基于二部图表示的属性网络社区发现算法

赵兴旺，薛晋芳

引用本文

赵兴旺，薛晋芳. 基于二部图表示的属性网络社区发现算法[J]. 计算机科学，2023，50(11)：107-113.

ZHAO Xingwang, XUE Jinfang. Community Discovery Algorithm for Attributed Networks Based on Bipartite Graph Representation [J]. Computer Science, 2023, 50(11): 107-113.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000226

基于二部图表示的属性网络社区发现算法

赵兴旺[1,2]　薛晋芳[1]
1 山西大学计算机与信息技术学院　太原 030006
2 山西大学计算智能与中文信息处理教育部重点实验室　太原 030006

摘　要　属性网络社区发现是网络数据分析中的一项重要研究内容。为了提高社区发现的准确性,现有算法大多通过融合拓扑信息和属性信息对属性网络进行低维表示,然后基于低维特征进行社区发现。然而,这类算法通常基于深度模型进行表示学习,缺乏一定的可解释性。因此,文中提出了一种基于二部图表示的属性网络社区发现算法,以提高社区发现结果的准确性和可解释性。首先,分别基于属性网络的拓扑信息和属性信息计算网络中各个节点作为代表点的概率,通过两类信息融合选出一定比例的节点作为代表点;其次,基于拓扑结构和节点属性计算各个节点到代表点的距离,构建二部图;最后,基于二部图利用谱聚类算法进行社区发现,得到最终结果。在人造属性网络和真实属性网络上与已有的属性网络社区发现算法进行实验比较分析。实验结果表明,所提算法在标准化互信息、调整兰德指数等评价指标上均优于已有算法。

关键词:属性网络;社区发现;二部图;融合
中图法分类号　TP391

Community Discovery Algorithm for Attributed Networks Based on Bipartite Graph Representation

ZHAO Xingwang[1,2] and XUE Jinfang[1]
1 School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China
2 Key Laboratory of Computational Intelligence and Chinese Information Processing,Ministry of Education,Shanxi University,Taiyuan 030006,China

Abstract　Community discovery in attributed networks is an important research content in network data analysis. To improve the accuracy of community discovery,most existing algorithms perform low-dimensional representation of attributed networks by fusing topological and attributed information,and then perform community discovery based on low-dimensional features. Such algorithms,however,are typically based on deep learning models for representation learning,which lack interpretability. Therefore,in order to improve the accuracy and interpretability of community discovery results,this paper proposes a community discovery algorithm for attributed networks based on bipartite graph representation. Firstly,the topological and attributed information of the attributed networks are used to calculate the probability of each node serving as a representative point in the network,and a certain proportion of nodes are chosen as representative points. Secondly,based on the topological structure and node attributes,the distances of each node to the representative points are calculated to construct a bipartite graph. Finally,based on the bipartite graph,the result is obtained by using the spectral clustering algorithm for community discovery. Experiments are carried out on artificial and real attributed networks to compare and analyze the proposed algorithm and the existing algorithms. In terms of evaluation indices such as normalized mutual information and adjusted rand index,experimental results show that the proposed algorithm outperforms the existing algorithms.

Keywords　Attributed networks,Community discovery,Bipartite graph,Fusion

1　引言

属性网络不仅包含节点与节点之间的拓扑关系,各个节点还拥有丰富的属性信息,例如对节点的文字描述、与节点相关的评论以及节点特有的图片。属性网络被广泛用于对现代信息系统的建模,如在线和离线的个人关系社交网络、基因、代谢物和神经元之间相互作用的生物网络等。属性网络的普适性和灵活性使其具有重要的研究意义和应用价值。

属性网络的一个显著特征是其蕴含一定的社团结构,即将节点划分成组,同组中的节点紧密连接或共享相似的特征。

到稿日期:2022-10-26　返修日期:2023-03-04
基金项目:国家自然科学基金(62072293,62272285)
This work was supported by the National Natural Science Foundation of China(62072293,62272285).
通信作者:赵兴旺(zhaoxw84@163.com)

社区发现是属性网络分析挖掘中的重要任务之一，对于网络拓扑结构分析、功能分析和行为预测具有重要意义[1-2]。近年来，社区发现方法已经在社会科学、生物医学和计算机科学等多个方面取得了重要应用[3]。

针对不同领域的应用需求，研究者近年来已经开展了广泛研究，并提出了系列属性网络社区发现算法[4]，主要包括基于节点属性加权的方法[5-8]、基于非负矩阵分解的方法[9-11]、基于概率模型的方法[12-14]和基于嵌入表示的方法[15-17]等。其中，基于嵌入表示的方法作为一种解决属性网络分析问题的有效方法，在保留属性网络原始信息最大化的情况下，将网络中每个节点的属性信息及其拓扑结构同时映射至一个联合的低维向量，进而基于低维向量表示，利用传统聚类算法进行社区发现。基于嵌入表示的属性网络社区发现方法可以更全面地探索网络中蕴含的结构信息，但也面临一定的挑战。例如，属性网络节点的拓扑结构和属性信息从内容到形式截然不同，呈现出了一定的异质性，使得两类信息难以有效地融合。如何高效地从这些异质信息中提取出同质的有用信息进行联合低维表示，是基于嵌入表示社区发现算法的核心所在。低维表示后的特征较为抽象，不具有解释性，存在模型构建和后期社区发现结果可解释性不足的问题。

针对上述问题，本文提出了一种基于二部图表示的属性网络社区发现算法，旨在提高社区发现结果的准确性和可解释性。具体地，(1)基于属性信息，借鉴密度峰值聚类算法[18]的思想计算各个节点作为代表点的概率，基于网络拓扑信息利用节点的结构度量指标计算每个节点作为代表点的概率，将两类信息融合后选出一定比例的节点作为代表点；(2)利用属性信息计算各个节点到代表点的余弦距离，利用结构信息计算各个节点到代表点的最短路径长度，将两类信息有效融合得到各个节点到代表点的距离，构建二部图；(3)基于二部图利用谱聚类算法[19]进行社区发现得到最终结果。

2 相关工作

2.1 密度峰值聚类算法

Rodriguez 等[18]于 2004 年提出了快速搜索和发现密度峰值的聚类算法，简称为密度峰值聚类算法。该算法基于两个假设：(1)同一簇中，簇中心被簇中其他密度较低的数据点包围；(2)不同簇的簇中心距离相对较远。为了找到满足这两个条件的簇中心，引入了局部密度的概念。假设数据点 x_i 的局部密度为 ρ_i，数据点 x_i 到局部密度比它大且距离最近的数据点 x_j 的距离为 δ_i。ρ_i 和 δ_i 的定义分别如下：

$$\rho_i=\sum_{j\neq i}\chi(d_{ij}-d_c) \tag{1}$$

$$\delta_i=\min_{j:\rho j>\rho i}(d_{ij}) \tag{2}$$

其中，d_{ij} 为 x_i 和 x_j 之间的距离；dc 为截断距离；$\chi(\cdot)$ 为逻辑判断函数，如果 $(\cdot)<0$，则 $\chi(\cdot)=1$，否则 $\chi(\cdot)=0$。式(2)中，当 xi 为局部密度最大的数据点时，$\delta_i=\max_{j\neq i}(d_{ij})$。

根据以上定义，构造 δi 相对于 ρ_i 的决策图，基于决策图确定簇中心，并将其余数据点分配到相应的类簇，得到最终的聚类结果。

2.2 谱聚类

谱聚类[19]是由 Luxburg 于 2007 年提出的一种基于图论的聚类算法，它在任意分布数据都能获得更好的聚类结果。假设 $X=\{x_1,x_2,\cdots,x_n\}^{\mathrm{T}}\in R^d$ 表示 d 维特征描述的 n 个样本，其中 x_i 表示第 i 个样本，谱聚类算法的具体过程如下。

(1)构建相似矩阵 $\boldsymbol{W}$。构建相似矩阵常常使用全连接法。一般使用不同的核函数来定义 w_{ij}，最常用的是高斯核函数。其定义如下：

$$w_{ij}=\exp\left(-\frac{\|x_i-x_j\|_2^2}{2\sigma^2}\right) \tag{3}$$

其中，σ 是一个尺度参数，控制高斯核函数的邻域宽度。

(2)构建度矩阵 $\boldsymbol{D}$。$\boldsymbol{D}$ 为对角矩阵，其定义如下：

$$d_i=\sum_{j=1}^{n}w_{ij} \tag{4}$$

即相似矩阵 $\boldsymbol{W}$ 的每一行元素之和。

(3)构建拉普拉斯矩阵 $\boldsymbol{L}$。其定义如下：

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W} \tag{5}$$

然后归一化拉普拉斯矩阵 $\boldsymbol{L}$，归一化后的拉普拉斯矩阵记为 $\boldsymbol{L}_{\mathrm{sym}}$。其定义如下：

$$\boldsymbol{L}_{\mathrm{sym}}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} \tag{6}$$

(4)计算归一化后的拉普拉斯矩阵 $\boldsymbol{L}_{\mathrm{sym}}$ 最小的 k_1 个特征值对应的特征向量 $\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_{k_1}$，形成一个 $n\times k_1$ 矩阵，记为 $\boldsymbol{U}=\{\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_{k_1}\}$。

(5)对 $\boldsymbol{U}$ 的每一行依次单位化，然后输入聚类个数 k_2，使用 K-Means 聚类方法进行聚类得到簇划分 $C=\{c_1,c_2,\cdots,c_{k_2}\}$。

2.3 属性网络社区发现算法

现有的属性网络社区发现算法大致分为基于节点属性加权的方法、基于非负矩阵分解的方法、基于概率模型的方法和基于嵌入表示的方法。

基于节点属性加权的方法将属性网络转化为加权图，边的权值包含节点之间基于属性信息的相似度。Steinhaeuser 等[5]提出了属性相似度的概念，当相邻节点存在相同属性值时，边的权重加 1；如果属性值是连续的，则将属性值归一化后计算差值作为节点间的属性相似度，即节点间边的权重。Combe 等[6]提出的基于属性信息聚类的模型利用属性信息计算节点间的余弦距离，将距离和边关联起来，再利用 K-Means 算法划分社区。Huang 等[7]构造加权图时，为了使边的权重不为 0，初始化阶段设置边的权重为 1，当两个节点有相同的属性值时，边的权重加 1，否则不加 1，然后利用传统的标签传播算法进行社区发现。Alinezhad 等[8]计算每个节点对的属性相似度，如果它们之间存在边，则属性相似度作为边的权重，否则权重为 0，然后通过 MILP 模型进行社区发现。该类方法主要通过节点的属性信息度量节点之间的相似性，未能有效平衡网络拓扑信息和节点属性信息在社区发现过程中的作用。

基于非负矩阵分解的方法利用低秩非负矩阵的乘积逼近高秩的非负矩阵，使 Frobenius 范数最小。Wang 等[9]提出了一种新颖的非负矩阵因子模型，该模型有两组参数，分别是社区成员矩阵和社区属性矩阵。Qin 等[10]在构造矩阵分解

模型过程中不仅融合了网络的拓扑结构和属性信息，而且考虑了拓扑结构和节点属性不匹配的问题。Pei 等[11]提出了一种基于用户、消息和两者交互的正则化非负矩阵三因子分解聚类框架，该框架能够将拓扑结构和属性信息无缝结合。该类方法中非负矩阵分解的优化过程与初始值有关，当矩阵初始值与最优解矩阵相差较大时，计算成本变大。

基于概率模型的方法指假定网络结构和节点属性服从选定的参数分布，从概率上推断网络中的社区成员分布。Xu 等[12]提出了基于贝叶斯概率模型的社区发现算法，该算法假设同属于一个社区的节点拓扑关系服从伯努利分布、节点属性服从多项式分布，从而将聚类问题转化为标准概率推理问题。Yang 等[13]提出了结合社区成员、网络拓扑结构和节点属性的概率模型，通过计算概率来获得聚类结果。Xu 等[14]提出了一种新颖的贝叶斯模型，用于进行社区发现，它不仅避免了先前方法中人为设计距离的弊端，还可以被应用在不同类型的属性网络中。这类方法在优化过程中涉及的参数较多，计算开销相对较大，此外该类方法的有效性在很大程度上依赖于对数据先验概率分布的预先估计，当属性网络模型的真实概率分布未知时，有效性将大大降低。

基于嵌入表示的方法通过融合网络的结构信息和属性信息，将其嵌入到低维向量空间，利用传统的聚类算法进行社区发现。Gao 等[15]提出了一种新颖的属性网络嵌入方法，该方法将节点的高阶邻接矩阵和节点属性矩阵作为输入，通过自动编码器获得节点的嵌入表示，并且提出了一种负采样策略，使得到的节点嵌入更准确。Liao 等[16]通过深度神经网络模型来获得节点的嵌入表示，在输入层早期融合网络的结构信息和节点属性信息，使后续训练过程包含两者的相互关系。Wang 等[17]提出了一种图自编码器算法，该算法将结构信息和属性信息集成到深度学习框架中，最后利用谱聚类算法进行社区发现。然而，该类算法得到的低维向量表示较为抽象，节点每维向量的关系的可解释性不足。此外，属性网络结构信息和属性信息的表现方式截然不同，如何高效利用这两类异质信息进行联合低维表示仍是此类算法问题的核心。

3 基于二部图表示的属性网络社区发现算法

3.1 问题描述

属性网络可表示为 $G=(V,\boldsymbol{E},\boldsymbol{A},\boldsymbol{F})$，其中 $V=\{v_i\}$ 表示节点的集合，$|V|=n$。$E=\{e_{ij}\}$ 表示边的集合，e_{ij} 表示节点 v_i 和 v_j 之间的边，$|E|=m$。$\boldsymbol{A}=[a_{ij}]_{n\times n}$ 表示属性网络的邻接矩阵，若节点 v_i 和 v_j 存在边，则 $a_{ij}=1$，否则 $a_{ij}=0$。$\boldsymbol{F}=[f_{ij}]_{n\times d}$ 表示属性网络的属性矩阵，f_{ij} 表示节点 v_i 的第 j 个属性，属性分为分类型和数值型，当属性为分类型时，属性值是离散值；当属性为数值型时，属性值是连续值。

属性网络的社区发现旨在将 G 中的 n 个节点划分成 k 个社区，即 $C(G)=(c_1,c_2,\cdots,c_k)$，并且划分的社区满足以下性质。

1）结构紧密性：即社区内部的节点结构连接紧密，社区之间的节点连接稀疏。

2）属性同质性：即社区内部的节点属性向量相似，社区之间的节点属性向量相异。

3.2 算法设计

本文算法的主要步骤如图 1 所示，包括代表点选取、二部图构建和谱聚类这 3 步。代表点选取阶段分别基于属性网络的拓扑信息和属性信息计算各个节点作为代表点的概率，融合两类信息选取一定比例的节点作为代表点；二部图构建阶段基于拓扑结构和节点属性计算各个节点到代表点的距离，构建二部图；谱聚类阶段基于二部图利用谱聚类算法进行社区发现，得到网络最终的社区结构。

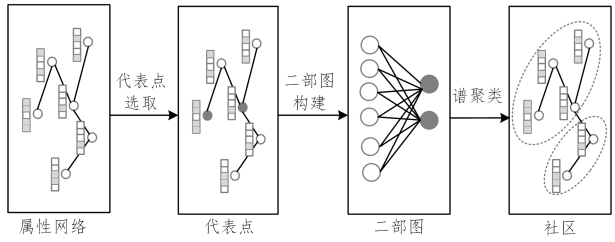


图 1 算法流程示意图

Fig. 1 Flow chart of the proposed algorithm

3.2.1 代表点选取

该步骤的主要目的是通过计算各个节点作为代表点的概率来选取一定的代表点。首先，基于结构信息计算每个节点作为代表点的概率。本文选取 pagerank[20]、特征向量中心性[21]和度中心性这 3 个指标，从不同角度度量节点的重要性。

pagerank 是衡量节点重要性的一种度量指标，其值表示节点在网络的全局影响力。其计算式如下：

$$PR(v_i)=\frac{(1-\omega)}{n}+\omega\sum_{v_j\in N(v_i)}\frac{PR(v_j)}{degree(v_j)} \tag{7}$$

其中，ω 取值为 0.85，$N(v_i)$ 表示节点 v_i 的邻居，$degree(v_j)$ 表示节点 v_j 的度数。初始化阶段，$PR(v_i)=\frac{1}{n}$，通过不断迭代来计算节点的 PR 值。

特征向量中心性指标通过邻居节点的数量和邻居节点的重要性来度量当前节点的重要性。其计算式如下：

$$EC(vi)=k_1^{-1}\sum_{v_j}a_{ij}EC(v_j) \tag{8}$$

其中，k_1 表示邻接矩阵 $\boldsymbol{A}$ 的最大特征值，a_{ij} 表示矩阵 $\boldsymbol{A}$ 的第 i 行第 j 列的元素值。在初始化阶段，对于所有节点 v_i，令 $EC(v_i)=1$。

度中心性表示该节点与其他节点相联系的程度。节点的度中心性值越大，该节点在网络中就越重要。其计算式如下：

$$DC(v_i)=\frac{degree(v_i)}{(n-1)} \tag{9}$$

通过计算节点的 pagerank 值、特征向量中心性值和度中心性值三者的和来度量节点作为代表点的概率。为了使 3 个指标对结果的贡献相同，这里将 3 个指标的取值分别归一化，$PR=\{PR(v_1),PR(v_2),\cdots,PR(v_n)\}$，$EC=\{EC(v_1),EC(v_2),\cdots,EC(v_n)\}$ 和 $DC=\{DC(v_1),DC(v_2),\cdots,DC(v_n)\}$ 分别表示所有节点的 PR 值，EC 值和 DC 值。归一化过程的计算式如下：

$$PR(v_i)'=\frac{PR(v_i)-\min(PR)}{\max(PR)-\min(PR)} \tag{10}$$

$$EC(v_i)'=\frac{EC(v_i)-\min(EC)}{\max(EC)-\min(EC)} \tag{11}$$

$$DC(v_i)'=\frac{DC(v_i)-\min(DC)}{\max(DC)-\min(DC)} \tag{12}$$

然后，求 3 个指标的均值，即：

$$p^s(v_i)=\frac{1}{3}(PR(v_i)'+EC(v_i)'+DC(v_i)') \tag{13}$$

借鉴密度峰值聚类算法的思想，利用属性信息计算每个节点作为代表点的概率。常见的数值属性数据计算距离的方法有欧氏距离和余弦距离。余弦距离由于通过两个向量夹角的余弦值来衡量两个向量差异的大小，因此更适合度量节点间基于属性信息的距离。故本文选取余弦距离来衡量节点基于属性信息的距离。具体过程如下：

1)计算节点 v_i 和 v_j 之间的余弦距离 $d^a(v_i,v_j)$，其计算式如下：

$$d^a(v_i,v_j)=1-\frac{f_{i1}f_{j1}+f_{i2}f_{j2}+\cdots+f_{id}f_{jd}}{\sqrt{f_{i1}^2+\cdots+f_{id}^2}\sqrt{f_{j1}^2+\cdots+f_{jd}^2}} \tag{14}$$

其中，$(f_{i1},f_{i2},\cdots,f_{id})$ 和 $(f_{j1},f_{j2},\cdots,f_{jd})$ 分别表示节点 v_i 和 v_j 的 d 维属性信息，当属性为分类型时，在 $(f_{i1},f_{i2},\cdots,f_{id})$ 和 $(f_{j1},f_{j2},\cdots,f_{jd})$ 中其属性值是离散的，当属性为数值型时，其属性值是连续的。

2)计算节点 v_i 的局部密度 ρ_i。将距离节点 v_i 小于 d_c 的节点的数量作为节点 v_i 的局部密度 ρ_i。本文中，d_c 取所有节点间的余弦距离按升序排列后的 1%～2%位置处的值。

$$\rho_i=\sum_{j=1,j\neq i}^{n}\chi(d^a(v_i,v_j)-d_c) \tag{15}$$

其中，$\chi(\cdot)$ 为逻辑判断函数，含义同式(2)。

3)计算节点 v_i 到比它局部密度大的其他节点的距离中的最小值 δ_i。将所有节点按照局部密度降序排列。当 v_i 不是局部密度最大节点时，δ_i 取与比其密度大且距离最近的节点之间的距离；当 v_i 是局部密度最大的节点时，δ_i 取节点 v_i 到其他节点的余弦距离的最大值。δ_i 的计算式如下：

$$\delta_i=\begin{cases}\min\limits_{j:\rho_j>\rho_i}(d^a(v_i,v_j)), & \text{if } \rho_i\neq\max\limits_j(\rho_j)\\ \max\limits_{j:j\neq i}(d^a(v_i,v_j)), & \text{otherwise}\end{cases},j=1,2,\cdots,n \tag{16}$$

4)计算基于属性信息节点 v_i 的成为代表点的概率 $p^a(v_i)$，用 $P^a=\{p^a(v_1),p^a(v_2),\cdots,p^a(v_n)\}$ 表示 n 个节点基于属性信息成为代表点的概率。

$$p^a(v_i)=(\rho_i\times\delta_i) \tag{17}$$

对 $p^a(v_i)$ 进行归一化，计算过程为：

$$p^a(v_i)'=\frac{p^a(v_i)-\min(P^a)}{\max(P^a)-\min(P^a)} \tag{18}$$

融合拓扑与属性两类信息得到每个节点成为代表点的概率 $p(v_i)$，权重为 α，定义为：

$$p(v_i)=\alpha p^s(v_i)+(1-\alpha)p^a(v_i)' \tag{19}$$

其中，α 表示融合两类信息计算节点作为代表点概率时结构信息所占的权重，取值范围为[0,1]。当 $\alpha>0.5$ 时，表明结构信息与属性信息相比，计算节点成为代表点的概率时前者更重要；当 $\alpha=0.5$ 时，表明此处结构信息和属性信息同等重要。

3.2.2 二部图构建

该步骤的目的是，在选取出代表点后，通过计算各个节点到代表点的距离来构建二部图。基于网络拓扑信息，本文利用最短路径长度计算所有节点到代表点的距离。具体过程为：计算所有节点到代表点的最短路径长度并进行归一化。用 $ST=\{ST(v_1,v_{p1}),ST(v_2,v_{p1}),\cdots,ST(v_n,v_m)\}$ 表示 n 个节点到代表点的最短路径长度。节点 v_i 到代表点 v_{pj} 的最短路径长度定义为：

$$d^s(v_i,v_{pj})=\frac{ST(v_i,v_{pj})-\min(ST)}{\max(ST)-\min(ST)} \tag{20}$$

基于属性信息方面，通过式(14)计算余弦距离，从而计算所有节点到代表点的距离。

通过融合网络拓扑信息和属性信息，得到节点 v_i 到代表点 v_{pj} 的距离，计算式如下：

$$d(v_i,v_{pj})=\alpha\times d^s(v_i,v_{pj})+(1-\alpha)d^a(vi,vpj) \tag{21}$$

其中，α 的含义与式(19)中 α 的含义相同，取值范围为[0,1]。为了保持前后一致性，并更好地研究 α 取不同值时对最终结果的影响，此处 α 的取值与式(19)中 α 的取值相同。基于式(21)得到任一节点与代表点之间的距离，即可构造节点与代表点之间的二部图。

基于以上步骤，本文对节点-代表点二部图构建的流程如算法 1 所示。

算法 1 二部图构建

输入：属性网络 G=(V,E,**A**,**F**)，权重 α，代表点比例 r

输出：节点-代表点二部图

1. for each node $v_i\in V$ do
2. for each node $vp_j\in$ 代表点 do
3. 利用式(20)计算节点 v_i 到 v_{pj} 基于结构信息的距离 $d^s(v_i,v_{pj})$；
4. 利用式(14)计算节点 v_i 到 v_{pj} 基于属性信息的距离 $d^a(v_i,v_{pj})$；
5. 利用式(21)计算综合距离 $d(v_i,v_{pj})$；
6. end for
7. end for
8. return 节点-代表点二部图

3.2.3 谱聚类

谱聚类使用较为广泛。与传统的 K-Means 算法相比，谱聚类使用了降维技术，因此更能解决复杂的聚类问题。该步骤利用谱聚类算法对上一步构建的二部图进行社区发现。具体过程为：通过二部图构造距离矩阵 $\boldsymbol{X}=[x_{ij}]_{n\times m}$，利用高斯核函数计算图的相似矩阵 $\boldsymbol{W}$，式(3)中令 $\sigma=\sqrt{0.5}$；然后特征向量的个数与社区个数的设置相同，均为 k；最后选用 discretize 算法[22]对行向量进行聚类。一般情况下，在谱聚类的最后阶段 K-Means 算法更受欢迎，然而该算法对初始聚类中心敏感，导致结果不稳定，discretize 算法的收敛速度更快，鲁棒性更好，因此本文选用 discretize 算法对 k 个特征向量进行聚类。

基于以上分析，本文算法的流程如算法 2 所示。

算法 2 基于二部图表示的属性网络社区发现算法

输入：属性网络 G=(V,E,**A**,**F**)，权重 α，代表点比例 r，社区个数 k

输出：k 个社区 $C=(c_1,c_2,\cdots,c_k)$

1. for each node $v_i \in V$ do

2. 利用式(13)计算 vi 基于结构信息成为代表点的概率 $p^s(v_i)$；

3. 利用式(18)计算 vi 基于属性信息成为代 表点的概率 $p^a(v_i)'$；

4. 利用式(19)融合二类信息计算节点成为代表点的概率 $p(v_i)$；

5. end for

6. 对概率进行降序排序，选取 rn 个代表点；

7. 利用算法 1 构建节点-代表点之间的二部图；

8. 针对二部图调用谱聚类算法。

3.2.4 时间复杂度分析

n 表示节点个数，d 表示属性个数，r 表示代表点比例，k 表示社区个数。本文算法的时间复杂度由 3 部分构成，第一部分为基于属性信息计算每个节点成为代表点的概率的时间复杂度，为 $O(n^2)$，利用结构信息计算每个节点成为代表点的概率的时间复杂度，为 $O(n)$；第二部分为计算各个节点到代表点的距离的时间复杂度，为 $O(rn^2)$；第三部分为对构建的二部图进行谱聚类的时间复杂度，为 $O(rn^2+kn^2)$。由于 $r \ll k$，因此总的时间复杂度为 $O(kn^2)$。

4 实验分析

为了验证本文算法的有效性，与已有的属性网络社区发现算法在人工合成属性网络和真实属性网络上进行了实验比较分析。

首先，通过人造属性网络数据集测试权重 α 和代表点比例 r 取不同值时对算法有效性的影响，选择算法较优时 α 和 r 的取值范围。其次，在 6 种真实属性网络上测试较优的 α 和 r 的取值范围下算法的有效性。最后，与目前先进的 4 种属性网络社区发现算法进行比较。算法有效性使用 NMI 和 ARI 这两个评价指标进行评估。

4.1 人工合成属性网络

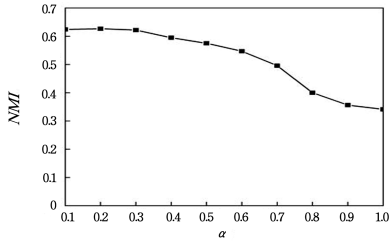
人工合成属性网络使用由 Elhadi 和 Agam[23] 于 2013 年提出的合成属性网络生成的属性数据集，该网络是 LFR 基准网络[24]的扩展。表 1 中参数的不同组合可以生成不同的数据集。生成网络是一个随机过程，每次运行都可能会产生不同的结果。本次实验选取了 3 组人工数据集来测试算法的有效性，通过 NMI 评价指标来进行评估。

表 1 合成属性网络数据集的参数设置

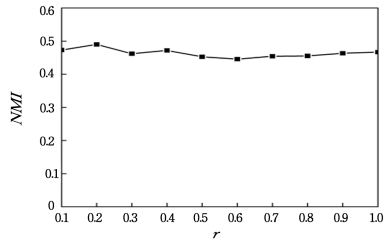
Table 1 Parameters information of artificial attributed networks

参数	参数值
n	1 000
k	10
max k	25
mu	0.3
t_1	2
t_2	1
min c	20
max c	50
on	0
om	0
$nattr$	4
$ainf$	1

在测试参数对算法有效性的影响时，取 3 组人工数据集运行结果的 NMI 平均值作为结果。先固定 r，令其等于 0.5，α 取[0,1]，实验结果如图 2(a)所示。然后固定 α，令其等于 0.5，r 取[0.1,1]，实验结果如图 2(b)所示。从图 2 中可以看出，α 在[0,0.3]时算法较优，r 在[0.2,0.4]时算法较优。



(a)$r=0.5$



(b)$\alpha=0.5$

图 2 人工网络的实验结果

Fig. 2 Experimental results on artificial attributed networks

4.2 真实属性网络

选取 6 个公开的真实数据集作为实验数据进行仿真，Cornell，Texas，Washington 和 Wisconsin 数据集分别是来自 4 所不同大学科学出版物的数据集，均属于 WebKB 网络的数据集。每个网络都有离散属性，所有属性的总个数为 1 703，节点分为 5 类，分别是课程、教室、学生、项目和工作人员。Terrorist_attack 数据集是由 UMD 的 MIND 实验室收集的恐怖主义相关信息的数据子集。该数据集包含 1 293 次恐怖袭击，共 106 维属性，每个攻击都用 0 或 1 来描述，其值表示属性的存在或缺失，节点分为纵火、轰炸、绑架、武器攻击、NBCR 攻击和其他攻击 6 类。Sinanet 网络是从新浪微博网站中提取的具有 3 490 个用户和 30 282 个关系的分类微博用户关系网络，每个节点都具有描述用户兴趣的 10 维数值属性。真实属性网络的数据信息如表 2 所列。

表 2 真实属性网络的信息描述

Table 2 Information description of real attributed networks

数据集	节点数	边数	属性个数	社区个数
Cornell	195	567	1 703	5
Texas	187	574	1 703	5
Washington	230	783	1 703	5
Wisconsin	265	938	1 703	5
Terrorist_attack	1 293	3 172	105	6
Sinanet	3 490	30 282	10	10

本文共选用了 4 种对比算法，分别是 SAS-LP 算法[25]、MGAE 算法[17]、CPRW-SI 算法[26] 和 AGC 算法[27]。SAS-LP 算法是一种新版本的标签传播算法，它将图中节点之间的结构相似度和属性相似度转化为边的权重，每个节点的中心性用 Laplacian centrality 来度量，将实验参数值设置为 Berahmand 等提供的源代码中的参数值。

MGAE算法基于图卷积网络得到节点的特征表示，该方法允许节点内容与网络特征交互，并在图自动编码器上下文中边缘化被破坏的特征，从而达到优化目的。实验参数的设置如下：噪音级别 p 设置为0.4，层数设置为3层，λ 设置为10^{-5}。CPRW-SI算法从内容传播的角度，提出了一种结合网络结构和节点属性的社区发现框架。它将网络视为一个动态系统，并将其社区视为节点之间交互的结果。该算法利用影响传播和随机游走两个原理来模拟相互作用。实验中，β 设置为1。本次选取了该文献中基于影响传播的CPRW算法，初始化策略为SI。AGC算法是一种自适应的图卷积方法。从图信号处理谱图理论的角度来理解GNN，利用高阶图卷积选择全局聚类结构，通过设计一个 k 阶的图卷积对节点特征进行低通滤波，从而获得平滑的特征表示。k 可以通过类内距离进行自适应选择，实验中设置 $max\_iter=60$。针对每个网络将各种算法重复运行50次，从中选取各种对比算法最优时NMI和ARI的取值作为最终的运行结果。针对本文算法，真实属性网络的结果取人工合成属性网络得到较优 α 和 r 取值范围内的最优值，实验结果如表3和表4所列。每列数据集中最优和次优的结果用粗体标记。

表3 不同算法的NMI值比较

Table 3 NMI performance of different algorithms

算法	数据集					
	Cornell	Texas	Washington	Wisconsin	Terrorist_attack	Sinanet
SAS-LP算法	**0.2231**	**0.1911**	**0.2148**	**0.2013**	0.0761	0.1160
MGAE算法	0.2099	0.1887	0.2128	0.1954	0.2089	0.2787
CPRW-SI算法	0.1389	0.0516	0.1574	0.1414	0.1930	0.5032
AGC算法	0.0448	0.0435	0.1186	0.1022	**0.2368**	**0.5657**
本文算法	**0.3267**	**0.3105**	**0.3830**	**0.4405**	**0.3172**	**0.6557**

表4 不同算法的ARI值比较

Table 4 ARI performance of different algorithms

算法	数据集					
	Cornell	Texas	Washington	Wisconsin	Terrorist_attack	Sinanet
SAS-LP算法	0.0819	**0.1501**	0.0863	0.0996	0.0353	0.0621
MGAE算法	**0.2969**	0.1302	**0.2223**	**0.1668**	0.1526	0.2026
CPRW-SI算法	0.0877	0.0944	0.1948	0.1036	**0.1632**	0.3711
AGC算法	0.0165	0.0165	0.1186	0.1171	0.1547	**0.4836**
本文算法	**0.3117**	**0.2555**	**0.3488**	**0.3881**	**0.3367**	**0.6307**

从表3和表4可以看出，本文算法在每个数据集上都取得了最大的NMI值和ARI值，总体上优于其他4种算法。SAS-LP算法在前4个数据集上表现较好，在Terrorist_attack数据集和Sinanet数据集上表现较差，这是因为Terrorist_attack数据集和Sinanet数据集都存在孤立点，特别是Terrorist_attack数据集，孤立点数量达到总节点数量的一半以上，这表明通过判断邻居节点标签来更新自身节点标签的SAS-LP算法的思想并不适用于存在孤立点的属性网络。由此可见，属性信息对社区划分具有重要意义，它能够有效解决由节点结构信息缺失造成的社区划分准确性低的问题。MGAE算法在各个数据集上的表现较为稳定，表明该算法在每种类型的属性网络中都可以有效地进行社区发现，而CPRW-SI算法和AGC算法在前4种数据集上表现一般，在Terrorist_attack数据集和Sinanet数据集上优于其他两种对比算法，说明CPRW-SI算法和AGC算法在属性网络中存在孤立点的情况下可以有效地进行社区发现。上述结果综合表明，无论属性网络是否存在孤立点，本文算法都表现出了较为良好的有效性，这表明该算法能够有效地挖掘代表点并指导社区发现。

结束语 本文提出了一种属性网络社区发现方法。首先，基于属性信息，借鉴密度峰值聚类算法的思想计算节点作为代表点的概率；其次基于拓扑信息，通过3个指标计算节点作为代表点的概率，融合节点两方面的概率信息，选取一定比例的节点作为代表点；然后利用属性信息计算节点之间的余弦距离，基于结构信息计算最短路径长度，融合两类信息构建二部图；最后利用谱聚类算法进行社区发现。实验结果表明，本文提出的基于二部图表示的属性网络社区发现算法在真实网络上优于目前先进的属性网络社区发现算法。然而，本文算法只适用于发现非重叠社区，在未来可将其扩展到属性网络的重叠社区发现方面。

参考文献

[1] BOTHOREL C, CRUZ J D, MAGNANI M, et al. Clustering attributed graphs: Models, measures and methods [J]. Network Science, 2015, 3(3): 408-444.

[2] FORTUNATO S, NEWMAN M E J. 20 years of network community detection [J]. Nature Physics, 2022, 18(8): 848-850.

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

[4] CHUNAEV P. Community detection in node-attributed social networks: A survey [J]. Computer Science Review, 2020, 37: 100286.

[5] STEINHAEUSER K, CHAWLA N V. Community detection in a large real-world social network [J]. Social Computing, Behavioral Modeling, and Prediction, 2008, 7: 168-175.

[6] COMBE D, LARGERON C, EGYED-ZSIGMOD E, et al. Combining relations and text in scientific network clustering [C]// Proceedings of the 2012 IEEE/ACM International Conference

on Advances in Social Networks Analysis and Mining. IEEE, 2012:1248-1253.

[7] HUANG B Y, WANG C K, WANG B B. NMLPA: Uncovering overlapping communities in attributed networks via a multi-label propagation approach [J]. Sensors, 2019, 19(2): 260-275.

[8] ALINEZHAD E, TEIMOURPOUR B, SEPEHRI M M, et al. Community detection in attributed networks considering both structural and attribute similarities: Two mathematical programming approaches [J]. Neural Computing and Applications, 2020, 32(8): 3203-3320.

[9] WANG X, JIN D, CAO X C, et al. Semantic community identification in large attribute networks [C]// Proceedings of the 13th AAAI Conference on Artificial Intelligence. Phoenix. AAAI Press, 2016: 265-271.

[10] QIN M, JIN D, HE D X, et al. Adaptive community detection incorporating topology and content in social networks[C]// Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2017: 675-682.

[11] PEI Y L, CHAKRABORTY N, SYCARA K. Nonnegative matrix tri-factorization with graph regularization for community detection in social networks[C]// Proceedings of the 24th International Joint Conference on Artificial Intelligence. AAAI Press, 2015: 2083-2089.

[12] XU Z Q, KE Y P, WANG Y, et al. A model-based approach to attributed graph clustering [C]// Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012: 505-516.

[13] YANG J, MCAULEY J, LESKOVEC J. Com munity detection in networks with node attributes [C]// Proceedings of the IEEE International Conference on Data Mining. IEEE, 2013: 1151-1156.

[14] XU Z Q, KE Y P, WANG Y, et al. GBAGC: A general Bayesian framework for attributed graph clustering [J]. ACM Transactions on Knowledge Discovery from Data, 2014, 9(1): 5. 1-5. 43.

[15] GAO H C, HUANG H. Deep attributed network embedding [C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. AAAI Press, 2018: 3364-3370.

[16] LIAO L Z, HE X N, ZHANG H W, et al. Attributed social network embedding [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(12): 2257-2270.

[17] WANG C, PAN S R, LONG G D, et al. MGAE: Marginalized graph autoencoder for graph clustering [C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, 2017: 889-898.

[18] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks [J]. Science, 2014, 344(6191): 1492-1496.

[19] LUXBURG U V. A tutorial on spectral clustering [J]. Statistics and Computing, 2004, 17(4): 395-416.

[20] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine [J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117.

[21] BONACICH P. Power and centrality: A family of measures [J]. American Journal of Sociology, 1987, 92(5): 1170-1182.

[22] YU S X, SHI J B. Multiclass spectral clustering [C]// Proceedings of the 9th IEEE International Conference on Computer Vison. IEEE, 2003, 2: 313-319.

[23] ELHADI H, AGAM G. Structure and attributes community detection: Comparative analysis of composite, ensemble and selection methods [C]// Proceedings of the 7th Workshop on Social Network Mining and Analysis. ACM, 2013, 10: 1-7.

[24] LANCICHINETTI A, FORTUNATO S, RADICCHI F. Benchmark graphs for testing community detection algorithms [J]. Physical Review E, 2008, 78(4): 046110.

[25] BERAHMAND K, HAGHANI S, ROSTAMI M, et al. A new attributed graph clustering by using label propagation in complex network [J]. Journal of King Saud University－Computer and Information Sciences, 2022, 34(5): 1869-1883.

[26] LIU L Y, XU L L, WANG Z, et al. Community detection based on structure and content: A content propagation perspective [C]// Proceedings of the 2015 IEEE International Conference on Data Mining. IEEE Computer Society, 2015: 271-280.

[27] ZHANG X T, LIU H, LI Q M, et al. Attributed graph clustering via adaptive graph convolution [C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. AAAI Press, 2019: 4327-4333.

ZHAO Xingwang, born in 1984, Ph.D, associate professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include data mining and machine learning.

（责任编辑：喻藜）