文章编号 1672-6634(2019)04-0072-09 **DOI** 10.19728/j.issn1672-6634.2019.04.012

一种基于网络嵌入的社区发现方法

王瑞国 叶雅玲 卜 湛

(南京财经大学 软件工程系,江苏 南京 210013)

摘 要 社区发现是社会网络分析的重要任务,有助于理解中观尺度的网络结构.现有的诸多社区发现方法仅考虑网络的拓扑信息,忽略了网络中每个节点所包含的属性信息.为此,本研究首先基于社会网络的拓扑结构信息与节点属性信息分别构建初始特征矩阵;然后基于网络嵌入模型,融合初始特征矩阵的主成分信息,构建共识嵌入矩阵;最后,给出社会网络中"领袖节点"的泛化定义形式,并提出一种改进的图聚类算法(LIK-means)挖掘社会网络中潜在的社区结构.实验表明,LIK-means 较其他经典算法有较好的可扩展性,同时在真实社会网络中的社区识别精度更高.

关键词 社区发现;网络嵌入;领袖节点;LIK-means 算法

中图分类号 N93 **文献标识码** A

0 引言

随着移动互联网的迅猛发展,人们的社交方式产生了日新月异的变化.以微信、微博为代表的新兴社交方式以井喷式的速度在人群中炸裂开来.从网络科学的研究视角出发,利用属性图模型对在线用户的交互行为进行语义建模,相较于传统的复杂网络模型,在属性图模型中,单个节点代表一位社交用户,节点附带的属性向量刻画社交用户的个体行为特征,节点间的边界则表示为社交用户间的友好关系.社会网络分析领域中一项热点研究议题是如何挖掘社交网络中的潜在社区,即在单个社区中,用户的行为偏好特征基本相似,同时用户间的交互关系比较紧密.挖掘社交网络中潜在社区不仅仅对理解分析社交网络的拓扑结构以及社交用户的偏好行为存在重要意义,更在针对社交平台优化广告投放、对社交用户进行个性化推荐等方面有着至关重要的作用.

最近的几十年间,越来越多的研究者致力于解决属性图潜在社区挖掘方面的研究课题,属性图聚类方法不失为一种解决本课题的方法,但仍存在着许多值得去探讨与解决的问题.首先,属性图中的拓扑结构与节点属性信息为两种独立异构数据形态,很难对两者进行有效地融合,并且此数据形态为不完全且有噪声的[1].虽然近期有学者们提出若干基于线性组合的方法[2,3],但对于加权参数(融合距离函数)的设置并不具有解释性;若干基于模型的方法[4,5]对于统计模型先验分布的选择上存在严重的依赖性,如何选择一个优秀的模型需要研究者更丰富的经验.其次,在传统的聚类算法中,质心的选择往往影响收敛的速度和效率,随机选取质心易导致糟糕的聚类结果且效率低下,并且属性图中离群点和孤立噪声点的存在将导致均值严重偏离,另外当属性图中数据量异常庞大时,传统的聚类算法时间开销很大,因此如何优先确定聚类算法中质心的位置依然是值得研究的问题.

为了降低属性图中拓扑结构与节点属性信息中存在的噪声问题,本文拟将采用网络嵌入的方法[6-8],分别将拓扑结构信息与节点属性信息映射到一个 m 维的欧式空间中,从而降低其数据中的噪声.其次,本文应当思考如下问题:如何最大限度地保留拓扑结构与节点属性信息并有效的将降噪处理后的信息进行融合?为了解决此问题,本文拟采用典型相关分析的方法最大化其关联程度,得到最终共识嵌入的表示形式.最后本文提出"领袖节点"定义,将"领袖节点"应用到 K-means 聚类算法的质心上,提高聚类的速度与效率,以期望挖掘出属性图中内部结构紧凑且节点属性相似的社区.

收稿日期:2018-10-10

基金项目:国家自然科学基金面上项目(71871109,71871233)资助

通讯作者:卜湛,男,汉族,博士,副教授,研究方向:社会计算、数据挖掘、博弈论, E-mail:buzhan@nuaa.edu.cn.

采用网络嵌入的方法，分别基于属性图中的拓扑结构信息和节点属性信息构建相应的 Laplacian 矩阵，进而根据谱分析原理将其分别映射到对应的 t 维嵌入空间，再基于典型相关分析理论最大化上述 t 维特征向量的相关性，形成最终的共识嵌入矩阵. 融合拓扑结构与节点属性信息后，本文给出“领袖节点”的泛化定义形式，通过计算节点的中心性指标，可识别网络当中的“领袖节点”，进而改进传统 K-means 聚类算法初始质心的选择，以提高社团发现结果的准确率.

为了有效挖掘属性图中隐含的社团结构，本研究采用网络嵌入的方法，融合属性图中节点的拓扑信息与属性信息形成共识嵌入矩阵，在此基础上挖掘属性图中的“领袖节点”，继而将这些“领袖节点”所对应的共识嵌入向量作为经典 K-means 算法的初始质心. 时间复杂度分析表明本研究所提出的 LIK-means 算法较其他经典算法具有更好的可扩展性（见表 2 所示）. 同时在八个真实的属性图数据集上的仿真实验表明，LIK-means 算法能够更准确的发现真实属性图中的社团结构，见图 2 所示.

1　相关工作

1.1　复杂网络潜在社区挖掘方法

在复杂网络分析中，被称为社区检测的传统的图聚类方法受限于简单图，早期的方法主要是基于度量过程中理想化的图生成模型. 因此，简单图中的社区检测问题可以转化为一个基于特定目标函数的全局优化问题. 模块度优化[9]是最著名的指标之一，可以用其衡量社区结构的优劣. 相比零模型中的期望值，社区内部边的密度越大，当前划分方式下的模块度得分就越高. 相比于理想化的图生成模型，李慧嘉等人[10]采用一种新颖的图生成模型，使得社团结构迭代快速探测算法可以在无参数的情况下方便高效的运行. 另外，不存在全局目标函数的社区检测方法通常采用自下而上的策略. 首先定义局部社区的某些特定属性，然后搜索图中具有预定属性的节点集合，最后通过融合局部结构来得到最终的全局社区结构. 例如，Clauset 等人[12]采用子图的边界节点来定义局部模块度，并提出了一个贪婪算法来对其进行优化. Lu 等人[13]将一个子图的内度与外度的比率作为一种度量指标. 由于以上两种度量指标包含较多的离群点，虽然可以实现较高的召回率，但其精确率却偏低. 另外，还有一些方法利用节点之间的相似性来度量局部社区质量，其中具有代表性的方法为基于多智能体自治计算的社区检测方法 AOCCM[14].

此外由于一些传统方法的不足，如模块度 Q 分辨率极限的问题，李慧嘉等人[15]结合 Potts 模型和 Markov 动态过程，通过计算每个节点的归属向量，识别网络中模糊的社区结构[15]. 除此之外，李慧嘉等人[16]利用迭代技术并引入一种新型的基于离散时间的动态系统，探索社团归属收敛的最优条件，并创新性的提出了划分指标函数的泛化形式，揭示了社团的层次结构与社团交互模式.

1.2　属性图聚类方法

相较于复杂网络潜在社区挖掘方法，属性图聚类方法不仅需要考虑属性图的拓扑结构信息，也要考虑节点属性信息. 大体可分为三种：(1) 基于距离的方法；(2) 基于模型的方法；(3) 基于子空间的方法.

基于距离的方法主要是设计一个距离/相似度度量方法，借此将节点的拓扑信息和属性信息相结合. 典型的距离函数可定义为：$d_{TA}(i,j)=\alpha d_T(i,j)+(1-\alpha)d_A(i,j)$. 其中 $d_T(i,j)$ 和 $d_A(i,j)$ 分别代表节点 i 和 j 之间的拓扑相似性和属性相似性，α 为权重因子，$0\leqslant\alpha\leqslant1$. 因此，基于距离的聚类方法可以应用于节点聚类，其中包括 SA-Cluster[15]和其相应的扩展 SA-Cluster-Opt[16]等.

基于模型的方法则基于特定的图生成模型，采用统一的方式融合属性与边，并将概率模型用于属性图的聚类过程. 当前最好的生成模型是 AGM[17]，其从当前图中学习到属性的相关性，并拓展现有的生成图模型用于更大样本上的推理预测. 类似的方法还有 CohsMix[18]、贝叶斯概率模型[19]、CESNA[20]和 MOEASA[21]等. 然而基于模型的方法却需要丰富的专家经验来选取统计模型的先验分布，且估计似然参数的时间复杂度通常较高.

基于子空间的方法致力于识别具有鉴别力的属性，继而得到良好划分的簇. 此类方法认为，随意使用所有的可得属性会导致糟糕的聚类结果. 因此，GAMer 方法[22]将密度、规模和簇的维度等相关质量属性考虑在内，通过权衡这些属性来实现网络中簇的检测. 此外，一个总体框架 EDCAR[23]也进一步被提出，继而可以更为有效地将子空间聚类和稠密子图挖掘联系起来. 尽管在已选子空间中的子图具有一定的实际意义，但其具有较高的计算成本.

1.3　基于网络嵌入的方法

近年来，网络嵌入吸引了许多研究者的关注. 其基本思想是基于不同的网络挖掘研究在嵌入的欧式空间

中将节点的信息最大限度的表现出来. Abbe 等人[24]采用随机块模型 SBM,设计新颖的算法,在准线性时间内将所有的社区恢复到最优阈值,根据结果的通用性处理重叠社区. Karyotis 等人[25]采用双曲线网络嵌入的方法修改了 GN 算法,其中节点距离通过嵌入的双曲线空间计算,并应用于感官数据聚类. 定义相对于的度量指标(HEBC)和修正 GN 的核心思想以便提高社区计算的速度. Keikha 等人[26]提出一种新颖的算法"CARE"适用于不同类型的网络(加权、有向),利用网络节点的局部邻居信息和社区信息代替网络社区中的局部和全局结构,并采用 Skip-gram 模型学习节点向量在空间的表示方法.

2 问题定义

首先定义本文中出现的符号,采用粗体大写字母表示矩阵(如 $\boldsymbol{A}$),粗体小写字母表示向量(如 a),普通小写字母表示标量(如 a),$\boldsymbol{A}'$代表$\boldsymbol{A}$ 的转置矩阵,1 表示为矩阵元素全为 1 的矩阵,$\boldsymbol{I}$ 表示为单位矩阵. 本文的符号说明见表 1.

社会网络可以建模为一个属性图,如 $g=(U,N,\boldsymbol{X})$,其中 $U=\{u_1,u_2,\cdots,u_n\}$表示 n 个节点的集合,$N=\{N_i:i\in\{1,2,\cdots,n\}\}$表示 n 个节点的邻居集合. $\boldsymbol{X}=[x_{ip}]\in\mathbf{R}^{n\times d}$, $\forall\, x_{ip}\in\{0,1\}$ 表示 n 个节点的属性矩阵,其中 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{id})\in\mathbf{R}^d$ 表示节点 i 的属性向量.

定义 1 属性图社团发现问题:旨在从 g 中寻找 $K(K\ll n)$ 个社团,即 $P=\{C_p\}_{p\in[1,K]}$,满足 $\bigcup_1^K C_p=U$, $\forall\, p,q,p\neq q,C_p\cap C_q=\varnothing$. 同时,理想的社团发现结果需满足以下两个基本条件:(1)同一社团内节点连接紧密,不同社团中节点连接稀疏;(2)同一社团内节点属性向量尽量相似,不同社团中节点属性向量相差较大.

表 1 符号说明

符号	定义
g	属性图
$\boldsymbol{A}$	属性图 g 的邻接矩阵
$\boldsymbol{X}$	属性图 g 的属性矩阵
$\boldsymbol{Y}$	属性图 g 的共识嵌入矩阵
$\boldsymbol{y}_i$	i 节点的共识嵌入特征向量
n	属性图 g 的节点个数
t	初始特征矩阵维度
m	属性图 g 的边界个数
d	属性向量维度
l	共识嵌入特征向量维度
K	社区数量
e_v	第 p 个社区
c_v	社区 p 的共识嵌入特征向量质心
P	社区结构
o	算法 1Part2 迭代次数

3 基于网络嵌入的社区发现方法

本文的总流程图如图 1 所示,例如给定一个有 7 个节点的属性图,每个节点附有 4 维属性信息,属性图的拓扑结构如图中所示,本文首先获取该属性图的信息构建节点的邻接矩阵 $\boldsymbol{A}$ 与属性矩阵 $\boldsymbol{X}$,其次根据网络嵌入模型得到共识嵌入矩阵$\boldsymbol{Y}$,然后通过"领袖节点"的泛化定义,选取出"领袖节点"(如图中的 1 节点与 4 节点),最终将"领袖节点"作为传统 K-means 算法的初始化质心节点,迭代聚类得到最终的社区结构.

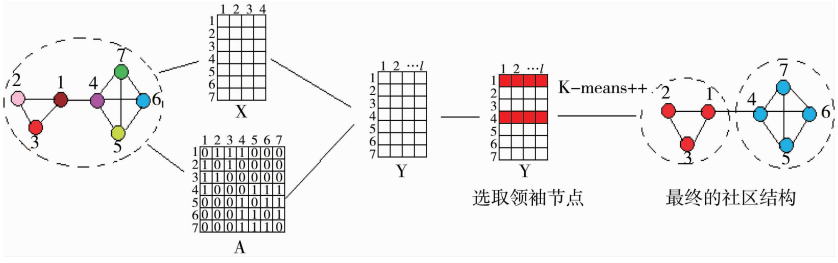

图 1 总流程图框架

3.1 网络嵌入模型

由于网络拓扑结构信息与节点属性信息为独立的异构数据表示形态,数据信息并不完整且含有噪声. 因此如何在最大限度保留节点拓扑信息与属性信息的同时对其进行降噪处理成为本文的一个难点,本文将采用网络嵌入模型(见图 1 中的网络嵌入框架)对属性图信息进行降噪处理,具体处理方法如下.

给定一个属性图网络 $g=(U,N,\boldsymbol{X})$,$\boldsymbol{A}\in\mathbf{R}^{n\times n}$表示为属性图的邻接矩阵,$\boldsymbol{D}_{\boldsymbol{A}}(i,i)=\sum_{j=1}^{n}\boldsymbol{A}(i,j)$表示为属性图节点的对角矩阵,$\boldsymbol{L}_{\boldsymbol{A}}=\boldsymbol{D}_{\boldsymbol{A}}-\boldsymbol{A}$ 表示为对应节点拓扑信息的 Laplacian 矩阵. 在一个网络嵌入中,根据谱定理将属性图网络中的节点映射到一个 t 维的嵌入空间 $\boldsymbol{y}_i\in\mathbf{R}^t(t\ll n)$中,理想的空间嵌入 $\boldsymbol{Y}_{\boldsymbol{A}}=[y_1,y_2,\cdots,y_n]\in\mathbf{R}^{n\times t}$本质为 $\min\frac{1}{2}\parallel\boldsymbol{y}_i-\boldsymbol{y}_j\parallel_2^2$,上述设置可以确保相连的节点在嵌入的空间中彼此相近. 故此问题可

以转化为求解特征向量问题 $\boldsymbol{L}_A\boldsymbol{a}=\boldsymbol{\lambda}\boldsymbol{D}_A\boldsymbol{a}$. 令 $\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n$ 为对应特征值 $0=\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_n$ 的特征向量,其中当 $\lambda_1=0$ 时,对应的特征向量为 1(向量中的元素都为 1). 从 $\boldsymbol{a}_2$ 开始选取 top-t 个特征向量组成嵌入空间 $\boldsymbol{Y}_A=[\boldsymbol{a}_2,\cdots,\boldsymbol{a}_k,\boldsymbol{a}_t+1]$,$\boldsymbol{Y}_A\in\mathbf{R}^{n\times t}$代表网络拓扑信息的嵌入空间. 在文中的以下部分,为了描述简练,本文将 t 个特征值与特征向量作为 top-t 个特征值与特征向量. 节点属性信息处理过程类似于网络中拓扑结构处理过程,首先标准化每个节点的属性值,得到余弦相似度矩阵 $\boldsymbol{W}$,求解特征向量问题 $\boldsymbol{W}\boldsymbol{b}=\boldsymbol{\lambda}_b$,选取 top-$t$ 个特征向量组成嵌入空间 $\boldsymbol{Y}_X=[\boldsymbol{b}_2,\cdots,\boldsymbol{b}_k,\boldsymbol{b}_{t+1}]$.

由上述得到的嵌入空间 $\boldsymbol{Y}_A$ 与 $\boldsymbol{Y}_X$ 为独立存在形式,并不能确保融合达到共识嵌入的目的. 本文采用典型相关分析的方法期望达到嵌入空间 $\boldsymbol{Y}_A$ 与 $\boldsymbol{Y}_X$ 的最大化关联程度,合理的定义两个投影向量 $\boldsymbol{p}_A$ 与 $\boldsymbol{p}_x\in\mathbf{R}^{t\times1}$,$\boldsymbol{Y}_A$ 与 $\boldsymbol{Y}_X$ 通过投影之后可以最大化关联程度,等价于解决以下优化问题

$$\begin{aligned}&\max_{\boldsymbol{p}_A{}'\boldsymbol{p}_X}\boldsymbol{p}'_A\boldsymbol{Y}'_A\boldsymbol{Y}_A\boldsymbol{p}_A+\boldsymbol{p}_A\boldsymbol{Y}'_A\boldsymbol{Y}_X\boldsymbol{p}_x+\boldsymbol{p}\boldsymbol{X}'\boldsymbol{Y}_X{}'\boldsymbol{Y}_A\boldsymbol{p}_A+\boldsymbol{p}_X{}'\boldsymbol{Y}_X{}'\boldsymbol{Y}_A\boldsymbol{p}_X,\\&s.t.\ \ \boldsymbol{p}'_A\boldsymbol{Y}'_A\boldsymbol{Y}_A\boldsymbol{p}_A+\boldsymbol{p}\boldsymbol{X}'\boldsymbol{Y}_X{}'\boldsymbol{Y}_X\boldsymbol{p}_X=1.\end{aligned}\tag{1}$$

构建 Lagrange 方程求解上述优化问题,令 Lagrange 方程偏导为 0,求出$[\boldsymbol{p}_A,\boldsymbol{p}_X]$的最优解,等价于求解以下特征向量问题

$$\begin{bmatrix}\boldsymbol{Y}'_A\boldsymbol{Y}_A & \boldsymbol{Y}'_A\boldsymbol{Y}\\ \boldsymbol{Y}'_X\boldsymbol{Y}_A & \boldsymbol{Y}_X\boldsymbol{Y}_X\end{bmatrix}\begin{bmatrix}\boldsymbol{p}_A\\ \boldsymbol{p}_X\end{bmatrix}=\boldsymbol{\gamma}\begin{bmatrix}\boldsymbol{Y}'_A\boldsymbol{Y}_A & 0\\ 0 & \boldsymbol{Y}_X{}'\boldsymbol{Y}_X\end{bmatrix}\begin{bmatrix}\boldsymbol{p}_A\\ \boldsymbol{p}_X\end{bmatrix},\tag{2}$$

其中 $\boldsymbol{Y}'\boldsymbol{A}$ 为 $\boldsymbol{Y}_A$ 的转置矩阵,γ 为 Lagrange 方程因子,选取 top-e 个特征向量,组成投影矩阵 $\boldsymbol{p}\in\mathbf{R}^{2t\times l}$,最终得到共识嵌入为 $\boldsymbol{Y}=[\boldsymbol{Y}_A,\boldsymbol{Y}_X]\times\boldsymbol{P},\boldsymbol{Y}\in\mathbf{R}^{n\times1}$.

3.2 “领袖节点”定义

真实世界中通常都存在着“领袖人物”,其在局部领域有着较高的“声誉”等级并与其他相关角色相互影响,网络中亦是如此. 网络中同样存在着“领袖节点”, 随着“声誉”的传播,“声誉”值低的节点对“声誉”值高的节点具有更强的依附趋势,进而促进了社区的形成,因此本文认为挖掘属性图中的“领袖节点”不仅对分析社区结构以及分析属性图聚类过程都具有重要意义. 将属性网络构建为属性图,其中网络中的节点可以相互影响,同时本文将做出如下假设

假设 1　“领袖节点”比其它节点具有更高的局部中心性.

假设 2　两个节点的邻居集越相似,则两个节点的相互影响越大.

假设 3　两个节点属性差异越大,其相互影响就越小.

当节点的局部“声望”越大时,越有可能成为“领袖节点”,并基于上述假设,给出节点的“声誉”定义

$$R_i=\sum_{j\in N_i}g_{ij}=\sum_{j\in N_1}\lambda_{ij}^{\blacksquare}\cdot\mathrm{e}^{-\omega(y_i,y_j)},\tag{3}$$

其中 $g_{ij}=\lambda_{ij}^{\blacksquare}\cdot\mathrm{e}^{-\omega(y_i,y_j)}$ 为节点 i 到节点 j 的影响力,$\lambda_{ij}^{\blacksquare}$表示节点 i 与节点 j 的拓扑信息的相似性,可从以下基于邻居相似性公式中选取

$$\text{Jaccard 系数 }\lambda_{ij}^{JC}=\frac{|N_i\bigcap N_j|}{|N_i\bigcup N_j|};\text{Salton 指标 }\lambda_{ij}^{Sa}=\frac{|N_i\bigcap N_j|}{\sqrt{|N_i||N_j|}};$$

$$\text{Sorensen 指标 }\lambda_{ij}^{SO}=\frac{2|N_i\bigcap N_j|}{|N_i|+|N_j|};\text{Hub Promoted 指标 }\lambda_{ij}^{HP}=\frac{2|N_i\bigcap N_j|}{\min(|N_i|,|N_j|)};$$

$$\text{Hub Depressed 指标 }\lambda_{ij}^{HD}=\frac{2|N_i|\bigcap|N_j|}{\max(|N_i|,|N_j|)};\text{Leicht-Newman 指标 }\lambda_{ij}^{LN}=\frac{2|N_i\bigcap N_j|}{|N_i||N_j|}.$$

$\bar{\omega}(\boldsymbol{y}_i,\boldsymbol{y}_j)$为可定义为经典的布雷格曼散度(Bregman divergence)形式,$\bar{\omega}(\boldsymbol{y}_i,\boldsymbol{y}_j)$定义

$$\bar{\omega}(\boldsymbol{y}_i,\boldsymbol{y}_j)=\varphi(\boldsymbol{y}_i)-\varphi(\boldsymbol{y}_j)-<(\boldsymbol{y}_i-\boldsymbol{y}_j),\nabla\varphi(\boldsymbol{y}_j)>,\tag{4}$$

其中 $\varphi(\cdot):\mathbf{R}^d\to\mathbf{R}_0^+$ 为连续可微非负凸函数,$\nabla\varphi(\boldsymbol{y}_j)$为评价 φ 在 $\boldsymbol{y}_j$ 的梯度向量,$\boldsymbol{y}_i-\boldsymbol{y}_j$ 为向量 $\boldsymbol{y}_i$ 和 $\boldsymbol{y}_j$ 的欧式距离,$<(\boldsymbol{y}_i-\boldsymbol{y}_j),\nabla\varphi(\boldsymbol{y}_i)>$为$(\boldsymbol{y}_i-\boldsymbol{y}_j)$与$\nabla\varphi(\boldsymbol{y}_j)$的内积. 基于上述函数的定义,可以根据 $\varphi(\cdot)$不同的凸性质,定义不同的 $\bar{\omega}(\cdot,\cdot)$的函数形式

欧式距离当 $\varphi(\boldsymbol{y}_i)=\|\boldsymbol{y}_i\|^2,\bar{\omega}(\boldsymbol{y}_i,\boldsymbol{y}_j)=\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2$.

余弦距离当 $\varphi(\boldsymbol{y}_i)=\|\boldsymbol{y}_i\|^2,\bar{\omega}(\boldsymbol{y}_i,\boldsymbol{y}_j)=\|\boldsymbol{y}_i\|-\dfrac{<\boldsymbol{y}_i,\boldsymbol{y}_j>}{\|\boldsymbol{y}_j\|}$.

3.3 算法描述

通过上述定义的“领袖节点”,可以找出融合矩阵中的初始质心,近而通过 K-means 聚类方法对节点进行聚类. 本文提出一种基于领袖节点挖掘的属性图聚类算法 (LIK-means),伪代码见算法 1.

给定一个包含 n 个节点的共识嵌入空间 $\boldsymbol{Y}$,其中 $\boldsymbol{y}_i \in \boldsymbol{Y}, i \in \{1,2,\cdots,n\}$,在 Part1:初始化质心中,第 2 步首先用公式(3)计算每个节点的"声誉".第 3 步根据每个节点的"声誉"降序排列.第 5 步到第 17 步,所有节点迭代访问:如果节点 u_1 没有被访问到,将 $\boldsymbol{y}_i$ 初始化为质心,并将 u_i 的邻居节点标记为已访问,并开始下一轮迭代过程,直到所有节点已被标记为访问或当初始化质心数量达到 K 时,迭代终止.Part1 算法在保证质心数量的同时,利用"声誉"初始化簇类质心.在 Part2:K-means++中,第 20 到第 28 步,计算所有节点与质心的距离,选取最短距离的簇加入,并更新簇的质心,直到所有的簇中节点不改变时迭代结束,此时得到所有的社区结构 p.

算法 1 基于领袖节点挖掘的属性图聚类算法(LIK-means).

输入 Y, K

```
1 ===Part1:初始化 K 个质心===
2 ∀ u_i ∈ u calculate R_i using Eq.(3);
3 sort u_i according to R_i in descending order;
4 k=1
5 for each u_i ∈ u do
6    if u_i has not been visited then
7        mark u_i as visited;
8        c_k = y_i;
9        for each j ∈ N_i do
10           mark u_j as visited;
11       end for
12       k=k+1;
13       if k>K then
14           break;
15       end if
16   end if
17 end for.
18 ===Part2:K-means++===
19 Repeat:
20   for each u_i ∈ u do:
21     for each k ∈ {1,2,…,K} do:
22       calculate distance between u_i and c_k;
23       k* = arg^min_{k∈(1,2,…,K)} w(y_i, c_k)
24       c_k*.add(u_i);
25       break;
26     end for
27 end for
28   update the c_k
29 Until ∀ c_k are not changed.
```

输出:p 社区结构.

3.4 算法时间复杂度分析

给定一个包含 n 个节点,m 条边,共识嵌入矩阵为 $Y \in R^{n\times 1}$ 的属性图,令 $\langle k\rangle = 2m/n$ 表示属性图的平均度,同时假设每个节点的邻居存储于有序的邻接链表.计算节点的"声誉"(Step 2)的时间复杂度为 $o(m\langle k\rangle + ml)$;初始化 K 个质心(Steps 3-17)需要的时间复杂度为 $O(n\log n + m)$;K-means++算法(Steps 19-29)的时间开销为 $O(nKlT)$,其中 T 表示算法的迭代次数.综上所述,LIK-means 算法的整体时间复杂度为 $O(m\langle k\rangle + ml + n\log n + m + nKlT)$.考虑真实网络中 $\log n \approx \langle k\rangle \approx K$,因此算法的时间复杂度可简化为 O(mlT).与其他经典社区发现算法的时间复杂度相比(如表 2 所示),LIK-means 算法有较好的可扩展性.

表 2 经典算法时间复杂度分析

算法	复杂度
CNM[35]	$O(n\log^2 n)$
DA[38]	$O(n^2\log n)$
Louvain[27]	$O(m\log n)$
OCR-HK[39]	$O(n^2)$
Bayesian inference[40]	$O(n\log^2 n)$
Variational Bayian[41]	$O(n^{1.44})$
RN Potts model[42]	$O(m^{1.3})$
Label Propagation model[43]	$O(m+n)$

4 实验分析

我们将 LIK-means 算法应用到八个真实的属性图数据集,并和其他经典社区发现算法进行性能对比.所有实验部署于一台安装 Linux 操作系统的计算机上执行,其配置为:主频 2.6 GHZ 的 4 核 E5-2650v2 处理器,128 G 内存,600 G 的 AS 硬盘和 240 G 的固态硬盘.

4.1 实验数据集

本文将基于领袖节点的社区发现算法应用到真实数据集当中,Twitter[36] 和 Google+[36] 为全球访问量比较大的社交网站,本文选取的真实属性图数据集来自于 Twitter 和 Google+的子网络,其中节点代表网络中的社交用户,属性刻画用户的行为特征.本文选取的子网络的优先级是先比较子网络的连通性,其次根据

子网络中节点被标记个数占子网络的节点个数比重，最后根据子网络中出现的社区个数来确定子网络是否被选取. 数据集中前 4 个数据集来自于 Twitter 的子网络，后 4 个数据集来自于 Google+的子网络.

表 3　真实属性图数据集

Twitter	n	m	d	K^-	Google+	n	m	d	K^-
Twitter-1	220	8354	1170	4	Google+-1	102	1335	92	4
Twitter-2	145	285	168	10	Google+-2	1520	238471	1211	2
Twitter-3	133	4577	847	9	Google+-3	1889	582827	995	3
Twitter-4	144	4002	712	16	Google+-4	511	39128	268	5

4.2　基准方法和评价指标

本文的对比潜在社区挖掘算法包括：Louvain[27]是一种基于模块度优化的社区挖掘算法；Infomap[28]是一种基于随机游走模型和加权模块度优化的社区挖掘算法；SCD[29]是一种基于社区聚类系数优化的社区抽取方法；BigClam[30]是一种基于标签传播的重叠社区发现方法；Oslom[31]是一种基于局部扩展及优化的思想的社区发现算法；Metris[32]是一种基于谱图切割模型的经典图聚类算法；Walktrap[33]是一种基于随机游走的社区挖掘算法；Clauset-Newman-Moore[34]是一种自底向上进行，采用凝聚的方式进行社区发现的算法. 上述基准方法中，BigClam 和 Metris 需要指定社团划分的社团个数 K，实验中我们选择真实数据集中的社团个数作为这两个算法的输入参数，其他基准方法采用默认参数输入.

已知真实社区结构 $p^1=\{C_q^1\}_{q\in\{1,2,\cdots,K\}}$，本文采用 AvgF1[35]评价指标对不同算法发现的社区结构 $p=\{C_p\}_{p\in\{1,2,\cdots,K\}}$ 进行评估

$$\mathrm{AvgF1}(pp^*)=\frac{1}{2k}\sum_{p=1}^{k}\max_{p\in[1,K^*]}F1(C_p^*,C_p^*)+\frac{1}{2k^*}\sum_{q=1}^{k^*}\max_{p\in[1,K^*]}F1(C_p^*,C_p^*). \tag{5}$$

令 $P(C_p,C_q^*=\frac{|C_p\bigcap C_q^*|}{|C_p|})$ 为精确率，$R(C_p,C_q^*)=\frac{|C_p\bigcap C_q^*|}{|C_q^*|}$ 为召回率，$F1(C_p,C_q^*)=\frac{2P(C_p,C_q^*)R(C_p,C_q^*)}{P(C_p,C_q^*)R(C_p,C_q^*)}$ 表示识别社区 C_p 和真实社区之间 C_q^* 之间的 $F1$ 指标. 上述指标越大，表明挖掘潜在的社区结构同真实的社区结构匹配程度越高，挖掘的潜在社区结构中社区内部节点越紧凑且节点间的属性越相似，社区结构越好.

4.3　实验结果

本文首先进行网络嵌入共识嵌入维度 l 参数的分析，实验中，初始化的社区数量设定为相应数据集中真实的社区数量，并固定“领袖节点”中“声誉”计算的节点邻居相似性函数以及 Bregman divergence 中的函数（本实验分别选取 Jaccard 系数以及欧式距离函数），本文分别选取网络嵌入共识嵌入维度为 5 维、10 维、20 维、30 维、40 维、50 维，表 3 分别表示在不同维度下不同数据集中对应的 AvgF1 指标，从表 3 中可以观察到当选取共识维度为 40 维时，除个别数据集外，在大多数数据集中比其他维度都表现出较为良好的 AvgF1 指标.

表 4　不同维度下不同数据集的 AvgF1 指标

网络	$l=5$	$l=10$	$l=20$	$l=30$	$l=40$	$l=50$
Twitter-1	0.178	0.182	0.175	0.157	0.124	0.115
Twitter-2	0.231	0.21	0.233	0.238	0.223	0.231
Twitter-3	0.215	0.212	0.196	0.232	0.341	0.259
Twitter-4	0.259	0.27	0.253	0.226	0.395	0.304
Google+-1	0.236	0.329	0.389	0.307	0.453	0.442
Google+-2	0.295	0.293	0.295	0.332	0.233	0.223
Google+-3	0.257	0.265	0.266	0.257	0.267	0.267
Google+-4	0.067	0.082	0.068	0.064	0.063	0.065

其次对“领袖节点”中“声誉”的节点邻居相似性函数的选择进行分析，实验中，初始化的社区数量设定为相应数据集中真实的社区数量，网络嵌入的共识嵌入维度选择 40 维，分别选取“声誉”中不同的节点邻居相似性函数，如 Jaccard 系数、Salton 指标、Sorensen 指标、Hub Promoted 指标、Hub Depressed 指标、Leicht-Newman 指标，分析不同节点邻居相似性函数对实验结果的影响，选取表现最佳的指标，具体实验结果如表 5 所示，从表 5 中不难得出在大部分数据集下 Jaccard 系数比其他系数指标表现优异.

表 5 不同节点邻居相似性函数下不同数据集的 AvgF1 指标

网络	Jaccard	Salton	Sorensen	Hub P	Hub D	Leicht-N
Twitter-1	0.124	0.118	0.104	0.109	0.134	0.109
Twitter-2	0.223	0.156	0.22	0.158	0.221	0.156
Twitter-3	0.341	0.262	0.284	0.262	0.201	0.272
Twitter-4	0.395	0.245	0.256	0.211	0.295	0.256
Google+-1	0.453	0.26	0.315	0.192	0.315	0.3
Google+-2	0.233	0.241	0.273	0.236	0.247	0.224
Google+-3	0.267	0.186	0.257	0.176	0.257	0.19
Google+-4	0.063	0.042	0.066	0.047	0.064	0.043

本文进一步对“领袖节点”中“声誉”的 Bregman divergence 的选择进行分析,实验中,初始化的社区数量设定为相应数据集中真实的社区数量,网络嵌入共识嵌入维度选择 40 维,节点邻居相似性函数选择 Jaccard 系数,分析 Bregman divergence 的欧式距离函数以及余弦距离函数对实验的影响,选取表现最佳的指标,具体实验结果如表 6 所示,从表 6 中可以观察到在其它实验参数固定时在大多数的数据集中选取欧氏距离函数比选取余弦距离所得到的实验结果要更好.

表 6 不同 Bregman divergence 的函数形式在不同数据集下 AvgF1 指标

Twitter	欧式距离	余弦距离	Google	欧式距离	余弦距离
Twitter-1	0.124	0.119	Google+-1	0.453	0.19
Twitter-2	0.223	0.171	Google+-2	0.233	0.241
Twitter-3	0.341	0.189	Google+-3	0.267	0.186
Twitter-4	0.395	0.196	Google+-4	0.063	0.042

最终选取上述实验的最优参数与基准方法做对比,为公平起见,每个算法在每个数据集中分别独立执行 30 次,图 2 分别记录了在数据集 1 到数据集 8 中不同方法最优社区划分结果的 AvgF1 指标.实验设定选取网络嵌入共识嵌入维度为 40 维,节点邻居相似性函数选取 Jaccard 系数,Bregman divergence 选取欧式距离函数.从图 2 中观察 AvgF1 指标,可以较为明显地观察到 LIK-means 在 Twitter1-2 优于 BigClam 算法以及 Walktrap 算法;在 Twitter-4 数据集中 LIK-means 仅逊色于 Oslom 算法以及 Clauset-Newman 算法;在 Google+-1 数据集中表现最好是算法依次为 Oslom、Clauset-Newman 和 LIK-means;在 Google+-2 数据集中 LIK-means 算法要优于其他算法.

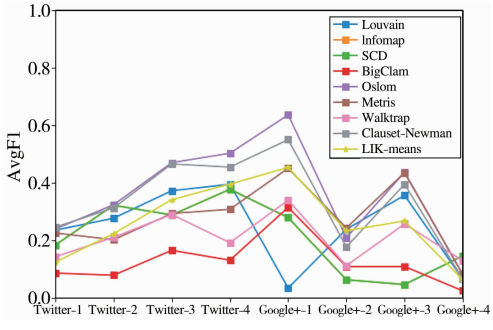

图 2 不同算法在不同数据集上的 AvgF1 值

5 结论

本文提出一种基于“领袖节点”挖掘的社区聚类算法用于挖掘属性网络中的潜在社区.为了简化研究问题,本文使用属性图对其进行建模,因此属性网络中的潜在社区挖掘问题转换为属性图聚类问题,将此属性图聚类问题分解为两个子问题:即属性图中节点拓扑信息与属性信息的融合以及挖掘“领袖节点”并初始化为 K-means 聚类的质心.为了解决这两个子任务,本文基于网络嵌入的思想,对属性图中的拓扑结构信息与节点属性信息进行降噪处理过后,使用典型相关分析的方法最大化嵌入信息的相关性,达到属性图中拓扑结构信息与属性信息融合的目的.合理的定义“领袖节点”,将“领袖节点”应用到 K-means 聚类的质心上.本研究提出的属性图社区发现算法(LIK-means)的时间复杂度为 $O(mlT)$,其中 m 表示属性图的边数,l 表示共识嵌入矩阵的特征维度,T 表示算法的迭代次数.对算法的时间复杂度分析显示 LIK-means 的执行时间同网络的边数 m 呈线性关系,这表明 LIK-means 更适用于稀疏网络.在真实属性图上的仿真实验表明,LIK-means 较其他基准方法能够更准确的发现属性图中隐含的社区结构,同时随着共识嵌入矩阵的特征维度 l 的增加,算法的识别精度也会得到相应提升.

参 考 文 献

[1] Lada A Adamic, Bernardo A Huberman. Power-law distribution of the world wide web[J]. Science, 2000, 287: 2115-2115.

[2] Zhou Y, Cheng H, Yu J X. Graph clustering based on structural/attribute similarities[J]. Proc VLDB Endowment, 2009, 2(1): 718-729.

[3] Cheng H, Zhou Y, Yu J X. Clustering large attributed graphs: a balance between structural and attribute similarities[J]. ACM Transactions on Knowledge Discovery from Data, 2011, 5(2): 190-205.

[4] Xu Z, Ke Y, Wang Y, et al. A model-based approach to attributed graph clustering[C]. //Proceedings of the ACM SIGMOD International Conference on Management of Data, Scottsdale, 2012.

[5] Yang J, Mcauley J, Leskovec J. Community detection in networks with node attributes[C]. //Proceedings of the IEEE International Conference on Data Mining, Shenzhen, 2014.

[6] Chen Mo, Yang Qiong, Tang Xiaoou. Directed graph embedding[C]. //IJCAI, 2007.

[7] Bryan Perozzi, Rami Al-Rfou, Steven Skiena. Deepwalk: online learning of social representations[J]. In KDD, 2014, 1: 701-710.

[8] Meng Jiantang , Wang Mingzhe, Ming Zhange, et al. LINE: large-scale information network embedding[C]. //In WWW, 2015.

[9] Newman M E, Girvan M. Finding and evaluating community structure in networks[J]. Phys Rev E Stat Nonlin Soft Matter Phys, 2004, 69(2): 026113.

[10] 李慧嘉，李爱华，李慧颖. 社团结构迭代快速探测算法[J]. 计算机学报，2017，40(4)：56-72.

[11] A Clauset. Finding local community structure in networks[J]. Physical Review E, 2005, 72(2): 026132.

[12] Luo F, Wang J Z, Promislow E. Exploring local community structures in large networks[J]. Web Intelligence & Agent Systems, 2006, 6(4): 387-400.

[13] Bu Z, Wu Z, Cao J, et al. Local community mining on distributed and dynamic networks from a multiagent perspective[J]. IEEE Transactions on Cybernetics, 2016, 46(4): 986-999.

[14] 李慧嘉，李慧颖，李爱华. 多尺度的社团结构稳定性分析[J]. 计算机学报，2015，38(2)：301-312.

[15] Zhou Y, Cheng H, Yu J X. Graph clustering based on structural/attribute similarities[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 718-729.

[16] 李慧嘉，严冠，刘志东，等. 基于动态系统的网络社团线性探测算法[J]. 中国科学：数学，2017，47(2)：241-256.

[17] Cheng H, Zhou Y, Yu J X. Clustering large attributed graphs: a balance between structural and attribute similarities[J]. ACM Transactions on Knowledge Discovery from Data, 2011, 5(2): 190-205.

[18] Iii J J P, Moreno S, Fond T L, et al. Attributed graph models: modeling network structure with correlated attributes[C]. //Proceedings of the 23rd International Conference on World Wide Web, Seoul, Republic of Korea, 2014.

[19] Zanghi H, Volant S, Ambroise C. Clustering based on random graph model embedding vertex features[J]. Pattern Recognition Letters, 2010, 31(9): 830-836.

[20] Xu Z, Ke Y, 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, Scottsdale, AZ, USA, 2012.

[21] Yang J, Mcauley J, Leskovec J. Community detection in networks with node attributes[C]. //IEEE 13th International Conference on Data Mining, Dallas, Texas, USA, 2013.

[22] Li Z, Liu J, Wu K. A multiobjective evolutionary algorithm based on structural and attribute similarities for community detection inattributed networks[J]. IEEE Transactions on Cybernetics, 2017, 10: 2720180.

[23] Gunnemann S, Farber I, Boden B, et al. Subspace clusteringmeets dense subgraph mining: a synthesis of two paradigms[C]. //IEEE 10th International Conference on Data Mining, Sydney, Australia, 2010.

[24] Gunnemann S, Boden B, Farber I, et al. Efficient mining of combined subspace and subgraph clusters in graphs with feature vectors[C]. //The 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Gold Coast, Australia, 2013.

[25] Abbe E, Sandon C. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms[J]. Foundations of Computer Science, 2015(3): 670-688.

[26] Karyotis V, Tsitseklis K, Sotiropoulos K, et al. Big data clustering via community detection and hyperbolic network embedding in IoT applications[J]. Sensors, 2018, 18(4): 1205.

[27] Keikha M M, Rahgozar M, Asadpour M. Community aware random walk for network embedding[J]. Knowledge-Based Systems, 2018, 23: 86-92.

[28] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. J Statist Mech, 2008, 10: 155-168.

[29] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure[J]. Proceedings of the National Academy of Sciences, 2008, 105(4): 1118-1123.

[30] Prat-Perez A,Dominguez-Sal D,Larriba-Pey J L. High quality,scalable and parallel community detection for large real graphs[C]. //Proceedings of the International Conference on World Wide Web,Seoul,2014.

[31] Yang J,Leskovec J. Overlapping community detection at scale:a nonnegative matrix factorization approach[C]. //Proceedings of the 6th ACM International Conference on Web Search and Data Mining,Rome,2013.

[32] Andrea L,Filippo R,Ramasco J J,et al. Finding statistically significant communities in networks[J]. PLoS ONE,2011,6(4):0018961.

[33] Karypis G,Kumar V. Multilevel k-way hypergraph partitioning[J]. Proceedings of the IEEE Conference on Design Automation,1999,11(3):285-300.

[34] Pons P,Latapy M. Computing communities in large networks using random walks[J]. Computer and Information Sciences,2005,3733:284-293.

[35] Clauset A,Newman,Moore M. Finding community structure in very large networks[J]. Phys Rev E70,2004,1:066111.

[36] Yang J,Leskovec J. Overlapping community detection at scale:a nonnegative matrix factorization approach[C]. //Proceedings of the 6th ACM International Conference on Web Search and Data Mining,2013.

[37] Mcauley J J, Leskovec J. Learning to discover social circles in ego networks[J]. Advances in Neural Information Processing Systems,2012,1:539-547.

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

[39] Boccaletti S,Ivanchenko M,Latora V,et al. Detecting complex network modularity by dynamical clustering[J]. Phys Rev E,2007,75(4):045102.

[40] Hastings M B. Community detection as an inference problem[J]. Phys Rev E,2006,74(3):035102.

[41]Hofman J M,Wiggins C H. Bayesian approach to network modularity[J]. Phys Rev Lett,2008,100(25):258701.

[42] Ronhovde P,Nussinov Z. Local resolution-limit-free potts model for community detection[J]. Phys Rev E,2010,81(4):046114.

[43] Raghavan U N,Albert R,Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Phys Rev E,2007,76(3):036106.

A Community Detection Approach Based on Network Embedding

WANG Rui-guo YE Ya-ling BU Zhan

(Department of Software Engineering,Nanjing University of Finances and Economics,Nanjing 210013,China)

Abstract In social network,the network topology information and the node attribute information exist in heterogeneous forms simultaneously. And how to effectively integrate the heterogeneous information for community detection has become a hot research topic in the field of social network analysis. Firstly,we use the topological structure information of social network and the node attribute information to construct the initial feature matrix respectively;Secondly,we build the consensus embedding matrix based on the network embedding model and the principal component information of the initial feature matrix;Finally,we provide the general definition of the "leader node" and propose an improved K-means algorithm (LIK-means)to mine the potential communities in the social networks. In the experiments,we choose eight classic community detection algorithms as the benchmark methods and verify the effectiveness of the LIK-means algorithm on real social network datasets.

Key words community detection;network embedding;leader node;LIK-means