计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190400160

一种基于粗糙集和密度峰值的重叠社区发现方法

张 琴 陈红梅 封云飞
西南交通大学信息科学与技术学院 成都 611756
西南交通大学云计算与智能技术高校重点实验室 成都 611756
(qinzhang@my.swjtu.edu.cn)

摘 要 随着互联网和社会的发展,各个领域每天都会产生大量相互关联、彼此依赖的数据,这些数据根据不同的主题形成了各种复杂网络。挖掘社区结构是复杂网络领域中的一项重要研究内容,因为其在推荐系统、行为预测和信息传播等方面具有极其重要的意义。社区结构中的重叠社区结构在生活中普遍存在,更具有实际研究意义。为有效发现复杂网络中的重叠社区,文中引入了粗糙集理论对社区进行分析,识别出重叠节点,进而提出了一种基于粗糙集和密度峰值的重叠社区发现方法 OCDRD(Overlapping Community Detection Algorithm Based on Rough Sets and Density Peaks)。该方法在传统网络节点局部相似性度量的基础上,结合灰色关联分析方法求出网络节点间的全局相似性,进而将其转化为节点间距离。将密度峰值聚类算法的思想应用于该算法中,以根据网络结构自动选取社区中心节点。依据网络中节点的距离比例关系,定义了社区的上近似、下近似以及边界域。最后,不断调整距离比率阈值并进行划分迭代,在每次迭代中针对社区的边界域进行计算,从而获得最佳重叠社区划分结构。在 LFR 基准人工网络数据集和真实网络数据集上,基于标准互信息(Normalized Mutual Information,NMI)和具有重叠性模块度 EQ 这两个评价指标,将 OCDRD 方法与近几年效果较好的其他社区发现算法进行测试比较。实验结果显示,OCDRD 方法在社区划分结构方面整体优于其他社区发现算法,表明了该算法的可行性和有效性。

关键词:重叠社区发现;粗糙集;密度峰值;灰色关联分析方法
中图法分类号 TP391

Overlapping Community Detection Method Based on Rough Sets and Density Peaks

ZHANG Qin,CHEN Hong-mei and FENG Yun-fei
School of Information Science and Technology,Southwest Jiaotong University,Chengdu 611756,China
Key Laboratory of Cloud Computing and Intelligent Technology,Southwest Jiaotong University,Chengdu 611756,China

Abstract With the development of the Internet and society,a large number of interrelated and interdependent data is produced in various fields every day,which form various complex networks according to different themes. Mining community structure of complex networks is an important research content,which has extremely important significance in recommendation system,behavior prediction and information spreading. Moreover,overlapping community structure of complex networks exists universally in life,which has practical research significance. In order to detect overlapping communities effectively in complex networks,an overlapping community detection method OCDRD based on rough sets and density peaks is proposed in this paper,in which rough set theory is used to analyze communities and identify overlapping nodes. Firstly,the global similarities among network nodes are obtained by using grey correlation analysis method based on the traditional local similarity measure of network nodes. Then the global similarities among network nodes are converted to distance among nodes. The center nodes of the community are automatically selected by the network structure by applying the idea of density peaks based clustering. Next,the lower approximation,the upper approximation,and the boundary region of the community are defined according to the distance ratio relation among nodes in the network. Finally,the threshold value of distance ratio is adjusted iteratively,and the boundary region of the community is calculated repeatedly in each iteration until the optimal overlapping community structure is obtained. The OCDRD algorithm is compared with other community detection algorithms that have achieved good results in recent years both on LFR benchmark artificial network datasets and real network datasets. By analyzing two common community detection evaluation indexes,NMI and overlapping module degree EQ,the experimental results show that OCDRD algorithm is superior to other community detection algorithms in community partition structure and it is feasible and effective.

Keywords Overlapping community detection,Rough set,Density peaks,Grey correlation analysis method

到稿日期:2019-04-18 返修日期:2019-08-13 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61572406)
This work was supported by the National Natural Science Foundation of China (61572406).
通信作者:陈红梅(hmchen@swjtu.edu.cn)

1　引言

随着互联网的发展，实际生活中每时每刻都会产生大量相互关联、彼此依赖的数据，这些数据构成了复杂系统。复杂网络是复杂系统的一种抽象表示，例如社交网络、蛋白质网络和交通网络等。社区是由相似程度高或连接紧密的个体所构成的集合，社区结构一般表现为属于同一个社区的节点连接紧密而不同社区间的节点连接稀疏[1]。社区是复杂网络的重要结构，能够帮助我们有效分析复杂网络。社区发现是根据网络中节点间联系的紧密程度来划分社区的过程，是复杂网络的重要研究内容之一。如果一个节点可同时划分到多个社区，则该节点为重叠节点，划分的社区结构称为重叠社区[2]。探测重叠社区结构具有实际意义，例如对于社交网络，用户可以根据兴趣爱好被划分到多个社区[1]；对于蛋白质网络，蛋白质可以根据功能作用被划分到多个社区[3]。因此，重叠社区发现算法具有重要的研究意义。

近年来，重叠社区发现算法的研究得到了广泛的关注和较好的发展。重叠社区发现算法包括派系过滤方法、基于边分割方法、局部扩展方法、模糊检测方法和动态方法[1]。为了处理大规模网络，Zhang 等提出了一种计算弱派系间相似性的方法，该方法可将弱派系融合为一个社区[4]。Sun 等提出了结合边的连接情况和标签传播算法划分重叠结构，然后再将边结构转化为点结构的边分割方法[5]。Yu 等提出了一种结合随机游走和种子扩展的重叠社区发现方法，先使用随机游走的策略来发现连接紧密的种子社区，然后根据节点到社区的相似性来扩展社区[6]。Wu 等提出了一种基于粗糙模糊聚类的社区发现算法，用于将蛋白质之间的模糊关系转化为模糊等价关系，最后探测重叠部分[3]。Bansal 等提出了一种快速社区检测算法，用于对动态网络数据进行实时划分[7]。

但上述重叠社区发现算法都未能很好地刻画出社区的重叠区域，因为上述算法一般是根据节点或者边的局部信息来扩展社区，不能很好地从全局把握重叠节点。社区的重叠区域由重叠节点构成，重叠节点在现实网络中普遍存在，因此对重叠社区划分进行研究具有重要的现实意义。粗糙集理论是由 Pawlak 教授于 1982 年提出的一种处理不确定性信息的软计算方法[8]，能够很好地应用于不确定性信息分析处理领域[9]，因此可以用粗糙集模型来划分重叠社区[10]。Zhang 等提出了一种基于粗糙集的描述社区模糊性的方法，其利用度中心性来确定中心节点，并结合 K-means 算法，进而划分得到含有模糊区域的社区结果[11]；但是该算法需要人工设置社区数，这对算法性能有较大影响。Zhu 等提出了一种粗糙聚类的方法，但是核心节点不能很好地被选择[12]。

密度峰值是近年来被提出的一种能够依据数据分布自动选取中心节点的聚类算法[13]。本文提出了一种基于粗糙集和密度峰值的重叠社区发现算法 OCDRD。OCDRD 算法的主要思想为：在传统局部节点间的相似性度量基础上，结合灰色关联分析[14]方法计算出节点间的全局相似性，由全局相似性定义节点间的距离。根据密度峰值的思想，由节点的局部密度和距离得到社区的数量及社区的中心节点最后依据距离的比率关系定义社区的近似集和边界域，针对社区边界域，通过调整距离比例阈值进行迭代计算，以获取最佳的重叠社区划分结构。在真实网络数据集和人工网络数据集上对所提算法进行对比实验，实验结果表明了该算法的可行性和有效性。

2　相关工作及定义

2.1　社区发现的含义

复杂网络可抽象为图的形式，本文仅考虑无向无权的网络图。图可以表示为 $G=(V,E)$，其中 $V=\{v_1,v_2,\cdots,v_n\}$ 表示网络中的节点集合，$E=\{e_1,e_2,\cdots,e_n\}$ 表示网络中的边集合。

社区是一个包含节点的子图，同一个社区内节点连接紧密，社区与社区间连接稀疏。社区结构在复杂网络中普遍存在[15]。挖掘复杂网络中社区结构的过程被称为社区发现。如果网络中存在某些节点能够同时属于多个社区，则为重叠社区发现。有效地划分重叠节点，是重叠社区发现的难点之一。

2.2　邻居节点和局部结构相似性

社区划分的主要依据是节点间的相互关系，相互关系可以通过共同邻居确定。邻居节点 $N(v_i)$ 和通过邻居节点确定的局部结构相似性 $sim(v_i,v_j)$ 定义如下。

定义 1（邻居节点）　已知复杂网络 $G=(V,E)$，$v_i\in V$，则节点 v_i 的邻居节点 $N(v_i)$ 定义为[16]：

$$N(v_i)=\{v_j\mid v_j\in V(i\neq j)\wedge(v_i,v_j)\in E\}\bigcup\{v_i\}\tag{1}$$

由节点间的邻接关系，可定义邻接矩阵 $\boldsymbol{AdjacentMatrix}=(a_{ij})_{n\times n}$，其中 $a_{ij}=\begin{cases}1, & v_j\in N(v_i)\\ 0, & \text{otherwise}\end{cases}$。

定义 2（局部结构相似性）　已知复杂网络 $G=(V,E)$，$\forall v_i,v_j\in V(i\neq j)$，则 v_i,v_j 间的局部结构相似性 $sim(v_i,v_j)$ 定义为[17]：

$$sim(v_i,v_j)=\frac{|N(v_i)\bigcap N(v_j)|}{\sqrt{|N(v_i)|\cdot|N(v_j)|}}\tag{2}$$

由式(2)可知，两个节点的共同邻居越多，其相似性 $sim(v_i,v_j)$ 越大，它们有很大的概率被划分到同一个社区。$sim(v_i,v_j)=0$ 时，表示两节点的局部结构完全不同；$sim(v_i,v_j)=1$ 时，表示两节点的局部结构完全相同；$sim(v_i,v_j)\in(0,1)$时，可以描述两节点的局部结构相似性。

根据节点间的局部结构相似性，可定义网络中所有节点之间的相似性矩阵 $\boldsymbol{SimMatrix}=(r_{ij})_{n\times n}$，其中 $r_{ij}=sim(v_i,v_j)$。局部结构相似性仅考虑了节点的邻居节点，没有考虑网络的全局结构，因此本文引入灰色关联分析来衡量节点间的全局结构相似性。

2.3　灰色关联分析方法的基本概念

本文采用灰色关联分析方法，来分析社区节点之间的全局相似性。灰色关联分析方法是灰色系统理论中的一个重要研究内容，其主要思想是通过观察曲线的几何形状来度量序列间联系的紧密程度[18]。灰色理论通过灰色关联度来度量两个数列的相互关系。

定义 3（灰色关联系数）　设有长度为 n 的数列 $\boldsymbol{X}=\{\vec{x}_i\mid\vec{x}_i=(x_i(1),x_i(2),\cdots,x_i(n)),i=0,1,\cdots,m\}$，其中 $\vec{x}_0$ 为参考数列，其余为比较数列。给定一个分辨系数 $\rho\in[0,1]$，则 $\vec{x}_i$ 与 $\vec{x}_0$ 的灰色关联度可表示为[19]：

$$\gamma(\vec{x}_i,\vec{x}_0)=\frac{1}{n}\sum_{j=1}^{n}\xi(x_0(j),x_i(j))\tag{3}$$

其中，$\xi(x_0(j),x_i(j))$被称为 $\vec{x}_i$ 与 $\vec{x}_0$ 的灰色关联系数，表示为：

$$\xi(x_0(j),x_i(j))=\frac{\min\limits_i\min\limits_j|x_0(j)-x_i(j)|+\rho\max\limits_i\max\limits_j|x_0(j)-x_i(j)|}{|x_0(j)-x_i(j)|+\rho\max\limits_i\max\limits_j|x_0(j)-x_i(j)|} \tag{4}$$

其中，$\xi(x_0(j),x_i(j))$满足 4 条灰色关系性质，即正则性、对称性、完整性和严密性。

2.4 粗糙集理论的基本概念

粗糙集是一种对不确定信息进行近似描述的数学范式，下面介绍其基本概念。

定义 4(决策信息系统) 已知论域 $U=(x_1,\cdots,x_i,\cdots,x_n)$为非空有限集合，$C$ 为条件属性，D 为决策属性，V 为属性的值域，$f:U\rightarrow V$ 为信息函数，$f(x_i,a_i)(a_i\in C\cup D)$表示对象 x_i 在属性 a_i 上的属性值，则 $DIS=(U,C\cup D,V,f)$表示决策信息系统。

定义 5(等价类) 已知决策信息系统 $DIS=(U,C\cup D,V,f)$，R 为 U 上的等价关系，$U/R=E_1,\cdots,E_i,\cdots,E_k$ 是由等价关系 R 形成的划分，E_i 称为等价类。

定义 6(下近似集/上近似集) 粗糙集中的下近似集表示确定属于某个任意集合，设 $\forall X\subseteq U$，则 X 的下近似集为 $\underline{R}(X)=\bigcup\{E_i\in U/R:E_i\subseteq X\}$；上近似集表示可能属于某个对象集合，设对象 $X\subseteq U$，则 X 的上近似集为 $\overline{R}(X)=\bigcup\{E_i\in U/R:E_i\cap X\neq\varnothing\}$。

3 基于粗糙集和密度峰值的重叠社区发现方法

重叠社区发现属于特殊聚类方法的一种，关键是要识别出聚类中心及个数，即社区中心节点；其次是根据节点间的相似性或者紧密性，能够很好地将节点进行社区划分，并且能准确地识别出社区的重叠节点。

为有效地度量节点之间的相似性以及社区距离，更合理地进行社区划分，本文运用灰色理论定义全局结构相似性的，运用密度峰值聚类算法确定聚类中心，并引入粗糙集理论对属于边界域的节点进行处理，以得到最佳的重叠社区结构。

3.1 社区近似集

粗糙集可以用于刻画不确定信息，因而可用于识别社区中的重叠节点。网络中社区的近似集定义如下。

定义 7(社区上近似集/下近似集) 已知网络 $G=(V,E)$，假设其社区划分为 $C=\{C_1,\cdots,C_l,\cdots,C_r\}$，则社区 C_l 的上近似集、下近似集分别定义为：

$$\overline{R^{\lambda}}(C_l)=\left\{v_i\ \middle|\ \frac{d(v_i,C_l)}{d(v_i,C_k)}\leqslant\lambda,\forall v_i\in V,\exists C_k\in C(k\neq l)\right\} \tag{5}$$

$$\underline{R^{\lambda}}(C_l)=\left\{v_i\ \middle|\ \forall C_k\in C(k\neq l),\frac{d(v_i,C_l)}{d(v_i,C_k)}\leqslant\lambda,\forall v_i\in V\right\} \tag{6}$$

其中，$0<\lambda\leqslant 1$ 为距离比例阈值，$d(v_i,C_i)$为节点 v_i 到社区 C_i 的距离。则边界域为：

$$Bnd(C_l)=\overline{R^{\lambda}}(C_l)-\underline{R^{\lambda}}(C_l) \tag{7}$$

$$Bnd(C)=\bigcup_{i=1}^{r}Bnd(C_i) \tag{8}$$

其中，$Bnd(C_l)$为社区 C_l 的边界域，$Bnd(C)$为所有社区划分 C 的边界域。下近似为社区的非重叠区域，社区的重叠区域为社区的边界域。

性质 1 已知网络 $G=(V,E)$，假设其社区划分为：$C=\{C_1,\cdots,C_l,\cdots,C_r\}$，$0<\lambda_1<\lambda_2\leqslant 1$，则以下性质成立：

(1) $\overline{R^{\lambda_1}}(C_l)\subseteq\overline{R^{\lambda_2}}(C_l)$；

(2) $\underline{R^{\lambda_1}}(C_l)\subseteq\underline{R^{\lambda_2}}(C_l)$。

通过调整阈值 λ，可得到多种重叠社区的划分结构。社区的划分过程是一个动态变化的过程。为有效衡量节点与社区之间的距离，需要定义动态的距离。

定义 8(节点与社区的距离) 在基于近似集的重叠社区发现中，节点与社区的距离可分为两种情况：当初始化社区 C_j 只有一个核心节点 v_j 时，节点 v_i 到社区 C_j 的距离可由距离矩阵 ***DistanceMatrix*** 求得；当社区发展为近似集社区时，由社区的下近似集和上近似集的不同重要性确定了不同的权重 α 和 β。因此，节点 v_i 与社区 C_j 的距离定义为：

$$d(v_i,C_j)=\begin{cases}d_{ij}, & 若|C_j|=1\\ \alpha*d_{\min}(v_i,\underline{R^{\lambda}}(C_j))+\beta*d_{\min}(v_i,\overline{R^{\lambda}}(C_j)), & 其他\end{cases} \tag{9}$$

其中，$|C_j|$表示社区 C_j 的节点个数；α 和 β 分别是社区下近似集和上近似集的权重，满足条件 $\alpha+\beta=1$ 且 $\alpha>\beta$；$d_{\min}(v_i,\underline{R^{\lambda}}(C_j))$和 $d_{\min}(v_i,\overline{R^{\lambda}}(C_j))$分别表示社区 C_j 下近似集与节点 v_i 之间的距离的最小值和社区 C_j 上近似集与节点 v_i 的距离的最小值，可由距离矩阵 ***DistanceMatrix*** 得到。$d_{\min}(v_i,\underline{R^{\lambda}}(C_j))$和 $d_{\min}(v_i,\overline{R^{\lambda}}(C_j))$为每次迭代中的动态距离。

3.2 基于灰色理论的全局结构相似性的定义

在网络中，节点和其他节点的相似性构成了一个数列，可运用灰色理论度量不同节点的相似性数列之间的相似性，以更好地描述全局结构的相似性。

定义 9(全局结构相似性) 已知网络 $G=(V,E)$，$v_i,v_j\in V(i\neq j)$，令 $\vec{v_i}$ 和 $\vec{v_j}$ 分别表示节点 v_i 和 v_j 由局部结构相似性构成的数列，即 $\vec{v_i}=\{sim(v_i,v_1),\cdots,sim(v_i,v_l),\cdots,sim(v_i,v_n)\}$，$\vec{v_j}=\{sim(v_j,v_1),\cdots,sim(v_j,v_l),\cdots,sim(v_j,v_n)\}$。$\xi(sim(v_i,v_r),sim(v_j,v_r))$表示 $\vec{v_i}$ 和 $\vec{v_j}$ 在数列第 r 维上的灰色关联系数，则节点 v_i 和节点 v_j 的全局结构相似性定义为：

$$S(\vec{v_i},\vec{v_j})=\frac{1}{n}\sum_{r=1}^{n}\xi(sim(v_i,v_r),sim(v_j,v_r)) \tag{10}$$

则包含 n 个节点网络的全局结构相似性矩阵为：***MatrixGS*** $=(S(\vec{v_i},\vec{v_j}))_{n\times n}$。将 ***MatrixGS*** 中全局结构相似性的值进行归一化，得到全局关联度矩阵为 ***GreyMatrix*** $=(g_{ij})_{n\times n}$。令 $d_{ij}=\frac{1}{g_{ij}}$，则距离矩阵 ***DistanceMatrix*** $=(d_{ij})_{n\times n}$。

3.3 社区发现中的密度峰值聚类算法

密度峰值聚类算法是由 Rodriguez 等于 2014 年在 *Science* 提出的一种聚类算法，其主要思想是类簇中心点具有较大的局部密度，且类簇中心点之间的距离应尽可能大[13]。本文将密度峰值聚类算法应用在社区发现中，其相关定义如下。

定义 10(节点局部密度) 社区核心点的局部密度较高，且大于邻近节点的密度。通过高斯核，结合节点间的距离和截断距离计算节点局部密度[20]。令 ρ_i 表示每个节点 i 的局部密度，d_c 为截断距离，d_{ij} 表示节点 i 与节点 j 之间的距离，则节点局部密度表示为：

$$\rho_i=\sum_j\exp\left(-\frac{d_{ij}^2}{d_c^2}\right) \tag{11}$$

其中，$d_{ij}=\frac{1}{\gamma_{ij}}$，$\gamma_{ij}$ 为节点 i 与节点 j 的灰色关联度。

定义 11（节点最小距离）　社区核心点与较高密度节点间的距离相对较大，则节点的最小距离表示为：

$$\delta_i=\begin{cases}\max\limits_{j}(d_{ij}), & \text{if } \rho_i=\max\limits_{k}(\rho_k)\\ \min\limits_{j:\rho_j>\rho_i}(d_{ij}), & \text{otherwise}\end{cases} \tag{12}$$

由式(12)可知，如果节点 i 是最大密度的节点，则最小距离为该节点 i 与其他节点之间的距离的最大值；如果节点 i 不是最大密度的节点，则最小距离为比节点 i 局部密度大的节点与节点 i 之间的距离的最小值。

定义 12（中心节点值）　因社区中心节点的局部密度较大，且与其他更高密度的节点间的距离较大，故中心节点值定义为[2]：

$$\zeta_i=\rho_i*\delta_i \tag{13}$$

定义 13（中心节点值变化率）　排序后的中心节点值中，中心节点和非中心节点的值变化率较大。为了自动选定中心节点，将中心节点值变化率定义为[21]：

$$rate_i=(i-1)\frac{\zeta_{i-1}-\zeta_i}{\zeta_i-\zeta_{i+1}} \tag{14}$$

由式(14)可知，根据每个节点的中心节点值变化率可得到具有最大变化率值的节点 i，则节点 i 前面的所有节点就被自动挑选出来作为中心节点。

3.4　OCDRD 算法框架

结合粗糙集理论、灰色理论和密度峰值聚类方法，本文提出了重叠社区发现方法，其主要思想如下。

首先，根据局部结构相似性的定义得到相似性矩阵，再通过灰色理论计算得出全局结构相似性，进而得到关联度矩阵，由关联度矩阵计算出节点之间的距离。然后，通过每个节点的局部密度和最小距离自动识别出社区的中心节点。接着，根据节点到社区的距离与阈值 λ 的关系，将非中心节点划分到由中心节点对应的上近似集、下近似集中，并计算所有社区的边界域。在迭代过程中，阈值 λ 以一定的步长增加；针对社区边界域进行反复计算，边界域不断缩小。每次迭代都计算相应的评价指标 NMI，直到阈值达到最大值时算法终止。最后，选取 NMI 最高的社区结构作为最终的划分结构。

OCDRD 算法可分为以下几个步骤：

(1) 读取网络数据，得到邻接矩阵 $\boldsymbol{AdjacentMatrix}=(a_{ij})_{n\times n}$；

(2) 由式(2)计算出节点间的局部结构相似性，得到相似性矩阵 $\boldsymbol{SimMatrix}=(r_{ij})_{n\times n}$；

(3) 由式(10)计算出全局结构相似性，得到关联度矩阵，归一化后的关联度矩阵为 $\boldsymbol{GreyMatrix}=(g_{ij})_{n\times n}$；

(4) 计算距离矩阵 $\boldsymbol{DistanceMatrix}=(d_{ij})_{n\times n}$；

(5) 密度峰值聚类算法识别社区中心节点 $Center=\{v_1,v_2,\cdots,v_k\}$，具体步骤见 3.4.1 节；

(6) 处理非中心节点，进行社区划分，详见 3.4.2 节。

3.4.1　自动识别中心节点的方法

在重叠社区发现方法中，选择中心节点会存在以下问题：错将噪声节点当作中心节点，选择的中心节点不恰当或者中心节点个数不正确。这都会导致社区发现算法不准确。社区发现算法选取中心节点的经典方法有度中心性[22]、介数中心性[23]、紧密度中心性[24]和特征向量中心性[25]等。但这些方法并不能正确选取社区中心，且划分的社区结果质量不高。密度峰值算法能够根据数据分布来衡量节点的局部密度，并能够依据节点的局部密度自动选取局部密度较大且相互间距离较远的节点作为聚类中心[13]，因此本文采用密度峰值聚类算法来自动选取社区中心点。

在社区中心点的选取过程中，首先计算网络中各节点的中心节点值，中心节点值越大的节点被选为中心节点的概率越大。中心节点与非中心节点的中心节点值具有较大的差异，因此其中心节点值变化率较大，可以用斜率的变化来自动确定中心节点的个数[21]。

自动识别中心节点的方法可分为以下几个步骤：

(1) 由式(11)和距离矩阵 $\boldsymbol{DistanceMatrix}$ 计算出每个节点的局部密度 ρ_i；

(2) 由式(12)和距离矩阵 $\boldsymbol{DistanceMatrix}$ 计算出每个节点的最小距离 δ_i；

(3) 由式(13)计算出每个节点的中心节点值，并将其按升序排列；

(4) 由式(14)计算出每个节点的中心节点值变化率，取最大变化率值的节点 i 前面的节点作为中心节点。

3.4.2　社区的划分

假设算法自动选取了 k 个社区中心节点 $Center=\{v_1,v_2,\cdots,v_k\}$，初始化社区结构 $C=\{C_1,C_2,\cdots,C_k\}$，此时 $C_k=\{v_k\}$。初始化 $Normal$ 用于存储网络中的非中心节点。

社区划分的步骤如下：

(1) 初始化 $\lambda=0.1$，$i=1$；

(2) 当 $i\leqslant Normal.size()$ 时，通过式(9)计算 v_i 到各个社区的距离，并按照定义 7 将 v_i 划分到各社区的近似集中，否则执行步骤 5)；

(3) 当 $\exists C_r$ 使得 $v_i\in\underline{R}^{\lambda}(C_r)$ 时，将 v_i 从 $Normal$ 中移除并将 v_i 加入 C_r，赋予 v_i 对应的下近似社区标签，否则赋予 v_i 对应的上近似社区标签；

(4) $i=i+1$，执行步骤 2)；

(5) 当 $\lambda\leqslant 1$ 时，存储当前社区结构，并计算当前结构的评价值，阈值 λ 以一定步长进行更新，并更新 $i=1$，清空 $Normal$ 中节点的社区标签（此时 $Normal$ 存储的是所有社区边界域），执行步骤 2)，否则输出具有最高评价值的社区结构，算法结束。

3.4.3　时间复杂度

网络中，节点个数用 n 表示，边的条数用 m 表示，社区个数用 k 表示。在 3.4.1 节自动识别中心节点的方法中，计算每个节点的局部密度、最小距离和中心节点值的时间复杂度为 $O(n*n)$。在 3.4.2 节社区的划分中，将非中心节点关联到 k 个社区的近似集的时间复杂度为 $O((n-k)*k)$。OCDRD 算法中，计算节点的相似性、关联度和距离的时间复杂度为 $O(n*n)$。因此，该算法总的时间复杂度为 $O(n*n)$。

4　实验结果与分析

为了验证 OCDRD 算法的性能和可行性，将其与近几年效果较好的其他社区发现算法进行对比。

(1) CDRS：一种新的基于粗糙集的社区发现算法，能够刻画出社区的模糊区域[11]。

(2) DCN：一种基于标签传播的社区发现算法，不需要

调整任何参数[26]。

(3)LDC:一种基于密度峰值和链接密度的社区发现算法,能够自动确定社区中心节点[27]。

(4)RFC:一种新的基于粗糙模糊聚类的社区发现算法,能够应用于蛋白质网络中[3]。

4.1 实验数据

对于人工合成网络数据集,本文采用由 LFR 基准网络生成工具生成的数据[28]。一般来说,LFR 基准网络的参数设置方法如下:节点的平均度 k 为{5,10},节点度分布参数 γ 为 2,社区大小分布参数 β 为 1,节点最大度 $k_{\max}$ 为{20,50},最小社区包含的节点个数 $minc$ 的范围为(10,50),最大社区包含的节点个数 $maxc$ 的范围为(20,100),表示节点与社区外部连接概率的混合参数 μ 为{0.1,0.3}[29]。在生成重叠网络时,每个重叠节点连接的社区数参数为 O_m,重叠节点比例参数为 O_n,其取值根据情况而定[28]。为了测试本文算法在不同复杂混合网络和含有不同比例重叠节点的网络中的性能,本文分别生成了 2 组 LFR 基准网络,固定参数 $N=500, k=10, \gamma=2, \beta=1, k_{\max}=50, minc=10, maxc=20, O_n=10\%$,其中 μ 的取值分别为 0.1 和 0.3,每组 LFR 基准网络又分别生成 7 个人工网络,分别将参数 O_m 设置为 2～8。生成的网络数据集信息如表 1 所列。

表 1 人工合成网络数据集信息

Table 1 Information of artificial synthesis networks

name	N	k	$k_{\max}$	$minc$	$maxc$	μ	$O_n/\%$
LFR1	500	10	50	10	20	0.1	10
LFR2	500	10	50	10	20	0.3	10

真实数据集采用的是由 Mark Newman 提供的网站上的数据[30],真实网络数据集信息如表 2 所列。

表 2 真实网络数据集信息

Table 2 Information of real networks

name	N	E	C
Karate	34	78	2
Dolphins	62	159	2
Polbooks	105	441	3
Football	115	613	12
Polblogs	1490	19090	4
Netscience	1589	2742	404

其中,N 表示网络节点数,E 表示网络边数,C 表示社区数。Karate 网络为美国大学空手道俱乐部网络,包含 34 个节点和 78 条边,其中节点表示俱乐部成员,边表示成员间的友谊关系[31]。Dolphins 网络为新西兰 Doubtful Sound 海峡海豚联系网络,包括 62 个节点和 159 条边,其中节点表示海豚,边表示海豚间的频繁联系[32]。Polbooks 网络为美国总统大选时出版的政治书籍网络,包括 105 个节点和 441 条边,其中节点表示卖出的政治书籍,边表示有人同时购买这两本书籍[33]。Football 网络为美国大学生足球联赛网络,包括 115 个节点和 613 条边,其中节点表示足球队,边表示两个球队有一场比赛[34]。Polblogs 网络为美国博客政治倾向网络,包含了 1490 个节点和 19090 条边,节点表示博客为民主派或者保守派。Netscinece 网络为科学家合作网络,其中包含了 1589 个科学家节点,边为科学家之间的联系。其中,Polblogs 和 Netscinece 数据集无真实划分结构,社区数目 C 根据可视化工具 Gephi 得出。

4.2 评价准则

对于真实网络和人工合成网络,本文分别采用不同的评价准则。

(1)人工合成网络的评价准则

标准互信息是一种使用互信息度量两个集合间相近程度的信息论方法,是社区发现中的一个重要衡量标准[35],其定义如下:

$$I(X,Y)=\frac{H(X)+H(Y)-H(X,Y)}{(H(X)+H(Y))/2} \tag{15}$$

其中,$H(X)$和 $H(Y)$分别是随机变量 X 和 Y 的熵;$H(X,Y)$为联合熵,定义如下:

$$H(X,Y)=H(X)+H(Y|X) \tag{16}$$

NMI 不能直接评价重叠社区,因为没有考虑社区被划分后社团间存在重叠节点的信息量。对此,Lancichinetti 等在 NMI 基础上提出了一种改进的 NMI[36]。在算法划分的社区结构 X 中,节点 x_i 是否隶属于 X 的不同社团用一个向量 $\vec{q}_i$ 表示,其长度为 K。$\vec{q}_i$ 的取值为 1 表示节点 x_i 属于第 i 个社区,否则取值为 0。将 $\vec{q}_i$ 的第 k 个元素看作一个随机变量 X_k。同样,在真实划分的社区结构 Y 中,节点对于第 r 个社区的概率分布为 Y_r。则随机变量 X_k 在所有 $Y_r(Y)$上的条件熵定义为:

$$H(X_k|Y)=\min_{r\in\{1,2,\cdots,R\}}H(X_k|Y_r) \tag{17}$$

因此,所有 $X_k(X)$在 Y 上的规范化条件熵定义为:

$$H(X|Y)=\frac{1}{|K|}\sum_{k=1}^{K}\frac{H(X_k|Y)}{H(X_k)} \tag{18}$$

类似地,可以计算 Y 在 X 上的规范化条件熵 $H(Y|X)$。因此,改进后的 NMI 定义如下:

$$NMI(X|Y)=1-\frac{H(X|Y)+H(Y|X)}{2} \tag{19}$$

(2)真实网络的评价准则

如果已知真实网络数据集中社区的真实划分结构,则使用 NMI 和具有重叠性的模块度 EQ[37]作为评价准则,否则评价指标为 EQ。EQ 也是评价重叠社区划分质量的一种常见方法,定义如下:

$$EQ=\frac{1}{2e}\sum_{l}\sum_{v\in C_i,w\in C_j}\frac{1}{Q_vQ_w}\left[\boldsymbol{A}_{vw}-\frac{k_vk_w}{2e}\right] \tag{20}$$

其中,e 为网络中边的条数,k_v 和 k_w 分别是节点 v 和 w 的度,$\boldsymbol{A}_{vw}$为邻接矩阵,C_i 表示第 i 个社区,Q_v 表示节点 v 所属社区的个数。

4.3 参数设置和实验环境

OCDRD 算法的截断距离 d_c 的取值为所有数据节点之间的距离升序排列后的 1%～2%。下近似权重 α 和上近似集权重 β 的取值满足 $\alpha+\beta=1$ 且 $\alpha>\beta$,本文 α 取 0.7,β 取 0.3。CDRS 算法的主要参数是指定真实社区的个数 K,K 可根据上述社区数目得出;决定社区边界区域的阈值 λ 取 0.909。DCN 算法中,切比雪夫不等式参数 ε 取 2。LDC 算法中,截断距离 d_c 取值为所有数据节点之间的距离升序排列后的 2%。RFC 算法中,阈值 λ_2 的取值范围为 $0.8\lambda_1\leqslant\lambda_2\leqslant0.9\lambda_1$。

所有实验的环境为:4GB 内存,Windows10 64bit 操作系统。算法编程语言均为 Java,IDE 环境为 Eclipse。

4.4 人工合成网络社区发现的实验结果

将本文所提算法分别与 CDRS,LDC 和 RFC 算法进行比较。在两组人工合成网络上的测试结果如图 1 和图 2 所示。

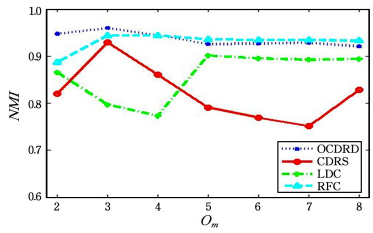

图 1　在 LFR1 数据集上的测试结果

Fig. 1　Test results on LFR1 datasets

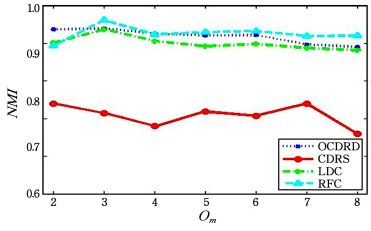

图 2　在 LFR2 数据集上的测试结果

Fig. 2　Test results on LFR2 datasets

从图 1 和图 2 可知，随着混合参数 μ 和重叠节点连接社区数参数 O_m 的增加，OCDRD 算法的 *NMI* 值都较稳定，且取值均在 0.9 以上。CDRS 算法在低复杂网络中，随着 O_m 的增加，震动幅度较大；在高复杂网络中，随着 O_m 的增加，*NMI* 值减小。LDC 和 RFC 算法在低复杂网络中划分重叠节点不准确，所以前期出现了不稳定的现象。

4.5　真实网络社区发现的实验结果

将本文所提算法与其他社区发现算法在真实网络数据集上进行测试，将 NMI 和具有重叠性的模块度 EQ 作为测量标准。为了直观地展示算法之间的准确性，本文还分别比较了真实网络数据集的 *NMI* 和 *EQ* 的均值。算法在真实网络数据集上的 *NMI* 和 *EQ* 测试结果分别如表 3 和表 4 所列；在真实网络数据集上的 *NMI* 和 *EQ* 的均值如图 3 和图 4 所示。

表 3　真实网络数据集上的 *NMI* 测试结果

Table 3　*NMI* test results on real network datasets

Algorithm	Karate	Dolphins	Polbooks	Football
OCDRD	**1.0**	**0.8982**	0.6310	**0.8979**
CDRS	0.8267	0.3684	0.4725	0.8677
DCN	**1.0**	0.8888	0.5979	0.3445
LDC	0.8510	0.8095	**0.6523**	0.8671
RFC	0.5236	0.5161	0.6411	0.8302

表 4　真实网络数据集上的 *EQ* 测试结果

Table 4　*EQ* test results on real network datasets

Algorithm	Karate	Dolphins	Polbooks	Football	Polblogs	Netscience
OCDRD	**0.3715**	**0.5142**	**0.5117**	**0.5811**	**0.2924**	**0.6676**
CDRS	0.3434	0.3717	0.3777	0.5510	0.1400	0.1209
DCN	0.3715	0.3787	0.4456	0.3534	0.1269	0.6663
LDC	0.2469	0.3693	0.4339	0.2415	0.0780	0.4716
RFC	0.3255	0.1869	0.4257	0.5179	0.1044	0.0707

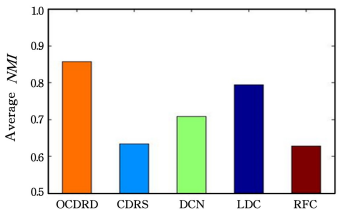

图 3　真实网络数据集上的 *NMI* 均值

Fig. 3　Average *NMI* on real network datasets

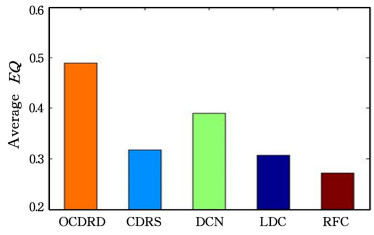

图 4　真实网络数据集上的 *EQ* 均值

Fig. 4　Average *EQ* on real network datasets

从图 3 和图 4 可知，在真实网络数据集上，不论是 *NMI* 均值还是 *EQ* 均值，本文提出的 OCDRD 方法均取得了最大值。

Karate，Dolphins，Polbooks 和 Football 数据集的社区真实划分结构均可知，因此可采用 NMI 和 EQ 评价指标；Polblogs 和 Netscience 数据集的社区真实划分结构未知，因此只采用 EQ 评价指标。从表 3 和表 4 可知，本文提出的 OCDRD 方法在 Karate，Dolphins 和 Football 数据集上，*NMI* 和 *EQ* 均取得最大值；在 Polblogs 和 Netscience 数据集上，*EQ* 也取得最大值。在 Karate 数据集上，DCN 算法和 OCDRD 算法的 *NMI* 和 *EQ* 均取得最大值。在 Polbooks 数据集上，LDC 算法的 *NMI* 取得最大值。总的来说，本文提出的 OCDRD 算法划分的重叠社区结构较其他社区发现算法更有效。

结束语　本文提出了一种基于粗糙集和密度峰值的重叠社区发现算法 OCDRD。该算法结合密度峰值的思想自动确定社区个数，避免了人工通过先验知识确定社区个数；结合粗糙集的思想来划分社区的不确定性区域，从而在不确定性区域挖掘重叠节点，得到划分更优的社区结构。在人工数据网络和真实数据网络中，将 OCDRD 算法与近几年效果较好的社区发现算法进行比较，结果证明 OCDRD 算法是有效的。OCDRD 算法的时间复杂度较高，在将较高复杂度的算法应用到大规模网络时，一般会将算法进行并行化设计，因此今后将重点针对算法并行化设计和社区划分质量进行相关研究。

参 考 文 献

[1]　NEWMAN M E J. Networks: An Introduction[EB/OL]. [2010-09]. http://www. oxfordscholarship. com/view/10. 1093/acprof:oso/9780199206650. 001. 0001/acprof-9780199206650.

[2]　BAI X Y, YANG P L, SHI X H. An overlapping community detection algorithm based on density peaks[J]. Neurocomputing, 2017, 226: 7-15.

[3]　WU H, GAO L, DONG J H, et al. Detecting overlapping protein complexes by rough-fuzzy clustering in protein-protein interaction networks[J]. PLoS One, 2014, 9(3): e91856.

[4]　ZHANG X Y, WANG C T, SU Y S, et al. A fast overlapping community detection algorithm based on weak cliques for large-scale networks[J]. IEEE Transactions on Computational Social Systems, 2017, 4(4): 218-230.

[5]　SUN H L, LIU J, HUANG J B, et al. LinkLPA: a link-based label propagation algorithm for overlapping community detection in networks[J]. Computational Intelligence, 2017, 33(2): 308-331.

[6]　YU Z Y, CHEN J J, QUO K, et al. Overlapping community detection based on random walk and seeds extension[C]// Pro-

ceedings of the 12th Chinese Conference on Computer Supported Cooperative Work and Social Computing (ChineseCSCW '17). New York, USA: ACM Press, 2017: 18-24.
[7] BANSAL S, BHOWMICK S, PAYMAL P. Fast community detection for dynamic complex networks[M]// Communications in Computer and Information Science. Springer, 2011: 196-207.
[8] PAWLAK Z. Rough sets [J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356.
[9] GRECO S, MATARAZZO B, SLOWINSKI R. Rough sets theory for multicriteria decision analysis[J]. European Journal of Operational Research, 2001, 129(1): 1-47.
[10] ZHANG W X, WU W Z. An introduction and a survey for the studies of rough set theory[J]. Fuzzy Systems and Mathematics, 2000, 14(4): 1-12.
[11] ZHANG Y, WU B, LIU Y. A novel community detection method based on rough set K-means[J]. Journal of Electronics and Information Technology, 2017, 39(4): 770-777.
[12] ZHU W Q, FU Y C. Community structure detection algorithm based on rough set[J]. Computer Engineering, 2011, 37(14): 41-43.
[13] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.
[14] DENG J L. The relational space in grey system theory[J]. Fuzzy Mathematics, 1985(2): 1-10.
[15] NEWMAN M E J, MICHELLE G. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 026113-026113.
[16] ZHOU X, LIU Y H, WANG J, et al. A density based link clustering algorithm for overlapping community detection in networks [J]. Physica A: Statistical Mechanics and Its Applications, 2017, 486: 65-78.
[17] XU X, YURUK N, FENG Z, et al. SCAN: a structural clustering algorithm for networks[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2007: 824-833.
[18] LIU S F, CAI H, YANG Y J, et al. Advance in grey incidence analysis modelling[J]. Systems Engineering-Theory & Practice, 2013, 33(8): 2041-2046.
[19] ZHANG Q, QIU Q, GUO W, et al. A social community detection algorithm based on parallel grey label propagation[J]. Computer Networks, 2016, 107(1): 133-143.
[20] WANG M M, ZUO W, WANG Y. An improved density peaks-based clustering method for social circle discovery in social networks[J]. Neurocomputing, 2016, 179: 219-227.
[21] YANG Z, WANG H J, ZHOU Y. A clustering algorithm with adaptive cut-off distance and cluster centers[J]. Data Analysis and Knowledge Discovery, 2018(3): 39-48.
[22] DWYER T, HONG S H, DIPK K, et al. Visual analysis of network centralities[C]// Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation. 2006: 189-197.
[23] BARTHELEMY M. [Lecture Notes in Morphogenesis] Morphogenesis of Spatial Networks ‖ Betweenness centrality[OL]. https://www.onacademic.com/detail/journal_1000040157249110_02f6.html.
[24] SABIDUSSI G. The centrality index of a graph [J]. Psychometrika, 1966, 31(4): 581-603.
[25] BONACICH P. Power and centrality: a family of measures[J]. American Journal of Sociology, 1987, 92(5): 1170-1182.
[26] DING J J, HE X X, YUAN J Q, et al. Community detection by propagating the label of center[J]. Physica A: Statistical Mechanics and Its Applications, 2018, 503: 675-686.
[27] HUANG L, WANG G S, WANG Y, et al. A link density clustering algorithm based on automatically selecting density peaks for overlapping community detection[J]. International Journal of Modern Physics B, 2016, 30(24): 1650167.
[28] LANCICHINETTI A, FORTUNATO S. Community detection algorithms: a comparative analysis[J]. Physical Review E, 2009, 80(5): 056117.
[29] XIE J R, KELLEY S, SZYMANSKI B K. Overlapping community detection in networks[J]. ACM Computing Surveys, 2013, 45(4): 1-35.
[30] NEWMAN M. Network data[EB/OL]. [2013-04-19]. http://www-personal.umich.edu/~mejn/netdata/.
[31] ZACHARY W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 33(4): 452-473.
[32] LUSSEAU D, SCHNEIDER K, BOISSEAU O J, et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations[J]. Behavioral Ecology and Sociobiology, 2003, 54(4): 396-405.
[33] KREBS V. Books about US Politics [EB/OL]. http://www.orgnet.com/.
[34] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[EB/OL]. http://arxiv.org/abs/cond-mat/0112110/.
[35] STREHL A, GHOSH J. Cluster ensembles: a knowledge reuse framework for combining partitionings[C]// Eighteenth national conference on Artificial intelligence. 2002: 93-98.
[36] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics, 2009, 11(3): 033015.
[37] NICOSIA V, MANGIONI G, CARCHIOLO V, et al. Extending the definition of modularity to directed graphs with overlapping communities[J]. Journal of Statistical Mechanics: Theory and Experiment, 2009(3): P03024.

ZHANG Qin, born in 1995, postgraduate, is a member of China Computer Federation. Her main research interests include database technology and data mining.

CHEN Hong-mei, born in 1971, Ph.D, professor, Ph.D supervisor, is a ember of China Computer Federation. Her main research interests include granular calculation, rough sets and intelligent information processing.