计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190800002

基于粗糙集和距离动态模型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 现实世界可被看作由许多不同的复杂系统组成。为了建模分析复杂系统中个体间隐藏的规律及功能,将复杂系统抽象为由节点和边组成的复杂网络。挖掘复杂网络中的社区结构在内容推荐、行为预测和疾病扩散等方面具有重要的理论意义和实际价值。随着复杂系统内个体的不断变化,多个社区间出现了重叠节点,有效且准确地挖掘社区中的重叠节点具有一定的挑战性。为了有效发现社区中的重叠节点,提出了一种基于粗糙集和距离动态模型的重叠社区发现方法(Overlapping Community Detection based on Rough sets and Distance Dynamics model,OCDRDD)。该方法首先根据网络的拓扑结构,结合节点度中心性和距离选出 K 个核心节点;然后按照定义的距离比率关系初始化社区的近似集和边界域,结合距离动态模型,迭代变化边界域节点与下近似集节点间相连的边的距离,且在每次迭代过程中将符合定义的距离比率关系的边界域节点划分到社区下近似集中,以缩小边界域节点(即缩小边界域的范围),直到找到最佳重叠社区结构;最后根据定义的两条规则处理"伪"重叠节点。在真实网络数据集和 LFR Benchmark 人工网络数据集上,以 NMI 和具有重叠性的模块度 EQ 作为评价指标,将 OCDRDD 方法与近几年具有代表性的社区发现方法进行实验测试比较,发现 OCDRDD 方法整体优于其他算法,结果表明该算法具有有效性和可行性。

关键词:重叠社区发现;粗糙集;距离动态模型;边界域
中图法分类号 TP391

Overlapping Community Detection Method Based on Rough Sets and Distance Dynamic Model

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 The real world is considered to be composed of many different complex systems. In order to model and analyze the hidden rules and functions among individuals in complex systems,the complex system may be abstracted as a complex network composed of nodes and edges. Mining community structures in complex networks has important theoretical significance and practical value in content recommendation,behavior prediction and disease spread. With the continuous changes of individuals in complex systems,overlapping nodes appear among multiple communities. How to effectively and accurately mine the overlapping nodes in communities has brought some challenges. In order to effectively detect the overlapping nodes in the community,an overlapping community detection method based on rough sets and distance dynamic model (OCDRDD) is proposed in this paper. First of all,according to the topology of the network,it selects K core nodes by combining node degree centrality and distance,then initializes the approximation sets and the boundary region of the community according to the distance ratio relationship. Combined with the distance dynamic model,the distances between boundary region nodes and the lower approximation set nodes are changed iteratively. During each iteration,boundary region nodes that conform to the defined distance ratio relationship are classified into the lower approximation of the community,and the boundary region nodes are reduced until the optimal overlapping community structure is found. Finally,the "pseudo" overlapping nodes are processed according to the two rules defined in this paper. NMI and overlapping module degree EQ are taken as evaluation indexes on real network datasets and LFR Benchmark artificial network datasets. The OCDRDD method is compared with other typical community detection methods in recent years both on real network datasets and LFR Benchmark artificial network datasets. The experimental results show that OCDRDD method is better

到稿日期:2019-07-31 返修日期:2019-11-11 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(61572406,61976182);四川省国际科技创新合作重点项目(2019YFH0097)
This work was supported by the National Natural Science Foundation of China (61572406,61976182) and Key Program for International Science and Technology Innovation Cooperation of Sichuan Province,China (2019YFH0097).
通信作者:陈红梅(hmchen@swjtu.edu.cn)

than other community detection algorithms on the whole. The results show that the proposed algorithm is effective and feasible.
Keywords Overlapping community detection, Rough set, Distance dynamic model, Boundary region

1 引言

自然界由许多不同的复杂系统组成。我们将一个复杂系统中的个体抽象为节点，个体与个体之间的联系抽象为边，则复杂系统可以抽象表示为复杂网络。比如，社交网络中将用户抽象为节点，用户之间的联系抽象为边；蛋白质网络中将蛋白质抽象为节点，蛋白质之间的相互作用抽象为边等。在复杂网络中，我们可以挖掘出具有重要研究意义的社区结构。社区结构表现为同一社区中的节点集合连接紧密或者相似度较高，不同社区间的节点集合连接稀疏或者相似度较低[1]。社区结构展现了网络的拓扑结构和功能结构。社区发现是根据网络节点间的连接紧密程度或者相似程度划分社区结构的过程。如果一个节点被划分到多个社区，则为重叠社区发现，该节点为重叠节点[2]。重叠节点作为信息量较多的节点，具有重要的实际研究意义。比如，社交网络中有些用户有多种喜好，其既喜欢篮球又喜欢足球；蛋白质网络中有些蛋白质具有多种功能结构。重叠社区发现已经被广泛应用在数据挖掘、社会学、计算机科学和生物学等领域，在功能预测、推荐系统、广告投放和信息传播等方面具有较好的效果，因此具有重要的研究意义[3]。

近几年，重叠社区发现方法是一个重要的研究问题，可大致分为以下几类：基于节点划分方法、基于边分割方法和基于动态过程的方法[4]。Bai 等提出了一种新的基于密度峰值的重叠社区发现方法，先基于相似度设置节点间的距离，然后根据定义好的最小距离和局部密度选取中心节点，最后分配其他的节点[5]。Zhou 等提出了一种基于密度的链接聚类重叠社区发现方法，该方法以边作为研究对象；并提出了核心边和基于核心密度可达的概念，以及未分类边的更新策略[6]。Sun 等提出了一种基于链接的标签传播重叠社区发现方法以及一种新的标签传播方法，将传统的基于节点的标签传播转化为基于边的标签传播算法，进而发现重叠节点[7]。Wu 等提出了一种基于模糊粗糙聚类方法并将其应用在蛋白质网络的重叠社区发现中，该方法将模糊理论和粗糙集理论相结合，采用粗糙集的上、下近似来自动处理重叠节点，建立蛋白质的模糊关系和模糊等价类，并定义了单个蛋白质与社区的相似性，从而进行社区划分[8]。Sun 等提出了一种新的基于信息动力学的重叠社区发现方法，该方法将网络看作一个动态系统，节点与其邻居节点通信并共享信息，进而根据网络的拓扑结构计算出每个节点信息，最后根据信息动力学划分社区结构[9]。

为了研究社区发现的动态过程，Shao 等于 2015 年提出了基于同步启发的距离动态模型，该模型将网络看作一个自适应动态系统，节点与邻居节点进行交互，由于相互作用与距离相互影响，因此实现了网络的动态划分[10]。该方法能够发现任意密度的社区，检测到噪音和异常点，但是不能发现重叠社区，且存在收敛慢等问题。Chen 等提出了一种基于距离动态模型的重叠社区发现方法，该方法将原始图转换为一个新的边图，然后利用改进的基于距离的动态模型进行社区划分，再将新的边图转化为原始图，从而得到重叠社区结构[11]。但是，该方法存在复杂度高、社区划分结构有待优化的问题。

社区发现是一个动态的过程，为此本文引入了距离动态模型，使得挖掘的社区结构更加合理。在动态过程中，为了找到社区的重叠节点，本文应用粗糙集理论来解决社区中重叠节点的挖掘问题，以避免全局计算，提高计算效率。粗糙集理论是由 Pawlak 提出的用于解决不确定、不准确和模糊数据的一种软计算方法[12]。粗糙集理论可以通过上、下近似集很好地刻画出重叠节点[8]。比如，Zhang 等通过粗糙集理论描述了社区的模糊区域，进而得到含有模糊区域的社区结构[13]。本文提出了一种基于粗糙集和距离动态模型的重叠社区发现方法 OCDRDD。OCDRDD 的主要思想为：首先根据网络拓扑结构，结合节点的度中心性和距离确定出 K 个核心节点；然后将定义的距离比率关系的阈值设为定值，并取值为最小值 1.1，以保证划分到社区的下近似集的节点是正确的，同时初始化社区的近似集和边界域。针对边界域节点，通过距离动态模型迭代调整边界域里的节点与社区下近似里的节点直接相连的边的距离，每次迭代将符合距离比率关系的节点划分到社区下近似集中，使得边界域逐渐减小，直到划分出最佳的重叠社区结构。此时识别出的重叠节点有部分实际上是非重叠节点，因为这类节点离所有的核心节点较远，且不满足比值关系，我们称这类节点为“伪”重叠节点。对“伪”重叠节点进行处理，将其划分到属于正确社区的下近似集。最后，基于真实网络数据集和人工网络数据集，将所提算法分别与近几年的重叠社区发现算法进行了比较。实验结果表明，本文所提算法是有效且可行的。

2 相关基础知识

本节将介绍社区发现的相关基本概念、粗糙集理论[12]和距离动态模型[10]等相关知识。

2.1 社区发现的相关定义

将复杂网络建模为图的形式，其中网络的个体看作图的节点，网络个体与个体间的联系看作图的边。我们可将图表示为 $G=(V,E)$，其中网络中的节点集合表示为 $V=\{v_1, v_2, \cdots, v_n\}$，边的集合表示为 $E=\{e_1, e_2, \cdots, e_n\}$。我们根据节点间的距离或者相似性得到划分后的社区结构，其中节点间的距离或者相似性可以根据节点的邻居节点确定。社区发现的相关定义如下。

定义 1（邻居节点） 给定复杂网络 $G=(V,E)$，$v_i \in V$，则节点 v_i 的邻居节点 $Nei(v_i)$ 表示为[6]：

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

定义 2（节点的度） 给定复杂网络 $G=(V,E)$，$v_i \in V$，则节点 v_i 的度 K_i 表示为[14]：

$$K_i=|v_j \mid v_j \in V(i \neq j) \wedge (v_i, v_j) \in E| \tag{2}$$

其中，$|*|$ 表示集合内元素的个数，节点的度表示在网络中与节点 v_i 直接相连的节点的个数。节点 v_i 与网络中其他节点

连接的数目越多，则其度越大，反映了节点 v_i 在网络中的重要性越大。

定义 3（Jaccard 局部相似性）　给定复杂网络 $G=(V,E)$，$\forall v_i,v_j\in V(i\neq j)$，则 v_i 和 v_j 的 Jaccard 局部相似性 $Sim(v_i,v_j)$ 定义为[15]：

$$Sim(v_i,v_j)=\frac{|Nei(v_i)\bigcap Nei(v_j)|}{|Nei(v_i)\bigcup Nei(v_j)|} \tag{3}$$

其中，$|*|$ 表示集合内元素的个数，$Sim(v_i,v_j)$ 表示节点 v_i 和 v_j 的结构化相似程度。两个节点的共同邻居越多，则两个节点越相似，$Sim(v_i,v_j)$ 值越大。

定义 4（Jaccard 距离）　给定复杂网络 $G=(V,E)$，$\forall v_i,v_j\in V(i\neq j)$，则 v_i 和 v_j 的 Jaccard 距离 $d(v_i,v_j)$ 定义为[15]：

$$d(v_i,v_j)=1-Sim(v_i,v_j) \tag{4}$$

由式(4)可知，Jaccard 距离根据 Jaccard 局部相似性得到，当节点 v_i 和 v_j 的相似性越大时，它们之间的距离越小。

2.2　粗糙集理论的相关知识

粗糙集是一种处理不确定性信息的数学方法，可描述信息的不确定性部分。

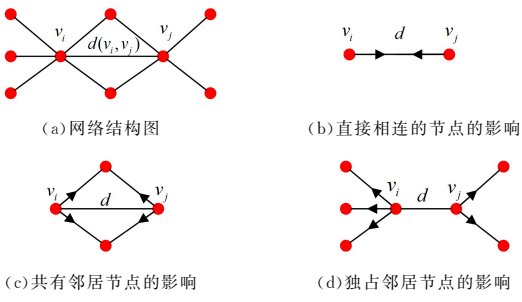
定义 5（决策信息系统）　决策信息系统由一个四元组 $DIS=(U,C\cup D,V,f)$ 表示。其中，论域 $U=\{x_1,\cdots,x_i,\cdots,x_n\}$ 为非空有限集合；C 为条件属性，D 为决策属性，所以 $C\cup D$ 为非空有限集合；V 为属性的值域；$f:U\rightarrow V$ 为信息函数，$f(x_i,a_i)(a_i\in C\cup D)$ 表示对象 x_i 在属性 a_i 上的属性值。

定义 6（等价类）　给定四元组决策信息系统 $DIS=(U,C\cup D,V,f)$，其中 R 为 U 上的等价关系，由等价关系 R 形成的划分为 $U/R=\{E_1,\cdots,E_i,\cdots,E_k\}$，$E_i$ 为等价类。

定义 7（上、下近似集）　给定四元组决策信息系统 $DIS=(U,C\cup D,V,f)$，设 $\forall X\subseteq U$，粗糙集中的下近似集表示确定属于某个任意集合，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，则 X 的上近似集为 $\overline{R}(X)=\bigcup\{E_i\in U/R:E_i\bigcap X\neq\varnothing\}$。

2.3　距离动态模型

距离动态模型的主要思想是：网络中节点间的距离是逐渐变化的，随着时间的推移，距离会存在两种变化方式，属于同一个社区的节点间距离会逐渐变小，属于不同社区间的节点间距离会逐渐变大，最终达到一个稳定的状态，使得社区结构呈现出来。距离动态模型是通过检测节点间距离的变化动态地发现社区结构，如图 1(a)所示，节点 v_i 和节点 v_j 间距离 $d(v_i,v_j)$ 的变化会受到 3 种不同邻居节点的影响，如图 1 中(b)－(d)所示，从而引出 3 种交互模式。



(a)网络结构图　(b)直接相连的节点的影响

(c)共有邻居节点的影响　(d)独占邻居节点的影响

图 1　距离动态模型的 3 种交互模式

Fig. 1　Three interaction modes of distance dynamic model

交互模式 1：来自直接链接节点的影响。直接链接节点是指直接相连的两个端节点 v_i 和 v_j 之间的影响。两个端节点之间会彼此吸引，使得距离 $d(v_i,v_j)$ 减小，如图 1(b)所示。直接链接节点的影响 DI 定义为：

$$DI(v_i,v_j)=-\left(\frac{f(1-d(v_i,v_j))}{K_i}+\frac{f(1-d(v_i,v_j))}{K_j}\right) \tag{5}$$

其中，$f(\cdot)$ 为耦合函数，一般为 $\sin(\cdot)$ 函数。$1-d(v_i,v_j)$ 表示节点 v_i 和 v_j 的相似程度，$d(v_i,v_j)$ 为 v_i 和 v_j 的 Jaccard 距离。节点 v_i 和 v_j 越相似，则它们之间来自直接链接节点的影响就越大。K_i 和 K_j 分别表示节点 v_i 和 v_j 的度；$\frac{1}{K_i}$ 和 $\frac{1}{K_j}$ 为归一化因子，用于衡量节点间不同度的影响程度。

交互模式 2：来自公共邻居的影响。公共邻居是指同时与两个端节点相连的节点集合。公共邻居会吸引两个端节点 v_i 和 v_j 向自身移动，从而导致距离 $d(v_i,v_j)$ 减小，如图 1(c)所示。公共邻居的影响 CI 定义为：

$$\begin{aligned}CI(v_i,v_j)=-\sum_{v_k\in CN}\Big(&\frac{1}{K_i}\cdot f(1-d(v_k,v_i))\cdot(1-d(v_k,v_j))+\\&\frac{1}{K_j}\cdot f(1-d(v_k,v_j))\cdot(1-d(v_k,v_i))\Big)\end{aligned} \tag{6}$$

其中，CN 表示节点 v_i 和 v_j 的共同邻居，表示为 $CN(v_i,v_j)=\{Nei(v_i)-v_i\}\bigcap\{Nei(v_j)-v_j\}$。

交互模式 3：来自排他邻居的影响。排他邻居是指只与一个端节点相连，与另一个端节点不相连的节点集合。因此我们将端节点 v_i 的排他邻居定义为 $EN(v_i)=\{Nei(v_i)-\{Nei(v_i)\bigcap Nei(v_j)\}\}$，端节点 v_j 的排他邻居定义为 $EN(v_j)=\{Nei(v_j)-\{Nei(v_i)\bigcap Nei(v_j)\}\}$。排他邻居只会吸引直接相连的端节点向自身移动。为了刻画排他邻居是否吸引非直接相连的端节点向自身移动，引入了凝聚力参数 λ，则在距离 $d(v_i,v_j)$ 上的影响积极度 $\rho(v_k,v_i)$ 定义为：

$$\rho(v_k,v_i)=\begin{cases}1-d(v_k,v_j), & 1-d(v_k,v_j)\geqslant\lambda\\ 1-d(v_k,v_j)-\lambda, & \text{otherwise}\end{cases} \tag{7}$$

则排他邻居的影响 EI 定义为：

$$\begin{aligned}EI(v_i,v_j)=&-\sum_{v_x\in EN(v_i)}\left(\frac{1}{K_i}\cdot f(1-d(v_x,v_i))\cdot\rho(v_x,v_i)\right)-\\&\sum_{v_y\in EN(v_j)}\left(\frac{1}{K_j}\cdot f(1-d(v_y,v_j))\cdot\rho(v_y,v_j)\right)\end{aligned} \tag{8}$$

最后，结合 3 种交互模式，节点 v_i 和节点 v_j 间的距离 $d(v_i,v_j)$ 随时间的动态变化为：

$$\begin{aligned}d(v_i,v_j,t+1)=&d(v_i,v_j,t)+DI(v_i,v_j,t)+CI(v_i,v_j,t)+\\&EI(v_i,v_j,t)\end{aligned} \tag{9}$$

3　基于粗糙集和距离动态模型的重叠社区发现方法

社区发现是根据节点间的连接紧密程度或者相似程度来划分社区，一般是由核心节点为中心去扩展社区，因此核心节点的选择对于提升算法的性能和准确率尤为重要。重叠社区发现是在社区发现方法中考虑到属于多个社区的重叠节点，因此挖掘出重叠节点也同等重要。

为了更有效地发现核心节点，考虑到复杂度问题，本文提

出了结合度中心性和核心距离来确定核心节点的方法。为了准确、合理地挖掘重叠节点，并且考虑到网络节点间的距离是动态变化的，我们引入了粗糙集理论和距离动态模型。由于边界域的节点与社区下近似集的节点直接相连的边的距离会随时间的变化而变化，属于单个社区的边界域节点与下近似节点的距离会越来越小，属于多个社区的边界域节点与下近似节点的距离会越来越大，因此随着时间的变化，我们会获得最佳的重叠社区结构。最后，对“伪”重叠节点进行处理。

3.1 结合度中心性和核心距离的核心节点选择方法

一般地，我们选取的核心节点是在网络中具有重要性的节点。度中心性是一种基于局部的中心性节点选择方法，复杂度低。但是，核心节点除了重要性高以外，还应该保证核心节点间的距离较大，以防止两个核心节点被划分到同一个社区。因此，本文提出了结合度中心性和核心距离进行核心节点选择的方法，其相关定义如下。

定义 8（节点欧氏距离） 给定复杂网络 $G=(V,E)$，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，v_j 由 Jaccard 距离构成的数列，即$\vec{v_i}=\{d(v_i,v_1),\cdots,d(v_i,v_l),\cdots,d(v_i,v_n)\}$，$\vec{v_j}=\{d(v_j,v_1),\cdots,d(v_j,v_l),\cdots,d(v_j,v_n)\}$，则$\vec{v_i}$和$\vec{v_j}$的欧氏距离 $Ed(v_i,v_j)$定义为：

$$Ed(v_i,v_j)=\sqrt{\sum_{k=1}^{n}(d(v_i,v_k)-d(v_j,v_k))^2} \tag{10}$$

由于定义 4 的 Jaccard 距离考虑的是节点间的局部距离，为了考虑节点间的全局距离，将节点扩展为由 Jaccard 距离构成的向量，再计算出节点的欧氏距离。

定义 9（节点核心距离） 社区核心节点与度中心性较高的节点的距离相对较大，则节点的核心距离 δ_i 表示为：

$$\delta_i=\begin{cases}\max\limits_{j}(Ed(v_i,v_j)), & \text{if } K_i=\max\limits_{k}(K_k)\\ \min\limits_{j:K_j>K_i}(Ed(v_i,v_j)), & \text{otherwise}\end{cases} \tag{11}$$

由式(11)可得，对于度中心性最高的节点，节点核心距离为该节点到其他节点距离的最大值；如果节点不是最大度中心性节点，则节点核心距离为该节点到比该节点度中心性大的节点的欧氏距离的最小值。

定义 10（中心节点值） 核心节点除自身重要性高以外，还应该保证核心节点间的距离较大，因此我们将节点欧氏距离与节点核心距离的乘积作为节点核心距离，定义如下：

$$\zeta_i=Ed(v_i,v_j)*\delta_i \tag{12}$$

3.2 社区近似集

粗糙集可以处理不确定性信息，因此可以很好地将粗糙集理论应用在社区发现中来识别重叠节点。网络中社区粗糙集的相关定义如下。

定义 11（社区上近似集、下近似集） 给定复杂网络 $G=(V,E)$，假定社区划分为 $C=\{C_1,\cdots,C_l,\cdots,C_r\}$，则社区的上近似集、下近似集分别定义如下：

$$\overline{R^{\lambda}}(C_l)=\left\{v_i\ \middle|\ \frac{d(v_i,C_l)}{d(v_i,C_k)}\leqslant Min,\forall v_i\in V,\exists C_k\in C(k\neq l)\right\} \tag{13}$$

$$\underline{R^{\lambda}}(C_l)=\left\{v_i\ \middle|\ \forall C_k\in C(k\neq l),\frac{d(v_i,C_l)}{d(v_i,C_k)}\leqslant Min,\forall v_i\in V\right\} \tag{14}$$

其中，$\frac{d(v_i,C_l)}{d(v_i,C_k)}$为距离比值，$Min$ 为社区上近似集和下近似集中的比例阈值，取最小值 1.1，以保证每次划分到社区下近似的为非重叠节点；$d(v_i,C_i)$表示节点 v_i 到社区 C_i 的距离。

因此，社区的边界域定义如下：

$$Bnd(C_l)=\overline{R^{\lambda}}(C_l)-\overline{R^{\lambda}}(C_l) \tag{15}$$

$$Bnd(C)=\bigcup_{i=1}^{r}Bnd(C_i) \tag{16}$$

其中，社区 C_l 的边界域表示为 $Bnd(C_l)$；$Bnd(C)$表示将每个社区的边界域求并集，得到的所有社区划分的边界域。

我们将社区上近似集和下近似集中的比例阈值取为最小值。社区划分是一个动态过程，节点间的距离会随着时间的变化而变化，因此本文定义一个动态的节点到社区的距离。

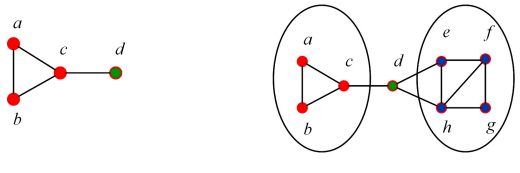
定义 12（节点到社区的距离） 在基于粗糙集理论的重叠社区发现中，节点到社区的距离分为两种情形：当初始化社区 C_j 中只有一个中心节点 v_j 时，非中心节点 v_i 到社区 C_j 的距离为节点 v_i 到节点 v_j 的 Jaccard 距离；随着迭代的进行，当社区中节点数多于 1 时，非中心节点 v_i 到社区 C_j 的距离为非中心节点 v_i 到社区 C_j 下近似集的距离占重及其到社区 C_j 上近似集的距离的权重和。因此，节点 v_i 与社区 C_j 的距离定义为：

$$d(v_i,C_j)=\begin{cases}d(v_i,v_j), & \text{若}|C_j|=1\\ \alpha*d(v_i,\underline{R^{\lambda}}(C_j))+\beta*d(v_i,\overline{R^{\lambda}}(C_j)), & \text{其他}\end{cases} \tag{17}$$

其中，α 和 β 分别是社区下近似集和上近似集距离度量的权重，满足条件 $\alpha+\beta=1$ 且 $\alpha>\beta$；$|C_j|$表示社区 C_j 的节点个数；$d(v_i,\underline{R^{\lambda}}(C_j))$和 $d(v_i,\overline{R^{\lambda}}(C_j))$是每次迭代中的动态距离，分别表示节点 v_i 到社区下近似节点集的 Jaccard 距离之和及其到社区上近似节点集的 Jaccard 距离之和。

3.3 对“伪”重叠节点的处理

对边界域节点进行迭代处理后，此时识别出来的重叠节点有部分是非重叠节点。产生这种现象的原因是这类节点离所有的核心节点较远，且不满足比值关系。我们称这类节点为“伪”重叠节点。“伪”重叠节点存在两种情况：1）只有一个邻居节点的叶子节点；2）该节点有多个邻居节点，且该节点只有一个邻居节点属于一个社区，大于或等于 2 个邻居节点属于另外一个社区。具体如图 2 所示。我们对这类节点的处理分两种情况，分别如图 2(a)和图 2(b)所示。为此，本文提出了两条对“伪”重叠节点进行处理的规则。



(a)伪重叠节点情形 1　　(b)伪重叠节点情形 2

图 2 “伪”重叠节点

Fig. 2 “Pseudo” overlapping nodes

规则 1 当“伪”重叠节点为只有一个邻居节点的叶子节

点时，该类节点被划分到其邻居节点所属的社区中。如图2(a)所示，节点{a,b,c}为社区 C，节点{d}为“伪”重叠节点，此时应该将节点{d}划分到社区 C。

规则2　当“伪”重叠节点为有多个邻居节点，且该节点只有一个邻居节点属于一个社区时，大于或等于2个邻居节点属于另外一个社区。如图2(b)所示，节点{a,b,c}为社区 C_1，节点{e,f,g,h}为社区 C_2，节点{d}为“伪”重叠节点，根据节点的链接紧密程度，此时应该将节点{d}划分到社区 C_2 中。

3.4　OCDRDD 算法框架

结合粗糙集理论和距离动态模型的方法，本文提出了一种重叠社区发现算法，解决了重叠节点的挖掘和社区动态距离变化的问题。

OCDRDD 的主要思想是首先通过度中心性和节点核心距离的乘积挑选出 K 个排序靠前的核心节点，然后根据定义的社区近似集比值关系初始化社区近似集，将非中心节点划分到由核心节点扩展的下近似集和上近似集，并求出每个社区的边界域。针对边界域节点，在每次迭代中都固定社区近似集比值关系，阈值取为最小值1.1，以保证非中心节点能被正确划分到社区下近似集。根据距离动态模型，边界域节点与社区下近似里的节点直接相连的边的距离在动态变化，所以边界域的节点到社区的距离也在动态变化，边界域在不断变小，最终达到一个稳定状态。每次迭代过程中都计算相应的评价指标 NMI，直到取得较好效果的社区结构，最后再处理“伪”重叠节点。OCDRDD 算法的大致过程如下：

1)由式(12)计算出节点中心值，排序选出前 K 个核心节点 $Core=\{v_i,v_2,\cdots,v_K\}$；

2)根据定义11定义的比值关系初始化由核心节点扩展的社区近似集，求出边界域，并计算此时的重叠社区评价指标 $index_1$；

3)针对边界域，依据距离动态模型更新边界域节点到社区下近似节点直接相连的边的长度，根据社区近似集的定义将边界域符合比值关系的节点划分到社区下近似集中，计算此时的重叠社区评价指标 $index_2$；

4)如果 $index_2>index_1$，则算法终止，否则继续迭代步骤3)，并将下一次的评价指标 $index_2$ 赋值给 $index_1$；

5)针对边界域的“伪”重叠节点，依据3.3节的两条规则进行处理。

OCDRDD 算法伪代码的详细描述如算法1所示。

算法1　OCDRDD 算法

Input：复杂网络 G=(V,E)，社区个数 K，凝聚力参数 λ，上、下近似集权值 α，β

Output：划分的重叠社区结构 $C=\{C_1,C_2,\cdots,C_K\}$，此时 $C_k=\{v_k\}$

1. For 节点 $v_i\in V$ do
2.　　利用式(12)计算节点 v_i 的节点中心值 ζ_i；
3. End for
4. 根据节点中心值 ζ_i 将节点降序排列，选择前 K 个节点作为核心节点 $Core=\{v_i,v_2,\cdots,v_K\}$；
5. For $v_j\in\{V-Core\}$ do
6.　　利用式(17)计算非核心节点 v_j 到由核心节点形成的社区 $C=\{C_1,C_2,\cdots,C_K\}$ 的距离 $d(v_j,C_i)$；
7. End for
8. 由定义11定义的距离比值关系初始化由核心节点扩展的社区近似集，初始化的社区结构为 $C=\{C_1,C_2,\cdots,C_K\}$，此时 $C_k=\{v_k\}$，所有社区的边界域为 Bnd(C)，计算此时的重叠社区评价指标 $index_1$；
9. 初始化变量 $index_2=index_1$；
10. Do{
11.　　$index_1=index_2$；
12.　　for $v_i\in$ Bnd(C) do //识别社区的重叠节点，边界域节点在变少
13.　　根据式(5)－式(8)计算边界域节点 v_i 与社区下近似节点 v_j 直接相连的边 e_i 的 $DI(v_i,v_j)$，$CI(v_i,v_j)$ 和 $EI(v_i,v_j)$，根据距离动态模型式(9)计算边 e_i 变化后的距离，边初始值距离为式(4)定义的 Jaccard 距离；
14.　　End for
15.　　根据定义11定义的距离比值关系，将符合条件的边界域节点划分到社区的下近似集中；
16.　　计算此时的重叠社区评价指标 $index_2$；
17. } while($index_1<index_2$)；
18. For $v_i\in$ Bnd(C) do
19.　　根据3.3节对“伪”重叠节点的处理识别出“伪”重叠节点 v_i，并根据两条规则处理“伪”重叠节点；
20. End for

4　实验结果与分析

在真实数据集和人工网络数据集上将 OCDRDD 算法与近几年的社区发现算法进行对比，以测试本文算法的有效性和可行性。参与对比的代表性社区发现算法分别是 CDRS[13]，LDC[16]，RFC[8]，LinkGraph[11]和 DCN[17]。

4.1　实验数据

本文选用的测试数据是 Mark Newman 提供的网上真实网络数据集[18]和 LFR Benchmark 网络数据集[19]。

1)真实网络数据集

本文选用了6个真实网络数据集进行验证，分别是 Zachary 美国大学空手道俱乐部网络(Karate)[20]、新西兰 Doubtful Sound 海峡海豚关系网络(Dolphins)[21]、美国政治书籍网络(Polbooks)[22]、美国足球联赛网络(Football)[23]、美国博客政治倾向网络(Polblogs)[24]和小说悲惨世界的人物关系网络(Lesmis)[25]。这6个真实网络数据集的描述如表1所列。

表1　真实网络数据集

Table 1　Real network datasets

Network	N	E	C
Karate	34	78	2
Dolphins	62	159	2
Polbooks	105	441	3
Football	115	613	12
Polblogs	1490	19090	4
Lesmis	77	508	6

其中，N 表示网络的节点数，E 表示网络中边的数目，C 表示网络中社区的数目。

2)LFR 基准网络数据集

LFR 基准网络数据集是由 LFR 生成工具生成的，该工具可以根据需求生成满足不同要求的模拟复杂网络。该数据

集不仅满足复杂网络的性质，而且具有已知社区结构，对评价算法的社区质量提供了很好的帮助。

LFR 基准网络数据集需要设置一些参数，以生成满足要求的复杂网络。其中参数 N 表示社区节点数目，参数 k 表示节点的平均度，参数 γ 表示节点分布参数，参数 β 表示社区大小分布参数，参数 k_{max} 表示节点最大度，$minc$ 表示社区包含的最小节点个数，$maxc$ 表示社区包含的最大节点个数，μ 表示节点与社区外部连接概率的混合参数。如果要生成重叠社区，还包含每个重叠节点连接社区的数目参数 O_m 和重叠节点比例参数 O_n。本文将固定参数 $N=1\,000$，$k=10$，$\gamma=2$，$\beta=1$，$k_{\max}=50$。其他参数，如 $minc$ 和 $maxc$ 表示社区的规模大小，μ 表示网络的混合复杂程度，O_n 和 O_m 表示重叠节点的比例情况，这些参数对网络生成的影响较大。为了测试这些参数对算法的影响，生成 8 组人工网络数据集，并且每组将参数 O_m 设置为 2～8。生成的人工网络数据集的描述如表 2 所列[26]。

表 2 人工网络数据集信息

Table 2 LFR benchmark network datasets

Network	N	k	$k_{\max}$	μ	$maxc$	$minc$	O_n
LFR1	1 000	10	50	0.1	50	10	100
LFR2	1 000	10	50	0.1	50	10	500
LFR3	1 000	10	50	0.3	50	10	100
LFR4	1 000	10	50	0.3	50	10	500
LFR5	1 000	10	50	0.1	100	20	100
LFR6	1 000	10	50	0.1	100	20	500
LFR7	1 000	10	50	0.3	100	20	100
LFR8	1 000	10	50	0.3	100	20	500

4.2 评价指标

采用改进的标准互信息 NMI[27] 和具有重叠性的模块度 EQ[28] 作为本文重叠社区发现算法的评价指标。对于不同的网络数据集，我们选用不同的评价指标。

1)真实网络数据集的评价指标

对于 Mark Newman 提供的真实网络数据集，本文采用常用的重叠社区评价指标——具有重叠性的模块度 EQ。EQ 的定义如下：

$$EQ=\frac{1}{2|E|}\sum_{w\in C_j}\sum_{v\in C_i}\frac{1}{N_vN_w}(\boldsymbol{A}_{vw}-\frac{K_vK_w}{2|E|}) \tag{18}$$

其中，$|E|$ 表示网络中边的数目，N_v 和 N_w 分别表示节点 v 和 w 属于的社区个数，$\boldsymbol{A}_{vw}$ 表示网络的邻接矩阵，C_i 和 C_j 分别表示第 i 个和第 j 个社区，K_v 和 K_w 分别表示节点 v 和 w 的度。

2)人工网络数据集的评价指标

对于由 LFR 合成的人工网络数据集，我们采用改进后的标准互信息 NMI 作为评价指标。NMI 是一种度量两个集合间差异的信息论方法。假设算法划分的社区结构用 Y 表示，任意节点 $y_i\in Y$ 可以用长度为 K 的向量 $\vec{m_i}$ 表示，其中向量 $\vec{m_i}$ 的分量取值为 1 表示节点 y_i 属于第 i 个社区，否则取值为 0。将 $\vec{m_i}$ 的第 k 个分量用随机变量 Y_k 表示。真实划分的社区结构用 X 表示，节点 $x_i\in X$ 被划分到第 r 个社区的概率为 X_r。随机变量 Y_k 在所有 X_r 上的条件熵定义如下：

$$H(Y_k|X)=\min_{r\in\{1,2,\cdots,R\}}H(Y_k|X_r) \tag{19}$$

则所有 $Y_k(Y)$ 在 X 上的规范化条件熵定义为：

$$H(Y|X)=\frac{1}{|K|}\sum_{k=1}^{K}\frac{H(Y_k|X)}{H(Y_k)} \tag{20}$$

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

$$NMI(Y|X)=1-\frac{H(Y|X)+H(X|Y)}{2} \tag{21}$$

4.3 参数设置及实验环境

OCDRDD 算法的社区个数参数 K 由网络的真实划分社区个数来设定；距离动态模型中的凝聚力参数 λ 的取值区间为[0.4,0.6][10]；社区近似集中上、下近似值 α 和 β 的取值满足条件 $\alpha>\beta$ 和 $\alpha+\beta=1$，本文中 $\alpha=0.7$，$\beta=0.3$。CDRS 算法中的社区个数参数 K 也是由网络的真实划分社区个数来设定的，粗糙集的阈值 λ 取值为 0.909。LDC 算法中，首先计算出所有节点间的距离，再将距离升序排序，取第 2%的值作为截断距离。RFC 算法中阈值 λ_2 的取值范围是参数 λ_1 的 0.8 倍和 0.9 倍之间，表示为 $0.8\lambda_1\leqslant\lambda_2\leqslant0.9\lambda_1$，其中 λ_1 由文献[8]中的均值方法求出。

本文的实验环境为：处理器 Intel (R) Pentium(R) CPU 2117U @1.8GHz，内存 4GB，操作系统为 Windows10 64 bit。本文所有算法的编程语言均为 Java，IDE 环境为 Eclipse。

4.4 实验结果及分析

1)真实网络数据集上的测试结果及分析

将本文提出的 OCDRDD 算法与 CDRS，LinkGrap，DCN，LDC 和 RFC 算法在真实网络数据集上进行测试，并采用 EQ 评价指标对社区划分结构进行评价。测试结果如表 3 所列。

表 3 真实网络数据集的 EQ 测试结果

Table 3 EQ test results on real network datasets

Algorithm	Karate	Dolphins	Polbooks	Football	Lesmis	Polblogs
OCDRDD	**0.3715**	**0.3801**	**0.4993**	0.3585	**0.4830**	**0.2967**
CDRS	0.3434	0.3717	0.3777	**0.5510**	0.4386	0.1400
LinkGraph	0.1612	0.0524	0.0038	0.0327	0.0423	0.0038
DCN	**0.3715**	0.3787	0.4456	0.3534	0.1102	0.1269
LDC	0.2469	0.3693	0.4339	0.2415	0.1597	0.0780
RFC	0.3255	0.1869	0.4257	0.5179	0.3469	0.1044

从真实数据集上的测试分析结果可知，OCDRDD 算法除了在数据集 Football 上的 EQ 低于 CDRS 算法和 RFC 算法的 EQ 外，在其他真实数据集上均取得了最大 EQ 值，其中 DCN 算法在 Karate 数据集上也取得了最大 EQ 值。综上可知，OCDRDD 算法划分的社区结构优于其他社区发现算法的划分结构，体现了 OCDRDD 算法的有效性和可行性。

图 3 给出了 OCDRDD 算法在 Dolphins 数据集上的划分结果，蓝色的节点被划分到一个社区，绿色的节点被划分到一个社区，红色节点{30，36}为两个社区的重叠节点。

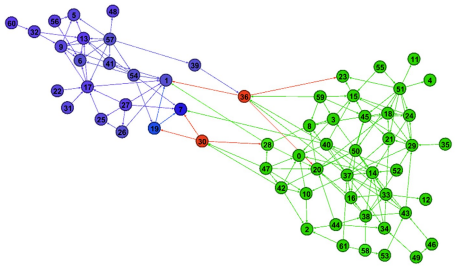


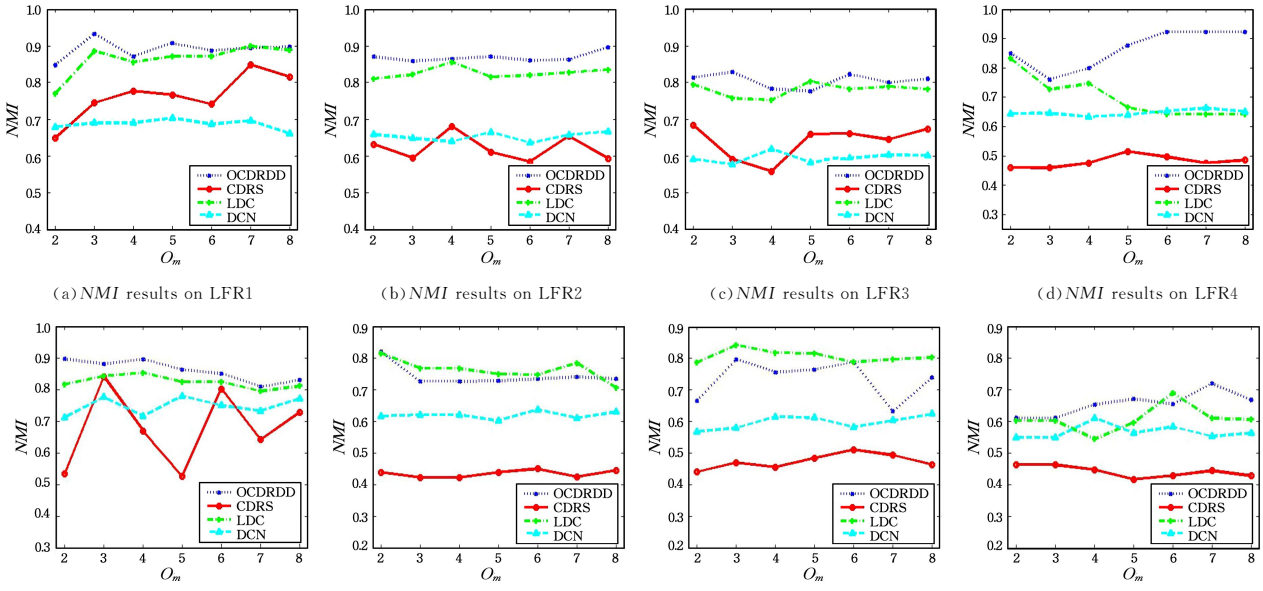
图 3　OCDRDD 算法在 Dolphins 数据集上的划分结果
（电子版为彩色）

Fig. 3　Classification results of OCDRDD algorithm on Dolphin dataset

2）人工网络数据集上的测试结果及分析

将 OCDRDD 算法与社区发现算法 CDRS，LDC 和 DCN 在上述提到的 8 组人工数据集上进行测试，评价指标采用改进后的标准互信息 NMI。测试结果如图 4 所示，其中横坐标表示参数 O_m 从 2 取值到 8，纵坐标为 *NMI* 值。

从图 4 的结果来看，随着参数 $minc$，$maxc$，μ，O_n 和 O_m 的改变，网络的复杂度由低到高，重叠节点由少到多，社区的规模由小到大。相比其他算法，OCDRDD 算法在 LFR1—LFR5 数据集上的 *NMI* 值都比较稳定，且取得了较好的效果；在 LFR6—LFR8 数据集上，LDC 算法划分的社区结构较好，OCDRDD 算法的效果仅次于 LDC 算法。因此，从网络中社区的规模大小、不同复杂网络和不同比例的重叠节点等因素来分析，本文提出的 OCDRDD 算法整体比较稳定，且具有有效性。



(a) *NMI* results on LFR1　(b) *NMI* results on LFR2　(c) *NMI* results on LFR3　(d) *NMI* results on LFR4

(e) *NMI* results on LFR5　(f) *NMI* results on LFR6　(g) *NMI* results on LFR7　(h) *NMI* results on LFR8

图 4　人工数据集上测试的 *NMI* 值

Fig. 4　NMI results on LFRs

结束语　本文提出了一种基于粗糙集和距离动态模型的重叠社区发现算法。该算法首先结合节点度中心性和距离选出 K 个核心节点。为了挖掘重叠节点，先应用定义的距离比值关系初始化社区近似集，然后迭代处理边界域的节点，结合距离动态模型动态改变边界域节点与下近似节点直接相连的边的长度，每次迭代都应用定义的距离比值关系将边界域符合条件的节点划分到社区下近似集中，以缩小边界域范围，直到找到最佳的重叠社区结构。最后，处理“伪”重叠节点。在人工网络数据集和真实网络数据集上的实验结果表明，相较于其他社区发现算法，OCDRDD 算法是有效且可行的。将来，我们将进一步研究异质网络中基于动态距离的重叠社区发现问题。

参 考 文 献

[1] COSCIA M，GIANNOTTI F，PEDRESCHI D. A classification for community discovery methods in complex networks[J]. Statistical Analysis & Data Mining the Asa Data Science Journal，2011，4(5)：512-546.

[2] KELLEY S，GOLDBERG M，MAGDON-ISMAIL M，et al. Defining and discovering communities in social networks[M] // Thai M T，Pardalos P M. Handbook of Optimization in Complex Network. New York：Springer，2012：139-168.

[3] CUI W，XIAO Y，WANG H，et al. Local search of communities in large graphs[C] // ACM SIGMOD International Conference on Management. 2014：991-1002.

[4] XIE J，KELLEY S，SZYMANSKI B K. Overlapping community detection in networks：the State of the Art and Comparative Study[J]. ACM Computing Surveys，2011，45(4)：1-35.

[5] BAI X，YANG P，SHI X，et al. An overlapping community detection algorithm based on density peaks[J]. Neurocomputing，2016，226：7-15.

[6] ZHOU X，LIU Y，WANG J，et al. A density based link clustering algorithm for overlapping community detection in networks

[J]. Physica A Statistical Mechanics & Its Applications, 2017, 486: 65-78.
[7] SUN H, LIU J, HUANG J, et al. A link-based label propagation algorithm for overlapping community detection in networks[J]. Computational Intelligence, 2016, 33(2): 308-331.
[8] WU H, GAO L, DONG J, et al. Detecting overlapping protein complexes by rough-fuzzy clustering in protein-protein interaction networks[J]. Plos One, 2013, 9(3): e91856.
[9] SUN Z, WANG B, SHENG J, et al. Overlapping community detection based on information dynamics[J]. IEEE Access, 2018, 6: 70919-70934.
[10] SHAO J, HAN Z, YANG Q, et al. Community detection based on distance dynamics[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 1075-1084.
[11] CHEN L, ZHANG J, CAI L J. Overlapping community detection based on link graph using distance dynamics[J]. International Journal of Modern Physics B, 2017, 32(3): 1850015.
[12] PAWLAK Z. Rough sets [J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356.
[13] ZHANG Y, WU B, LIU Y, et al. A novel community detection method based on rough set K-means[J]. Journal of Electronics & Information Technology, 2017, 39(4): 770-777.
[14] DWYER T. Visual analysis of network centralities[C]// Proceedings of Asia-Pacific Symposium on Information Visualization (APVIS 2006). 2006: 189-197.
[15] HENNIG C, HAUSDORF B. Design of dissimilarity measures: a new dissimilarity between species distribution areas[C]// Data Science and Classification. Berlin: Springer, 2006: 29-37.
[16] HUANG L, WANG G, 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): 15.
[17] DING J, HE X, YUAN J, et al. Community detection by propagating the label of center[J]. Physica A: Statistical Mechanics and its Applications, 2018, 503: 675-686.
[18] NEWMAN M. Network data [EB/OL]. [2013-04-19]. http://www personal. umich. edu/~mejn/netdata/.
[19] LANCICHINETTI A, FORTUNATO S. Community detection algorithms: A comparative analysis[J]. Physical Review E, 2009, 80(5): 056117.
[20] ZACHARY W. Information-flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 33: 452-473.
[21] LUSSEAU D, SCHNEIDER K, BOISSEAU O, et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations[J]. Behavioral Ecology & Sociobiology, 2003, 54(4): 396-405.
[22] KREBS V. Books about US Politics [EB/OL]. http://www.orgnet. com/.
[23] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[EB/OL]. http://arxiv. org/abs/cond-mat/0112110/.
[24] ADAMIC L A, GLANCE N. The political blogosphere and the 2004 US Election[C]// Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem. 2005.
[25] KNUTH D E. The stanford graphBase: A platform for combinatorial computing[M]. Addison-Wesley, 1993.
[26] ZHU M, MENG F R, ZHOU Y. Density-based link clustering algorithm for overlapping community detection[J]. Journal of Computer Research and Development., 2013, 50(12): 2520-2530.
[27] 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.
[28] NICOSIA V, MANGIONI G, CARCHIOLO V, et al. Extending modularity definition for directed graphs with overlapping communities[OL]. https://wenku. baidu. com/view/f1432a2d0066f5335a81219b. html.

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 member of China Computer Federation. Her main research interests include granular calculation, rough sets and intelligent information processing.