社区划分个数。

关键词 二分网络,社区划分,谱聚类,相似性矩阵
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.04.034

Detecting Community from Bipartite Network Based on Spectral Clustering

ZHANG Xiao-qin[1] AN Xiao-dan[1] CAO Fu-yuan[2]
(School of Mathematics Sciences, Shanxi University, Taiyuan 030006, China)[1]
(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)[2]

Abstract Bipartite network is a special kind of network, which plays an important role in exploring the deep structure of the network. However, the methods of dividing the bipartite network community still have some problems, such as low precision of division. Through the application of normalized spectral clustering algorithm, an algorithm of detecting community- spectral clustering interaction (SPCI) was proposed. First, a similarity matrix is constructed based on the relationship between two kinds of nodes. Then, a cluster is clustered by spectral clustering algorithm. Finally, the community partition of two points network is realized by using two kinds of node's interaction index. Through the verification on artificial data and real data, the result shows that SPCI not only has higher accuracy and modularity than the algorithm based on resource distribution matrix, edge clustering coefficient and spectral co-clustering, but also can accurately determine the number of community partition.

Keywords Bipartite network, Community partition, Spectral clustering, Similarity matrix

1 引言

自然界中的许多生物系统和社会关系都存在复杂网络,如社会网络、技术网络、生物网络等。一般地,复杂网络都可以简化为一个图,图中的节点对应所研究的对象,对象之间的关系对应图中的连边。而且,这些网络大部分都会呈现出社区结构,即社区内部的连边分布较密集,社区间的连边分布较分散。社区发现对挖掘复杂网络的结构、预测对象的行为特征、提取网络的有用信息具有重要的研究意义。

二分网络是网络的一种重要类型,现实中许多真实的网络,如用户-产品购买关系网[1-2]、新陈代谢网络[3]、疾病-基因网络[4]等,都可以表示为二分网络。二分网络能体现出一些深层网络结构的特点,对网络结构特性的研究具有非常重要的作用[5]。

目前,主要有两种二分网络社区划分方法。一种是将其转化到单模网络上进行聚类和分析,如Melamed[6]提出的双映射方法、Murata[7]提出的二分网络模块度优化算法和高曼等[8]提出的PLP算法,但是这些方法会损失大量的网络信息。另外一种是直接在二分网络上进行划分,这些方法可以有效地保留原网络的大部分信息。例如,Barber等[9]在单模网络的基础上提出了适合二分网络社区划分的BRIM算法,虽然BRIM算法在小规模的二分网络上表现良好,但其不适合大规模的二分网络。Lehmann等[10]将K-clique算法[11]推广到了二分网络,并提出了bi-clique二分网络社区发现算法。bi-clique算法的主要优点是可以发现重叠社区和不同连接密度的社区;缺点是只挖掘了那些密度超过指定阈值的二分网络[12]。陈伯伦等[13]应用矩阵分解提出了基于矩阵分解的二分网络社区发现算法,但该算法依然存在划分精度不高的问题。

到稿日期:2018-02-29 返修日期:2018-05-28 本文受国家自然科学基金(61573229),山西省回国留学人员科研资助项目(2017-020),山西省基础研究计划项目(201701D121004),山西省高等学校教学改革创新项目(J2017002)资助。
张晓琴(1975-),女,博士,硕士生导师,主要研究方向为统计机器学习,E-mail:zhangxiaoqin@sxu.edu.cn(通信作者);**安晓丹**(1991-),女,硕士生,主要研究方向为统计机器学习;**曹付元**(1974-),男,教授,博士生导师,主要研究方向为数据挖掘与机器学习。

由于二分网络社区划分方法仍存在准确性不高的问题，因此有必要对二分网络社区发现做进一步的研究。本文拟将标准化谱聚类算法应用到二分网络社区发现中，结合交互度指标提出一种新的二分网络社区发现算法。由于谱聚类算法与数据分布无关，因此只需要考虑数据点间的相似性；另外，谱聚类算法具有能有效处理稀疏数据聚类的特点，能够克服二分网络中同类节点间没有连边这一缺点，理论上可较好地划分二分网络中的一类节点。利用交互度将另一类节点加入已划好的社区中，两个较好的算法集成，可较好地保证社区内部的连边相对紧密而社区间的连边相对稀疏的要求。通过实验部分的验证，新算法确实能够在一定程度上提高聚类精度。在一个区间范围内，通过迭代更新模块度得到最优社区划分，可实现社区个数的自主确定。

2　二分网络的相关介绍

2.1　二分网络的有关概念

二分网络有两种类型的节点，且只有不同类型的节点间有连边。因此，二分网络可以表示为如图 1 所示的简单无向图 $G(X,Y,E)$，其中 X,Y 分别表示二分网络的两类节点，E 表示节点 X 和节点 Y 间的连边。

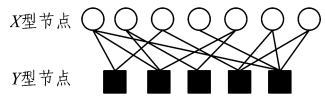

图 1　二分网络示意图

Fig. 1　Sketch map of bipartite network

设 X 型和 Y 型节点的个数分别为 m 和 n，由于 $X(Y)$ 型节点间没有连边，因此二分网络的邻接矩阵可以表示为：

$$A=\begin{bmatrix} 0_{m\times m} & \widetilde{A}_{m\times n} \\ \widetilde{A}^{\mathrm{T}}_{n\times m} & 0_{n\times n} \end{bmatrix}$$

其中，$0_{m\times m}$ 和 $0_{n\times n}$ 分别为 m 阶、n 阶的零矩阵；$\widetilde{A}_{m\times n}=(\widetilde{a}_{ij})_{m\times n}$，$\widetilde{a}_{ij}=\begin{cases}1, & \text{若节点 } x_i \text{ 与节点 } y_j \text{ 有连边}\\ 0, & \text{否则}\end{cases}$。

2.2　二分网络的模块度

为了度量社区划分结果，Guimera 等[14] 和 Barber[9] 基于 Newman 等[15] 提出的单模网络模块度，提出了二分网络的模块度。还有许多学者[7,16-17] 对二分网络模块度进行了改进。本文仅对 Barber 的二分网络的模块度做简单介绍。

鉴于二分网络同类节点间没有连边的特点，Barber 定义节点 x_i 与节点 x_j 的连边概率矩阵如下：

$$P=\begin{bmatrix} 0_{m\times m} & \widetilde{P}_{m\times n} \\ (\widetilde{P}^{\mathrm{T}})_{n\times m} & 0_{n\times n} \end{bmatrix}$$

其中，$\widetilde{P}_{m\times n}=(\widetilde{P}_{ij})_{m\times n}$，$\widetilde{P}_{ij}=k_{x_i}k_{y_j}/M$，$k_{x_i}$ 和 k_{x_j} 分别为节点 x_i 和节点 x_j 的度。二分网络的模块度定义为：

$$Q=\frac{1}{M}\sum_{i=1}^{m}\sum_{j=1}^{n}(\widetilde{A}_{ij}-\widetilde{P}_{ij})\delta(x_i,y_j) \tag{1}$$

其中，$\widetilde{A}_{ij}$ 是邻接矩阵中的元素；$\delta(x_i,y_j)=\begin{cases}1, & \text{若节点 } x_i \text{ 与节点 } y_j \text{ 属于同一个社区}\\ 0, & \text{否则}\end{cases}$。

2.3　二分网络的互信息指标

二分网络的互信息指标[14] 被用来比较两种社区划分结果的一致性，假设有两种社区划分 C 和 D，其互信息定义如下：

$$I_{CD}=\frac{-2\sum_{i=1}^{N^C}\sum_{j=1}^{N^D}n_{ij}^{CD}\lg\left(\frac{n_{ij}^{CD}M}{n_i^C n_j^D}\right)}{\sum_{i=1}^{N^C}n_i^C\lg\left(\frac{n_i^C}{M}\right)+\sum_{j=1}^{N^D}n_j^D\lg\left(\frac{n_j^D}{M}\right)}$$

其中，M 表示网络的总连边数；N^C 和 N^D 分别表示 C 和 D 划分的社区个数；n_i^C 表示算法 C 第 i 个社区的节点数；n_j^D 表示算法 D 第 j 个社区的节点数；n_{ij}^{CD} 表示算法 C 第 i 个社区和算法 D 第 j 个社区的公共节点数。

$$I_{CD}=\begin{cases}1, & \text{若划分 } C \text{ 与 } D \text{ 完全一致}\\ 0, & \text{若划分 } C \text{ 与 } D \text{ 不相关}\end{cases}$$

2.4　二分网络的交互度指标

应用交互度指标[18] 将 Y 类型中的节点加入到 X 类型的节点社区中。对于二分网络 $G(X,Y,E)$，假设 $\forall G_s\in G$，$y_j\in Y$，$G_s=X_s\cup Y_s$，$Y_s\subset Y$，$X_s\subset X$，那么交互度 (IND) 定义为：

$$IND(y_j,G_s)=\frac{\sum_i\sum_j A_{s'}(i,j)}{|X_s||Y_{s'}|} \tag{2}$$

其中，$A_{s'}$ 表示加入节点 y_j 后的邻接矩阵；$s'=s\cup j$，标号 j 为 Y 型节点的序号；$Y_{s'}$ 表示 $G_{s'}$ 加入节点 y_j 后 Y 型节点的集合；$|Y_{s'}|$，$|X_s|$ 分别表示 $Y_{s'}$ 和 X_s 包含的节点个数。

3　谱聚类交互算法

与传统聚类方法相比，谱聚类算法能够简化算法在高维数据上的表达，且可快速收敛到全局最优解。基于谱聚类算法的优良特性，Tang 等[19] 将谱聚类算法的思想推广到二分网络，对二分网络的邻接矩阵进行奇异值分解，提出了联合谱聚类算法。但是，联合谱聚类算法依然存在精确度不高和无法自主确定社区个数的问题。在此启发下，本文提出了一种新的社区发现算法——谱聚类交互算法（Spectral Clustering Interaction，SPCI）。与其他社区划分算法相比，SPCI 算法应用了聚类效果较好的标准化谱聚类算法，并构造了新的相似度矩阵，通过最大化模块度，自主地确定社区划分个数。该算法将分类效果较好的谱聚类算法与交互度集成，提高了二分网络聚类的精确度。

在给出本文的新算法 SPCI 之前，先给出如下 X 型节点的相似性矩阵的概念。

3.1　相似性矩阵的构造

相似性矩阵是通过两种类型节点间的连边获得的。对于二分网络 $G(X,Y,E)$，其相似性与共同邻居和节点度相关。节点 x_i 与节点 x_j 的共同邻居的定义如下：

$$N(x_i,x_j)=|N(x_i)\bigcap N(x_j)|$$

其中，$N(x_i)=\{y_k\in Y|(x_i,y_k)\in E\}$ 表示节点 x_i 的邻居节点集[20]。节点 x_i 与节点 x_j 的相似性[18] 为：

$$s_{ij}=N(x_i,x_j)\left(\frac{1}{k_{x_i}}+\frac{1}{k_{x_j}}\right) \tag{3}$$

其中，k_{x_i} 和 k_{x_j} 分别表示节点 x_i 和节点 x_j 的度。由于 X 型节点间没有连边，因此 X 型节点相似性矩阵中的元素由式 (3) 获得。为了更精确地应用谱聚类划分 X 型节点，将 X 型

节点的相似性矩阵定义为：

$$W=\begin{bmatrix} 0 & S_{12} & \cdots & S_{1p} \\ S_{21} & 0 & \cdots & S_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ S_{p1} & S_{p2} & \cdots & 0 \end{bmatrix} \tag{4}$$

其中相似性矩阵 W 的对角线元素全为 0，且矩阵 W 为对称矩阵。

3.2 谱聚类算法

谱聚类算法源于图划分问题：对于给定的无向图 $G(X,E)$，找到一种合理的分割，将节点分割为互不相交的集合。将谱聚类算法应用到社区划分中，则是要使得两个社区间切割的边数尽可能少。本文将标准化的谱聚类[21]应用到 X 型节点的划分中。

给定网络 $G(X,Y,E)$，对 X 型节点应用标准化谱聚类。其相似性矩阵如式(4)所示，节点 x_i 的相似度为：

$$d_i=\sum_{i=1}^{p}W_{ij}$$

则 X 型节点的相似度矩阵 D 为：

$$D=\begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix} \tag{5}$$

拉普拉斯矩阵为：

$$L=D-W$$

标准化拉普拉斯矩阵为：

$$L_{sym}=D^{-1/2}LD^{-1/2}=I-D^{-1/2}WD^{-1/2}$$

其中，I 为单位矩阵。标准化分割的目标函数为：

$$Ncut(C_1,\cdots,C_k)=\frac{1}{2}\sum_{i=1}^{k}\frac{W(C_i,\overline{C}_i)}{vol(C_i)}=\sum_{i=1}^{k}\frac{cut(C_i,\overline{C}_i)}{vol(C_i)}$$

其中，C_i 为第 $i(i=1,2,\cdots,k)$ 个分类，$\overline{C}_i$ 为 C_i 的补集；$W(C_i,\overline{C}_i)$ 为所切割的连边的权重和；$vol(C_i)$ 为属于 C_i 的节点的连边的权重和。

目标函数可等价为：

$$\min_{C_1,C_2,\cdots,C_k}\mathrm{Tr}(T'D^{-1/2}LD^{-1/2}T)$$

其中，$\mathrm{Tr}(\cdot)$ 表示矩阵的迹；T 的列向量为标准化拉普拉斯矩阵 L_{sym} 的前 k 个特征向量，且满足 $T'T=I$。

3.3 算法设计

本文提出了二分网络社区发现方法 SPCI。首先，根据二分网络不同类型节点间的连边情况，构建 X 型节点的相似性矩阵；然后，应用标准化谱聚类算法对 X 型节点进行划分；最后，由交互度指标将 Y 型节点合并到 X 型节点社区。通过分析二分网络的社区结构，获取谱聚类节点分类个数的取值范围，并根据此范围确定社区个数的初始值，经过迭代更新使得模块度最大化，即可得到二分网络社区划分的最终结果。算法的具体步骤如算法 1 所示。

算法 1 SPCI 算法

输入：二部图 G(X,Y,E)

输出：二分网络社区划分的最终结果 $F_1,F_2,\cdots,F_k$

Step 1 分析二分网络结构，依据二分网络图结构或经验，给出社区个数 k 的初值。

Step 2 计算 X 型节点的相似性矩阵 W。

Step 3 标准化拉普拉斯矩阵，得到 L_{sym}。

Step 4 计算标准化拉普拉斯矩阵 L_{sym} 的前 k 个特征向量 $u_1,u_2,\cdots,u_k$，并将 $u_1,u_2,\cdots,u_k$ 作为矩阵 $U\in R^{n\times k}$ 的列。

Step 5 标准化矩阵 U 的行向量，形成矩阵 T，使得 $t_{ij}=u_{ij}/\sum_k u_{ik}^2$。

Step 6 利用 k-means 对矩阵 T 的行向量进行聚类，得到 X 型节点分类 $C_1,C_2,\cdots,C_k$。

Step 7 对于每个 $y_i\in Y$，选择最优的节点满足 $l=\arg\max_k(|N(y_i)\cap C_k|)/|C_k|$，将 y_j 加入 C_l，更新 C_l；如果 $l\geqslant 2$，先不将其加入任何一个 C_l，将其余节点加入后，根据式(2)计算交互度 $e=\arg\max_s IND(y_i,F_s')$，然后将节点 y_j 加入社区 F_e'。

Step 8 依据式(1)计算二分网络的模块度。

Step 9 迭代更新，直到二分网络的模块度最大，模块度达到最大时的分类结果即为社区划分的最终结果 $F_1,F_2,\cdots,F_k$。

3.4 算法复杂度

算法 1 中 Step 2－Step 6 为标准化谱聚类算法，其时间复杂度[22]为 $O(m^3)$；Step 7 的时间复杂度为 $O(nk)$；Step 8 的时间复杂度为 $O(mn)$。因此，谱聚类交互算法的总时间复杂度为 $N(O(m^3)+O(nk)+O(mn))\approx N(O(m^3))$。$m$ 表示 X 型节点数，n 表示 Y 型节点数，N 为更新迭代次数，k 为社区划分个数。

假设节点的共同邻居均为 g，边集聚系数算法的时间复杂度为 $O(mng^2)$，资源分布的时间复杂度为 $O(m^2)$，虽然边集聚系数和资源分布的复杂度较低，但是其精确度同样也较低。联合谱聚类算法的时间复杂度与 SPCI 相同，也为 $O(m^3)$，但联合谱聚类算法比 SPCI 算法的聚类精度低。

4 实验结果与分析

为了验证 SPCI 的性能，本文分别在人工网络和真实网络上进行了实验，并将 SPCI 与 3 种经典算法(联合谱聚类算法(SPEC)、边集聚系数算法[23]、资源分布算法[24])作比较。

4.1 人工网络实验

为检验 SPCI 算法在给定网络结构下对社区划分的准确性，本文构造了 8 个人工二分网络进行测试。构建的每个网络中节点的度均为 $k=64$，$k=k_{in}+k_{out}$，k_{in} 表示节点与社区内部节点的连边数，k_{out} ($k_{out}=0,1,\cdots,7$) 表示节点与其他社区节点的连边数。具体的网络信息如表 1 所列。

表 1 构建的人工网络信息

Table 1 Constructed artificial network information

二分网络	节点总数	每个社区的节点数	社区个数	社区中每种类型的节点数	k_{out}
网络 1－网络 8	512	128	4	64	0～7

分别在 k_{out} 为 0～7 的二分网络上应用 SPCI 算法、联合谱聚类算法、边集聚系数算法和资源分布算法，并将互信息指标作为评价指标。如图 2 所示，不论在 $k_{out}=0$ 还是 $k_{out}=7$ 的人工网络上，SPCI 算法总能准确且自主地划分社区。由于 SPCI 算法始终是在给定社区个数 k 的范围内寻找使得模块度最大的社区划分，因此在每次实验中，SPCI 算法总能精确地找到社区划分个数 4；求得的互信息指标 $I=1$，故划分的准确性均为 1；而且算法稳定，不含可变参数。SPCI 算法不仅准确性比其他 3 种算法高，而且可以自主地确定社区划分的个数。其中，边集聚系数算法在人工网络数据集上的划分准

确性出现了较大的波动，且在 $k_{out}=4$ 时，正确率仅为 0.5338；资源分布算法和联合谱聚类算法在 k_{out} 为 0～7 的网络上虽无明显波动，但其正确率也不及 SPCI 算法。

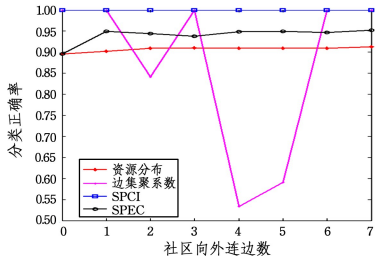

图 2　4 种算法的分类正确率比较

Fig. 2　Partition accuracy comparison of 4 algorithms

4.2　真实网络上的实验

本节分别在 Southern Women[25]，CEO Clubs[26]，Scotland[27]，Moreno Crime[1] 4 个真实数据上验证 SPCI 的精确度，并将 SPCI 算法与 3 种经典算法（联合谱聚类算法、边集聚系数算法、资源分布算法）进行模块度值的比较。

实验 1　Southern Women 数据集是由 Divas 等收集的，被广泛应用于二分网络的社区发现算法的检测中。此数据来自于 1930 年的密西西比州，是由 18 位妇女和 14 个活动组成的。若将妇女和活动之间的关系简化为一个图，则可构成一个二分网络。网络中的两类节点分别表示妇女和活动，网络中的连边表示妇女参加了相应的活动。

社区分布图如图 3 所示，将数据集中的妇女用圆形表示，编号为 1－18；活动用矩形表示，编号为 1－14。Southern Women Data 数据集上，SPCI 算法将其划分为 2 个社区。第一类社区为妇女{1,2,3,4,5,6,7,9}与活动{1,2,3,4,5,6,7,8}；第二类社区为妇女{8,10,11,12,13,14,15,16,17,18}与活动{9,10,11,12,13,14}。以模块度作为评价标准，其模块度值为 0.3291。在图 3 中，同一个社区中的节点用相同的灰度表示。

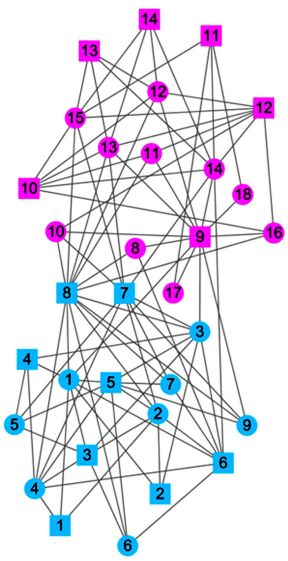

图 3　SPCI 在数据集 Southern Women 上的社区划分

Fig. 3　Community partition of SPCI on Southern Women

在 Southern Women 数据集上，分别采用边集聚系数和资源分布算法划分社区，并计算两种算法对应的模块度值。实验多次，取其最大的模块度值与 SPCI 进行比较。以模块度作为评价标准，比较结果如表 2 所列。

表 2　Southern Women 数据集上的模块度比较

Table 2　Modularity comparison on Southern Women

边集聚系数	资源分布	SPCI
0.3036	0.3082	0.3291

由表 2 可知，在 Southern Women 数据集上，由 SPCI 算法求得的模块度值比边集聚系数算法和资源分布算法求得的模块度值略高，社区结构更加明显，划分准确性也相对较高。

实验 2　CEO Clubs 数据集是由 Galaskiewicz 收集的，该数据集由 26 位 CEO 和 15 个俱乐部组成。同样，该数据可以构成一个二分网络，网络中两种类型的节点分别代表 CEO 和俱乐部，网络中的连边表示 CEO 和俱乐部之间的关系。

社区分布图如图 4 所示，在 CEO Clubs 数据集上，SPCI 算法将其分为 4 个社区。将 CEO 用圆形表示，编号为 1－26；俱乐部用矩形表示，编号为 1－15。SPCI 算法划分的第 1 个社区为 CEO {1,2,10,11,13,19}与俱乐部{3,5,7,9}；划分的第 2 个社区为 CEO{3,14,17,18,20,26}与俱乐部{11,12,15}；划分的第 3 个社区为 CEO{4,5,6,15,16,22,23,25}与俱乐部{2,6,13,14}；划分的第 4 个社区为 CEO {7,8,9,12,21,24}与俱乐部{1,4,8,10}。相应的模块度值为0.3183，在图 4 中，同一个社区中的节点用相同的灰度表示。

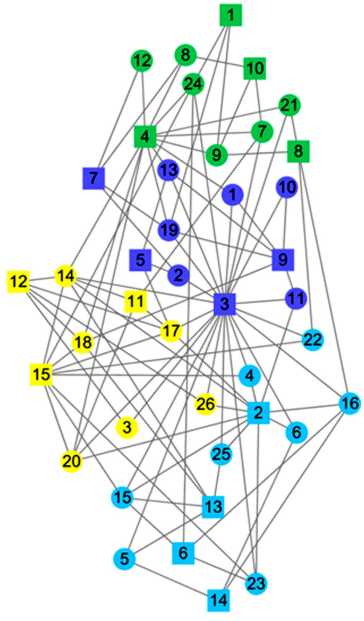

图 4　SPCI 在数据集 CEO Clubs 上的社区划分

Fig. 4　Community partition of SPCI on CEO Clubs

同样，分别采用边集聚系数、资源分布和联合谱聚类 3 种算法划分社区，同时计算这 3 种算法对应的模块度值。进行多次实验，取 3 种算法求得的最大模块度值与 SPCI 进行比较。以模块度作为评价标准，在 CEO Clubs 数据集上的模块度值比较结果如表 3 所列。

[1] http://konect.uni-koblenz.de/networks/moreno_crime

表 3 CEO Clubs 数据集上的模块度比较

Table 3 Modularity comparison on CEO Clubs

边集聚系数	资源分布	联合谱聚类	SPCI
0.1558	0.2640	0.2787	0.3183

由表 3 可知，在 CEO Clubs 数据集上，SPCI 算法的模块度值大于 3 种经典算法的模块度值，因此 SPCI 算法划分的社区结构更加明显，划分准确性也相对较高。

实验 3 Scotland 是一个较典型的二分网络数据集，常被用作二分网络算法的检测。该数据集描述了苏格兰早期 108 家连锁企业与 136 位股东之间的任职情况。将该数据表示成一个二分网络，网络中的两类节点代表公司和股东，节点间的连边表示该股东在公司任职。

在 Scotland 数据集上，SPCI 将其分为 14 个社区。例如，公司{64}与股东 {34,107,81}为一个社区；公司{3,16,36,131}与股东{63,93,103,104}为一个社区；公司{1,2,4,5,6,7,8,23,69}与股东{1,7,16,21,23,50,53}为一个社区。相应的模块度值为 0.6857。

分别采用边集聚系数、资源分布和联合谱聚类 3 种算法划分社区，同时计算这 3 种算法对应的模块度。进行多次实验，取最大的模块度值与 SPCI 进行比较。以模块度作为评价标准，在 Scotland 数据集上的模块度值比较结果如表 4 所列。

表 4 Scotland 数据集上的模块度比较

Table 4 Modularity comparison on Scotland

边集聚系数	资源分布	联合谱聚类	SPCI
0.2453	0.4963	0.6599	0.6857

由表 4 可知，在 Scotland 数据集上，SPCI 算法的表现更佳。SPCI 算法的模块度值均大于 3 种经典算法的模块度值，这充分说明 SPCI 算法可较好地挖掘社区结构，社区划分的准确性更高。

实验 4 Moreno Crime 数据集由刑事案件和与刑事案件有关的人员构成。一类节点表示案件中的一个受害者、目击者或嫌疑人，另一类节点是相应的案件。这个数据集包含有 829 名人员和 551 起案件，总连边数为 1476 条。

在 Moreno Crime 数据集上，SPCI 将其分为 43 个社区。例如，人员{513}与案件{464,465,47}为一个社区；人员{84,242,255,398,543,625,725,752}与案件{138,284,287,303,304,305,484,485}为一个社区；案件{117,252,253,351,352,353,495,496,497,498,520}与人员{193,194,333,592,692}为一个社区。相应的模块度值为 0.9116。

分别采用边集聚系数、资源分布和联合谱聚类 3 种算法划分社区，同时计算这 3 种算法对应的模块度。在 Moreno Crime 数据集上的模块度值比较结果如表 5 所列。

表 5 Moreno Crime 数据集上的模块度比较

Table 5 Modularity comparison on Moreno Crime

边集聚系数	资源分布	联合谱聚类	SPCI
0.1046	0.1713	0.9064	0.9116

由表 5 可知，在 Moreno Crime 数据集上，SPCI 算法的精确度高于 3 种经典算法的精确度，其能够较好地识别社区。

结束语 针对目前二分网络社区划分精确度不高、社区个数难以确定的问题，本文提出了二分网络社区划分方法——谱聚类交互法(SPCI)，其在一定程度上缓解了社区划分时出现的上述两个问题。SPCI 通过构造相似性矩阵，应用标准化谱聚类和交互度指标对二分网络社区进行划分。在真实网络数据集和人工网络数据集的验证中，SPCI 的精确度和模块度值都相对较高，而且能够较准确地确定社区个数。但是 SPCI 算法中社区个数的确定要先通过分析社区结构给出范围，再由最大化模块度得到最终的社区划分个数。社区个数的确定在后续学习中还需做进一步的探索性研究。

参 考 文 献

[1] SHANG M, LU L, ZHANG Y C, et al. Empirical analysis of web-based user-object bipartite networks[J]. Epl, 2012, 90(4): 1303-1324.

[2] YANG Z, ZHANG Z K, ZHOU T. Anchoring bias in online voting[J]. Computer Science, 2012, 100(6): 1-6.

[3] JEONG H, TOMBOR B, ALBERT R, et al. The large-scale organization of metabolic networks [J]. Nature, 2000, 407(6804): 651-654.

[4] JUAN, SHENG, ZHANG, et al. Traffic characteristics based dynamic radio resource management in heterogeneous wireless networks [J]. China Communications, 2014, 11(1): 1-11.

[5] CHEN D, YAN Y, WANG D, et al. Community detection algorithm based on structural similarity for bipartite networks[C]// IEEE International Conference on Software Engineering and Service Science. IEEE, 2017: 98-102.

[6] MELAMED D. Community structures in bipartite networks: a dual-projection approach[J]. Plos One, 2014, 9(5): 1-5.

[7] MURATA T. Detecting communities from bipartite networks based on bipartite modularities[C]// International Conference on Computational Science and Engineering. IEEE Computer Society, 2009: 50-57.

[8] GAO M, CHEN L, XU Y C. Projection based algorithm for link prediction in bipartite network [J]. Computer Science, 2016, 43(2): 118-123. (in Chinese)
高曼，陈崚，徐永成. 基于投影的二分网络链接预测[J]. 计算机科学，2016，43(2)：118-123.

[9] BARBER M J. Modularity and community detection in bipartite networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2007, 76(2): 1-9.

[10] LEHMANN S, SCHWARTZ M, HANSEN L K. Biclique communities[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2008, 78(2): 1-9.

[11] GREGORI E, LENZINI L, MAINARDI S. Parallel k-clique community detection on large-scale networks[J]. IEEE Transactions on Parallel & Distributed Systems, 2013, 24(8): 1651-1660.

[12] LIU X, MURATA T. Community detection in large-scale bipar-

tite networks[C]// IEEE /WIC /ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies,2009. WI-LAT:IEEE,2009:50-57.

[13] CHEN B L,CHEN L,ZOU S R,et al. Detecting community structure in bipartite networks based on matrix factorization [J]. Computer Science,2014,41(2):55-58. (in Chinese)
陈伯伦,陈崚,邹盛荣,等. 基于矩阵分解的二分网络社区挖掘算法[J]. 计算机科学,2014,41(2):55-58.

[14] GUIMERA R,SALESPARDO M,Amaral L A N. Module identification in bipartite and directed networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics,2007,76(3 Pt 2):1-8.

[15] NEWMAN M E J,GIRVAN M. Finding and evaluating community structure in networks [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics,2004,69(2):1-15.

[16] LI Z,ZHANG S,ZHANG X. Modularity and community detection in bipartite networks[J]. American Journal of Operations Research,2015,5(5):421-434.

[17] DORMANN C F,STRAUSS R. A method for detecting modules in quantitative bipartite networks[J]. Methods in Ecology & Evolution,2014,5(1):90-98.

[18] GUO G G,QIAN Y H,ZHANG X Q,et al. Algorithm of detecting community in bipartite network with autonomous determination of the number of communities[J]. PR & AI,2015,28(11):969-975. (in Chinese)
郭改改,钱宇华,张晓琴,等. 自主确定社区个数的二模网络社区发现算法[J]. 模式识别与人工智能,2015,28(11):969-975.

[19] TANG L,LIU H. 社会计算:社区发现和社会媒体挖掘[M]. 文益民,闭应洲,译. 机械工业出版社,2013:78-83.

[20] ZHOU Y,SUN G,XING Y,et al. Local community detection algorithm based on minimal cluster[J]. Applied Computational Intelligence and Soft Computing,2016,2016(2):1-11.

[21] LUXBURG U V. A Tutorial on spectral clustering [J]. Statistics & Computing,2007,17(4):395-416.

[22] YAN D,HUANG L,JORDAN M I. Fast approximate spectral clustering [C] // ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. ACM,2009:907-916.

[23] ZHANG P,WANG J,LI X,et al. Clustering coefficient and community structure of bipartite networks[J]. Physica A Statistical Mechanics & Its Applications,2008,387(27):6869- 6875.

[24] WU Y J. A Clustering algorithm for bipartite network based on distribution matrix of resources[J]. Journal of Beijing Normal University,2010,46(5):643-646. (in Chinese)
吴亚晶,狄增如,樊瑛. 基于资源分布矩阵的二分网聚类方法[J]. 北京师范大学学报(自然科学版),2010,46(5):643-646.

[25] DAVIS A,GARDNER B B,GARDER M R. Deep South:a social anthropological study of caste and class[J]. American Journal of Sociology,1941,2(3):117-118.

[26] SCOTT J,HUGHES M,MACKENZIE J. The anatomy of Scottish capital:Scottish companies and Scottish Capital,1900-1979 [J]. Economic History Review,1980,33(4):271-275.

[27] WASSERMAN S,FAUST K. Social network analysis:methods and applications[J]. Contemporary Sociology,1994,91(435):219-220.