

一种基于可变半衰期的机会网络社团兴趣值更新策略

严禹道 刘林峰

(南京邮电大学计算机学院 南京 210023)

(江苏省无线传感网高技术研究重点实验室 南京 210003)

摘要 为了掌握节点移动的历史信息,提出了MDIR (Mass-group Detected by Interest-value)算法。该算法引入了社团的概念,将节点的移动规律与其它节点的关系进行关联。理论上,在社会网络中节点的移动可以归结为在不同社团中移动的过程。因此在该算法中,消息更倾向于向包含目标节点的社团转发。此外,考虑到社会关系的动态性,算法还引入兴趣值概念来更新网络拓扑中的社团结构。实验测试数据表明,在不同的节点密集度和网络资源有限的情况下,相较于Epidemic、BDCR以及SREP算法,MDIR算法可通过计算效用值进行路由转发来产生较低的转发能耗以及稳定的送达率。

关键词 机会路由算法,社会网络,社团结构,兴趣值

中图分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.2.028

Update Strategy of Communities Information Based on Variable Half-lives in Opportunistic Networks

YAN Yu-dao LIU Lin-feng

(School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

(Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing 210003, China)

Abstract Due to the incomplete use of historical information of nodes' movements, to introduce the concept of group, every single person can be put into the connect of the community. Theoretically, the movement of nodes can be abstracted as movement from one group to another. We proposed a new method mass-group detected with interest value (MDIR) to compute the utility value in opportunistic routing in social network. With the concept of interest value, dynamic communities can be detected precisely. Experiments indicate that in different intensiveness, MDIR performs better stability and efficiency compared with Epidemic, BDCR and SREP.

Keywords Opportunistic networks routing, Social networks, Community structure, Interest value

1 概述

随着大量的移动便携式设备的出现,机会网络(Opportunistic Network)也慢慢成为一个新的研究热点。其与传统的移动热点网(Mobile Ad hoc Network, MANET)的不同之处在于它对延迟的容忍特性,因此也是一种延迟容忍网(Delay Tolerant Network, DTN)^[1]。机会网络不需要源节点和目的节点之间存在完整路径,是一种利用节点移动带来的相遇机会实现通信的新型网络体系结构,节点以“存储”、“携带”、“转发”的通信模式灵活地实现了节点之间的数据传输。机会网络可以以动物、车辆、人等具象移动节点为载体,而其他载体不同,每个人在不同社会环境中都扮演着特定的角色。因此以人为载体的机会网络节点移动应当表现出社会网络的特性^[2]。

一些经典的应用在DTN网络中的算法,比如PROPHET^[3]是基于对节点移动模型规律的研究,并提出对将来移动的预测。文献[4]提出了一种基于贪婪的分布式路由转发协议(Beam width and Direction Concerned Routing, BDCR),

通过这些估计转发节点与目标节点的位置关系,得到一个大致转发方向。文献[5]提出了一种在网络中将地理位置抽象成社区,并在一定周期更新节点的社区信息的算法(Social Relationship Enhanced Predicable, SREP),该算法将整个网络的地理位置划分为多个社区,同时每个节点保存自己的社区信息和对应其它节点的社会关系强度。与陌生人相比,人更容易和亲戚、朋友等有一定社会关系的个体接触,这样,每个人便形成了自己生活的规律——“社团”。通过观察以人为载体的机会网络拓扑结构,可以得到体现人社会特性的社会网络社团结构,如图1所示。由于社团内部结构会随着时间的推移而改变,比如社团成员的加入和离开,这一特性会增加社团结构用于分析机会网络的复杂度,从而导致应用社团结构进行机会路由的效率大大降低。社团的拓扑结构体现了人与人之间的社会关系,并且暗示了网络中节点移动潜在的规律性。但在社会网络中社团是动态变化的,一个有效的社团拓扑结构,解决了哪些特定节点可能发生相遇的问题(当前存在哪些社团),但是如要将其应用到机会路由的算法中,还需要解决社团内部成员的地理上相遇的时间间隔问题(哪些社团

最近最有可能发生相聚)。文献[6]提出了一种在复杂网络中通过对节点聚类关系研究来检测社团拓扑结构的方法,这种方法针对图中节点的度和社团的规模进行社团的静态检测。

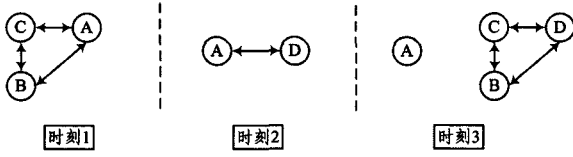


图1 不同时刻节点参与不同的社团聚集

对机会网络中人的移动行为的预测类似于对人的其它行为预测,目前已有的数据挖掘算法中的推荐系统概念通过人之间的相关性,把人从一个孤立的节点抽象成了整体环境里的一部分,从而更准确地把握人的行为^[7]。本文提出了一种基于兴趣值的社团更新策略(Mass-group Detected by Interest-value, MDIR),通过检测社团内部成员的聚集并加入时间参数更新节点对社团的兴趣值,再用兴趣值计算可以体现一定时间内发生相遇的概率值作为机会转发的效用值。本算法旨在利用人在社会关系中的依赖性和稳定性,解决对社团内部节点相遇时间的预测,得到一个较为准确的机会路由效用值的参考。

2 相关研究

在机会网络中,主流的路由转发算法分为:基于冗余、基于效用两个基础策略。基于冗余的机会路由算法是把待发送消息扩散至网络中,形成多消息备份的转发策略,该策略的代表算法为 Epidemic。基于效用值的机会转发算法,总体上是通过效用值计算,为消息携带节点提供一个或多个合理的转发节点。

在算法 BDCR(Beam wedth and Direction Concerned Routing)中,每个消息携带节点默认知道本节点与目标节点的大致位置以及到目标节点的方向。节点对数据包是否转发取决于一个效用值,在此算法中称为转发因子。转发节点通过计算本节点到目标节点的距离以及将消息发给自己的“上一跳”到目标节点的距离,可得出校正因子。最后,消息携带节点决定对消息携带节点通信半径内的其它节点是否转发消息。BDCR 算法提出了一种利用节点距离和位置计算效用值的方法,涉及到并能加以利用的节点移动历史信息过于少量且简单。若要做出更加准确的效用值判断,还需要更加全面而庞大的历史数据分析。近几年,随着智能设备存储材料的发展,通信设备逐渐趋于容量大、速度快等特征。如今的通信设备完全可以更高效地满足基于更大数据量的复杂算法的硬件要求。

在算法 SREP(Social Relationship Enhanced Predicable)中,提出了一种基于地理位置的社区概念。该算法首先针对节点的运动模型分析得到网络中的社区信息,同时得出节点之间的共同社区,称为节点之间的社会关系强度。经过长时间运动后,检测节点是否偏离原运动规律,并更新节点对应的社区信息。在路由转发时,消息携带节点根据社区信息和节点之间的社会关系强度来计算发往通信半径内其它节点的概率。

3 基于兴趣值的社团更新与机会转发算法

本文首先定义 3 个概念:社团 M (Mass-group),区别于社

区,社团仅代表节点的集合,每个社团中包含几个可能发生聚集的节点集 $\{V_a, V_b, V_c, \dots\}$;每个移动节点 V 也同样属于几个社团 $\{M_a, M_b, M_c, \dots\}$;兴趣值 INV (Interest Value),每个节点对它曾经相遇过的社团保留一个兴趣值,该兴趣值会随时间变化而变化^[8]。

所有节点对应的所有社团是多对多的关系,因此为了模拟整个网络的拓扑结构,在每个节点 i 内部保存了一张兴趣值表 $LIST$,其中包含 K 个块,每一个块里存放的 V_i 有着相应兴趣值的社团信息。 V_i 社团表中第 x 个块的信息 $LIST_x$ 存放信息为 $\{V_i_INV_M_j | \alpha \times (K-x) < V_{i_INV_M_j} < \alpha \times (K-x+1), M_j \in M, x < K\}$ 。取一个极小值 φ ,所有兴趣值小于 φ 的节点被删除。这样,除了被社团表删除的兴趣值小于 φ 的社团外,与 V_i 相关的社团信息便在 V_i 的社团表内被按兴趣值大小近似平均地分成了 K 份。

为了更新兴趣值,并体现时间变化对兴趣值的影响,本文借用了半衰期 HLP(Half-Life Period)的物理学概念^[9]。描述了每个节点 V_i 对它的所有所属社团兴趣值的更新——若一个 HLP 时间内某 V_i 节点没有检测到它的某个所属社团 M_j 没有发生社团聚集,则在 V_i 节点的社团兴趣表中对应的社团 M_j 兴趣值衰减。

本文算法的理论基础是从移动节点所拥有的社会性出发的。人的时间,除了独处的社团状态—— M_0 以外,剩下的时间便被他所参与的社团占有。因此若有节点 V_i ,在 T_0 时刻产生一个发往 V_d 的信息,假设待发送信息及副本在网络中传播的有总效生存期为 Dl_{upp} ,那么本算法的目标便是将信息发往 V_d 节点从 T_0 到 $T_0 + Dl_{upp}$ 时刻内最有可能参与的社团。映射到每个节点的兴趣表中便可以描述为,将信息尽可能地发往 V_d 兴趣值高的社团。带着这样的目的性,理论上便可以在缩小网络中副本数的同时保证较高的送达率。

3.1 社团聚集检测

节点可分为携带信息和不携带信息两种状态,若节点未携带信息,则以时长 T_1 检测社团聚集,否则检测周期为 T_2 。若节点连续两次检测到相同的一个或多个节点在通信半径内,即在一个间隔周期内,这些节点保持相聚状态,则将这些节点生成一次记录,并在节点中各个块内部检查是否有某个社团与本次记录对比,通过式(1)计算本次相聚和节点已有社团 M_j 的相似度 SIM 。

$$SIM = \beta \times \frac{VS}{VE} + (1-\beta) \times \frac{VS}{VM_j} \quad (1)$$

式中, $\beta \in (0, 1)$, VE 是本次相聚实际节点个数, VM 是 V_i 节点中待计算相似度的社团节点个数, VS 是本次相遇的所有节点与待计算相似度的社团的节点交集。当 $SIM > \gamma$ 时,确定本次节点聚集为社团 M_j 的一次聚集。

3.2 节点对社团的兴趣值更新

对于节点 V_i ,若检测到一个社团聚集并与其社团表中所有社团均不相似^[10],则判断生成一个新的社团^[11],同时本次聚集的所有节点对该社团兴趣值置为初始兴趣值 $INV_0 \in (0, 1)$ 。之后, V_i 节点每次检测到社团 M_j 的聚集,则记录 V_i 节点历史中 M_j 聚集次数加 1, $FM_j = FM_j + 1$,同时更新本 V_i 对社团 M_j 的兴趣值:

$$V_i_INV_{M_j} = V_i_INV_{M_j} + \Delta \quad (2)$$

式中, Δ 为本次更新的兴趣值增量,计算方法为式(3)。

$$\Delta = (1 - V_i_INV_{M_j})^{\frac{1}{FM_j}} \quad (3)$$

其中, x 代表 M_j 在 V_i 节点社团表中 K 个历史信息块中的第 x 块。 ϵ 取值为自然数, 是决定历史聚集次数和实时兴趣值对兴趣值增量的影响程度的影响因子。节点之间的连接分为强连接与弱连接, 强连接指同事、家人等高频联系, 这也是节点应该拥有最高兴趣值的社团。弱连接范围更广, 但接触频率低, 一般节点间社会关系是朋友、同学等。在算法中需对强连接有足够的重视, 同时还要尽量保留有价值的弱连接。式(3)可以得到一个可变的兴趣值增量 Δ , 历史相遇次数越多, 体现了节点与社团越可能存在的强连接性, x 值越大则说明距离上一次聚集的时间越长。两个参数同时作用, 一定程度上利用了社会网络弱连接的特点。

节点对社团的兴趣值的衰减可以由式(4)得到:

$$INV_{M_j} = INV_{M_j} \times C^{\frac{\sigma}{h}} \quad (4)$$

式中, C 是半衰系数, 在物理学中放射性物质的半衰期 $C = 0.5$ 。 σ 为衰减因子, 取值为自然数。

当节点 V_i 未携带转发数据时, 每个监听周期 T_1 执行社团信息维护更新策略, 步骤如下。

- 步骤 1: 检查是否有到达半衰期的社团, 用式(4)对 V_i 节点中所有到达半衰期的社团更新兴趣值, 转步骤 2;
- 步骤 2: 记录本次相聚的节点号, 从第 1 块开始, 依次对各块社团执行相似度分析, 即式(1), 若没相似社团则进入步骤 3, 否则进入步骤 4;
- 步骤 3: 新建对应兴趣初值为 INV_0 的社团 M_j , 加入到 V_i 节点的对应块内, 执行步骤 5;
- 步骤 4: 在节点 V_i 的社团表中查找到社团 M_j 后, 用式(2)和式(3)更新社团兴趣值, 执行步骤 5;
- 步骤 5: 与本次社团相聚的所有节点交换对社团 M_j 的兴趣值, 同时用其它节点广播的 M_j 值信息更新 V_i 节点中 M_j 值, 至此, 社团表维护周期结束。

3.3 机会转发策略

当节点 V_i 携带消息时, 检测到周围有其它节点。节点将待转发消息的目标节点号 V_d 广播给通信半径中其它节点, 其它节点收到目标节点号后, 在所有的块中查找 V_d 以及其相关的所有社团 M_j , 其中 $j \in \{z | V_d \in M_z\}$ 。所有节点返回一个关于 V_y 到节点 V_d 的效用值 $V_y - V_d$, 计算方式为:

$$V_y - V_d = \sum_{j \in \{z | V_d \in M_z\}} V_y - INV_{M_j} \quad (5)$$

若节点 V_i 搜集到通信半径内所有关于 V_d 的兴趣总值, 则可计算出将消息转发一份拷贝副本给通信半径内相应节点 V_y 的概率:

$$FP_y = (1 - \frac{1}{h+1})^{\frac{V_y - V_d}{V_y - V_d}} \quad (6)$$

式中, h 是消息的当前跳数, 通过计算式(5)并将结果代入式(6), 可以得到节点 V_i 将消息转发给节点 V_y 的概率。

MDIR 算法动态地更新了社会网络结构中的社团信息, 通过节点对社团的兴趣值描述了节点出现在该社团中的概率, 进而得到节点之间进行路由转发的效用值。因此算法的时间复杂度直接取决于节点社团表的长度 n 。由于网络中节点可能存在反常运动, 或检测到违反节点运动规律的干扰社团, 算法应取一适当的 φ 值以约束社团表中存储社团的数量。检测到的干扰社团 M_i 在经过 m 个衰减周期后兴趣值 $V_i - INV_{M_i} = INV_0 \times C^{\sum_{x=0}^m [K - K \times INV_0] - x}$, 当 $\varphi > V_i - INV_{M_i}$ 时, 干扰社团 M_i 从社团表内删除, 即 $\log_{0.5} \frac{\varphi}{INV_0} > \sum_{x=0}^m [K - K \times$

$INV_0] - x$ 。当约束因子 φ 取 0.2 时, m 的数学期望值为 2, 即节点最多经过两个半衰周期后干扰社团便从社团表中删除。在 MDIR 算法中节点能耗与消息转发次数线性相关, 假设初始节点数为 m_0 , 区域内节点数量增加 k 倍, 则每个社团中节点数目也近似增加 k 倍, 同时社团数量增加倍数小于 k 。由于社团数量有所增加, $\sum_{j \in \{z | V_d \in M_z\}} V_i - INV_{M_j}$ 与 $\sum_{j \in \{z | V_d \in M_z\}} V_y - INV_{M_j}$ 也会相应地增加一个正值 ρ_1 与 ρ_2 。我们期望跳数 h 越大, 则转发节点的社团信息越接近目的节点, 因此根据式

$$(2) \text{和式(3)可知 } \rho_1 > \rho_2, (1 - \frac{1}{h+1})^{\frac{\rho_1 + \sum_{j \in \{z | V_d \in M_z\}} V_i - INV_{M_j}}{\rho_2 + \sum_{j \in \{z | V_d \in M_z\}} V_y - INV_{M_j}}} <$$

$(1 - \frac{1}{h+1})^{\frac{\sum_{j \in \{z | V_d \in M_z\}} V_i - INV_{M_j}}{\sum_{j \in \{z | V_d \in M_z\}} V_y - INV_{M_j}}}$, 每个消息携带节点对周围节点的消息转发概率 FP_y 会减小。由上文可知, 在节点密集程度增加时, MDIR 算法的能耗增加主要体现在社团检测时的小报文通信, 相较于其他算法转发能耗的线性增长, MDIR 在送达率约束下, 可以保持较低的能耗。

4 实验模拟及性能分析

模拟实验在两个环境中实现, 基于 ONE(the Opportunistic Networking Environment)^[12] 机会网络模拟平台。实验一基于 ONE 自带城市地图, 用 156 个节点进行测试, 实验二是在实验一的基础上将节点数增加到 300 个。在两种节点稀疏程度下分别执行 Epidemic、BDCR、SREP 以及以本文中 MDIR 算法计算结果作效用值的路由算法, 同时统计各算法运行过程中送达率、能耗等性能指标来进行比较。

4.1 模拟环境设置

实验一中, 模拟环境基于 ONE 平台中自带的城市地图, 节点总数为 156 个, 分为移动模型为 Working Day Movement^[13] 的行人节点、移动模型为 Shortest Path Map Based Movement^[14] 的汽车节点以及移动模型为 Map Route Movement^[15] 的轨道列车节点。实验二在实验一的基础上, 将节点数量增加至 300。MDIR 的参数取值如表 1 所列。

表 1 参数设置

参数名	值
K	6
D _{lupp}	300min
HLP	720min
T ₁	3min
T ₂	1min
INT ₀	0.6
β	0.4
γ	0.6
ϵ	1
σ	1
φ	0.2

4.2 实验性能指标及结果分析

图 2 是在节点稀疏时 4 项算法送达率的比较。可以看出, 消息缓存区为 200kB 左右时, 由于 MDIR 系统缓存空间基本都用来维护社团表, 转发系统不工作。而当缓存空间开始上升时, 由于社团表已经在缓存区内完善, MDIR 算法的送达率具有最大的上升速度。在消息缓冲区为 2MB 左右时, MDIR 具有最高的相对交付比率。MDIR 算法可以有效缩减副本数量, 随着缓冲区扩大, 各个算法送达率均有不同程度的增加。从图 2 可以看出 MDIR 的副本数具有最小增长幅度。

其中 Epidemic 算法送达率上升幅度最大,并且从图 2—图 4 的曲线可以看出 Epidemic 送达率的上升与其副本数的上升是成正比的。通过图 3 可以看出 Epidemic 算法占有最多的传输能耗,由于 MDIR 节点间需要经常交换历史信息,但交换的信息总量远小于一个消息的长度,而 MDIR 的信息交互是个常量,即不管是否有报文需要传递,MDIR 社团表必须保持维护。由于社团表的作用(见式(5)、式(6)),MDIR 转发系统的副本数可以约束在线性增长以内,因此能耗远小于 Epidemic,与 SREP 算法相当。

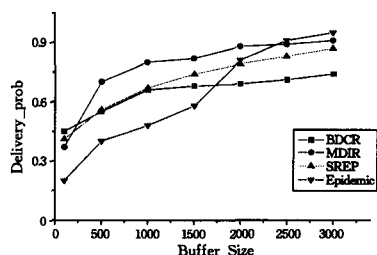


图 2 节点稀疏时不同算法对送达率的影响

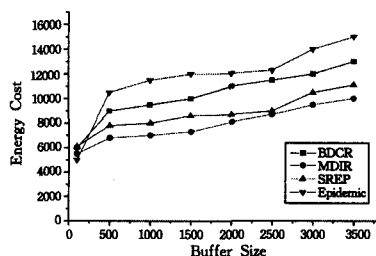


图 3 节点稀疏时不同算法对能耗的影响

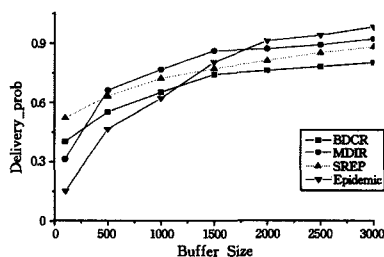


图 4 节点密集时不同算法对送达率的影响

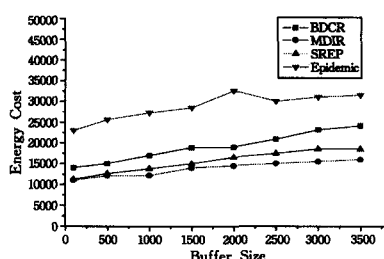


图 5 节点密集时不同算法对能耗的影响

从图 4 可以看出,Epidemic 算法在节点密集时,随着缓存区的增加送达率上升最明显。图 5 中可以看出 Epidemic 在缓存区可以运行,但略显不足是节点能耗达到峰值。图 5 的 BDCR 能耗与 Epidemic 曲线最为接近,这是由于 BDCR 的贪婪策略,即理论上每次转发是倾向发往目标节点方向的节点,同时对副本的冗余并无有效约束。可以看出在节点密集时 BDCR 的转发策略会退化为 Epidemic 广播算法。SREP 算法中由于地理位置是固定的,因此社区也是静态的。相较于 MDIR 路由表维护策略,SREP 并没有直接反映节点之间的直接关系,而是间接地研究节点地理位置信息,因此在节点转

发时会有少量偏差。体现在仿真结果里,则显示出 SREP 算法送达率比 MDIR 略低。MDIR 算法由于需要额外存储社团信息,因此在缓存区较小时送达率极低,但只要缓存区超过 1MB,根据 MDIR 的路由策略,副本数得到相应约束,便可用相对较低的能量消耗达到 0.8 以上的送达率。

结束语 本文根据社会网络拓扑结构中节点的社会社团属性,尽可能地利用相关节点运动的历史信息,借鉴数据挖掘与推荐系统的方法提出一个基于计算效用值的机会网络路由策略 MDIR。在 ONE 平台上分别运行 Epidemic、BDCR 以及 SREP 和 MDIR 算法的 JAVA 程序,统计各项性能指标。可以看出,在一个完整的社会网络体系下,MDIR 算法可以在保持传输率的同时有效降低网络中的副本数量,降低网络带宽能耗等代价,进而提高机会网络路由转发的效率。

参考文献

- [1] Kevin Fall. A delay-tolerant network architecture for challenged internets[C]// Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications(SIGCOMM'03). New York, NY, USA: ACM, 2003:27-34
- [2] Xiong Yong-ping, Sun Li-min, Niu Jian-wei. Opportunistic Network[J]. Journal of Software, 2009(1): 124-137(in Chinese)
熊永平,孙利民,牛建伟. 机会网络[J]. 软件学报, 2009(1): 124-137
- [3] Lindgren A, Doria A, Schelén O. Probabilistic Routing in Intermittently Connected Networks[M]// Service Assurance with Partial and Intermittent Resources. Springer Berlin Heidelberg, 2004:239-254
- [4] Guo Z, Peng Z, Wang B, et al. Adaptive routing in underwater delay tolerant sensor networks[C]// 2011 6th International ICST Conference on Communications and Networking in China (CHINACOM). IEEE, 2011:1044-1051
- [5] Xie X, Zhang Y, Dai C, et al. Social Relationship Enhanced Predictable Routing in Opportunistic Network[C]// 2011 Seventh International Conference on Mobile Ad-hoc and Sensor Networks (MSN). 2011:268-275
- [6] Andrea L, Santo F. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2009, 80(1Pt2): 145-148
- [7] Riedy J, Bader D A, Meyerhenke H. Scalable Multi-threaded Community Detection in Social Networks[C]// 2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW). IEEE, 2012:1619-1628
- [8] Kim J K, Kim H K, Oh H Y, et al. A group recommendation system for online communities [J]. International Journal of Information Management, 2010, 30(3): 212-219
- [9] Yuan Shuang-gui, Yang Wei-fan, Xu Yan-bing. The study of half-life theory from experimental data[J]. Science Technology and Engineering, 2004, 4(3): 179-180(in Chinese)
袁双贵,杨维凡,徐岩冰. 从实验数据看计算半衰期的理论[J]. 科学技术与工程, 2004, 4(3): 179-180
- [10] Zhang Ke-jia. Outlier nodes detection in Sensor Networks [D]. Harbin: Harbin Institute of Technology, 2007(in Chinese)
张可佳. 传感器网络中离群点的检测[D]. 哈尔滨: 哈尔滨工业大学, 2007

- [11] Xie Zhou, Wang Xian-fan. A fast algorithm for detecting local community structure in complex networks [J]. Computer Simulation, 2007, 24 (11): 82-85
- [12] Kernen A, Ott J, Krkkinen T. The ONE simulator for DTN protocol evaluation[C]//The 2nd International Conference on Simulation Tools and Techniques. Rome, Italy, 2009: 1-10
- [13] Oh S. An advanced taxi movement model in the working day movement for delay-tolerant networks[J]. Cluster Computing,

2014, 17(3): 751-756

- [14] Gewali L, Roman V. Generalization of Shortest Path Map[C]//2010 Seventh International Conference on Information Technology; New Generations (ITNG). IEEE, 2010: 296-300
- [15] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure[C]//Proceedings of the National Academy of Sciences of the United States of America. 2008: 1119-1123

(上接第 97 页)

首先,我们用 Iris 数据集验证算法。它有 150 个对象,包含 4 个条件属性和一个决策属性,决策属性对条件属性 a_1, a_2, a_3, a_4 的依赖度为 0.2133, 0.1267, 0.7800, 0.6533。使用 AP 算法对属性依赖度聚类,得到 3 类: $\{a_1, a_2\}, \{a_3\}, \{a_4\}$ 。通过选取代表属性, Iris 数据集较粗的属性粒子产生了 $\{a_2, a_3, a_4\}$ 。由此可以看出, AGAP 算法可以产生正确的属性粒化结果,并且其效率大大提高,如表 6 所列。

表 6 AR 算法与 AGAP 算法比较

Data set	Attribute granulation results		Runtime (s)	
	AR	AGAP	AR	AGAP
Iris	$\{a_2, a_3, a_4\}$	$\{a_2, a_3, a_4\}$	0.6875	0.2832
Glass	$\{a_1, a_3, a_5\}$	$\{a_1, a_3\}$	8.0756	2.0312
Identification	$\{a_6, a_7\}$	$\{a_6, a_7\}$		
Ionosphere	$\{a_{14}, a_{16}, a_{28}\}$	$\{a_1, a_4\}$	5.2116E+002	5.9639
		$\{a_{14}, a_{16}, a_{28}\}$		
Mushroom	$\{a_2, a_3, a_4\}$	$\{a_2, a_3, a_4\}$	7.2308E+003	87.3126
	$\{a_6, a_8, a_{10}\}$	$\{a_5, a_6, a_8\}$		
	$\{a_{13}, a_{14}, a_{15}\}$	$\{a_{10}, a_{13}, a_{14}\}$		
	$\{a_{16}, a_{21}, a_{22}\}$	$\{a_{15}, a_{21}\}$		

实验表明,对 Iris 数据集, AGAP 算法和 AR 算法有相同的属性粒化结果 $\{a_2, a_3, a_4\}$, 但是 AGAP 算法的效率显然高于传统的属性约简算法 AR。对于 Glass Identification 数据集, AGAP 算法与传统的属性约简算法相比丢失了属性 a_5 , 但是算法效率明显提高。对于 Ionosphere 数据集, AGAP 算法比 AR 算法多了两个属性 a_1, a_4 , 算法结果有了冗余。对于 Mushroom 数据集, 与 AR 算法相比, AGAP 算法丢失了两个属性 a_{16}, a_{22} , 又多了一个属性 a_5 , 虽然算法的结果没达到 100% 的精度, 但是算法运行时间只有原来算法的 1%。对于大数据集来说, 大数据时代降低算法的精度来换取算法的时效性具有重要意义。

结束语 本文提出的 AGAP 算法首先依据属性区分能力得到属性相似性度量, 然后使用 AP 算法对属性聚类, 最后从各属性簇类中依据信息熵方法提取代表属性, 从而达到对属性集合粗粒化的目的。算法效率相比传统属性粒化方法大大提高, 但有时精确度不高, 需要进一步寻找度量属性相似度的合适方法, 以提高算法的准确率。

参 考 文 献

- [1] Wang Guo-yin, Yu Hong, Yang Da-chun. Decision table reduction based on conditional information entropy[J]. Chinese Journal of Computers, 2002, 25(7): 759-766 (in Chinese)
- 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759-766
- [2] Xu Zhang-yan, Liu Zuo-peng, Yang Bing-ru, et al. A quick at-

tribute reduction algorithm with complexity of $\max\{O(|C||U|), O(|C|^2|U/C|)\}$ [J]. Chinese Journal of Computers, 2006, 29 (3): 391-399 (in Chinese)

徐章艳, 刘作鹏, 杨炳儒, 等. 一个复杂度为 $\max(O(|C||U|), O(|C|^2|U/C|))$ 的快速属性约简算法[J]. 计算机学报, 2006, 29 (3): 391-399

- [3] Tong C, Ding S F, Zhu H, et al. A Granularity Attribute Reduction Method Based on Binary Discernibility Matrix[J]. IJACT; International Journal Advancements in Computing Technology, 2012, 4(12): 213-221
- [4] Qian Y H. Positive approximation: an accelerator for attribute reduction in rough set theory[J]. Artificial Intelligence, 2010, 174(9-10): 597-618
- [5] Wang Sa, Zheng Lian. Feature selection method based on fisher criterion and feature clustering [J]. Computer Applications, 2007, 27(11): 2812-2840 (in Chinese)
- 王飒, 郑链. 基于 Fisher 准则和特征聚类的特征选择[J]. 计算机应用, 2007, 27(11): 2812-2840
- [6] Ma Xi-ao, Wang Guo-yin, Yu Hong. Heuristic method to attribute reduction for decision region distribution preservation[J]. Journal of software, 2014, 25(8): 1761-1780 (in Chinese)
- 马希霄, 王国胤, 于洪. 决策域分布保持的启发式属性约简方法[J]. 软件学报, 2014, 25(8): 1761-1780
- [7] Ji Su-qin, Shi Hong-bo, Lv Ya-li. An attribute reduction algorithm based on granular computing and discernibility [J]. PR&AI, 2015, 28(4): 327-334 (in Chinese)
- 冀素琴, 石洪波, 吕亚丽. 基于粒计算与区分能力的属性约简算法[J]. 模式识别与人工智能, 2015, 28(4): 327-334
- [8] Zhang Wen-xiu, Wu Wei-zhi, Liang Ji-ye, et al. Rough set theory and method[M]. Beijing: Science Press, 2001 (in Chinese)
- 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001
- [9] Frey J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976
- [10] Jia S, Qian Y T, Ji Z. Band selection for hyperspectral imagery using affinity Propagation[C]//Proceedings of the 2008 Digital Image Computing: Techniques and Applications. Canberra, ACT: IEEE, 2008: 137-141
- [11] Li G, Guo L, Liu T M, et al. Grouping of brain MR images via affinity propagation [C]// IEEE International Symposium on Circuits and Systems, 2009 (ISCAS 2009). Taipei: IEEE, 2009: 2425-2428
- [12] Dueck D, Frey B J, Jojic N, et al. Constructing treatment portfolios using affinity propagation[C]//Proceedings of 12th Annual International Conference (RECOMB 2008). Berlin: Springer, 2008: 360-371
- [13] Kelly K. Afinity Program Slashes Computing Times[EB/OL]. (2007-02-15). <http://www.news.utoronto.ca/bin6/070215-2952.asp>