基于邻居扩展的社区发现算法

吴 帆[1]
WU Fan

摘 要 诸多真实场景下的数据集可通过图形式存储与表达，图中节点构成的社区是图数据的核心指标，其结构可在多尺度下评估复杂系统的粗粒度。但社区检测仍面临挑战，现有算法多依赖全局模型或计算开销较高。为此，文章提出了一种基于邻居扩展的社区发现算法。通过邻居扩展的方式从高度数节点开始向外扩展社区，同时通过边平衡的方式防止邻居扩展的结果陷入局部最优解。并且基于邻居扩展的社区发现算法与现有的多种经典的社区发现算法在多个真实数据集上进行了对比实验。实验结果表明，在不需要其他额外信息只需要局部节点信息的情况下，基于邻居扩展的社区发现算法拥有更低的计算开销而且准确性在数个数据集上也能取得较高的分数。

关键词 数据集；多尺度；粗颗粒

doi：10.3969/j.issn.1672-9528.2025.11.016

0 引言

诸多真实场景的数据集，核心是由多个对象及对象间的关联构成的集合，基于该集合可进一步构造出对应的图结构。目前，图结构已经成为模拟复杂系统的一种基本方法，重点是将数据集中对象变为图中的点，对象之间的关系变为图中的边。例如，在社交网络图的数据中，图中点和边表示社交网络中每一个个体以及这个个体的社交关系。但在每个数据集构建起的网络中，现实世界的数据往往表现出群体性和不均匀性，在这些不均匀的分布中，在空间表示中的高密度区域，称为集群，或者网络中的高密度子图，称为社区。社区是许多网络的属性，其中一个网络可能有多个社区，社区中的节点是密集连接的多个社区间的节点可以互相重叠。集群或社区结构提供了复杂系统中底层的粗粒度表示[1]，通常与数据中不同的功能高度关联并影响其集体行为[2]，因此关于社区发现的无监督检测方法在数据科学的不同领域中至关重要。

对此，文中提出了一种基于邻居扩展的社区发现算法（以下简称 NCD 算法），这种算法通过邻居扩展的方法从高度数的节点出发逐步扩展邻居节点，同时保证划分的多个社区之间大小大致均等。该算法使用纯粹的局部拓扑信息和广度优先搜索，并在线性时间内运行，该算法不需要依赖于全局模型的目标函数的启发式优化[3]或计算成本高的扩散动态[4]，此外该算法也不依赖于相似性度量[5]。还和多个经典的社区发现算法在多个具有已知地面真实社区标签的标准数据集上进行了对比。

1 算法分析

1.1 算法背景介绍

本文算法是一种基于邻居扩展的社区发现算法，只需节点的局部信息，并依据邻居节点的信息进行扩展，同时为防止陷入局部最优解的困境，选择邻居扩展时，计算将它加入通过边平衡的方法来确保各个划分的各个社区之间大致均等。

接下来，将以图 1 中的图结构为例来详细介绍基于邻居扩展的社区发现算法的各个步骤，其中图 1 中各个节点的不同颜色代表的是它们属于不同的社区，将图 1 中的图命名为 $G(V,E)$，其中 V 是图 G 中所有的节点组成的集合，而 E 则是图 G 中所有的边组成的集合。

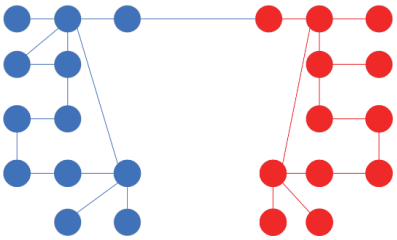

图 1 图结构示例图

1.2 算法步骤解析

本文算法的第一步如图 2 所示。从图中最高度节点中随机选取一个作为起始节点，同时定义社区边界集 S 与社区核心集 C。图 2 中矩形透明框对应社区边界集 S 的范围，图 3 中椭圆无色透明框对应社区核心集 C 的范围——S 中包含社

1. 贵州财经大学信息学院 贵州贵阳 550025

区备选节点，C 中为已完成社区划分的节点，后续邻居扩展均围绕该起始节点展开。

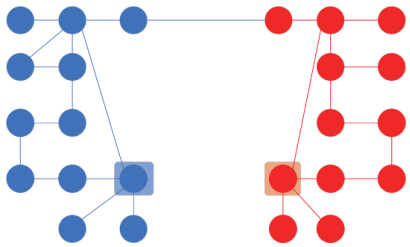

图 2 算法步骤一

确定 S 和 C 后，启动邻居节点扩展：将待划入核心集的节点设为候选节点 x，候选集合为 S/C。若 $S/C=\emptyset$（即边界集 S 中所有节点均已划入核心集 C），则从 V/C（图中未划入核心集的剩余节点）中选取度数最大的节点作为 x（图 2~3 即执行此操作）；若 $S/C\neq\emptyset$，则按式（1）在图 G 中筛选候选节点 x。

$$x=\arg\min_{v\in s/c}\left|N(v)/S\right| \tag{1}$$

式中：$N(v)$ 表示将节点 v 加入核心集后，节点 v 的邻居边会有多少条被纳入边界集 S。意味着需选择能使最少邻居边进入边界集 S 的节点 v 作为候选节点 x，确定后再将候选节点 x 的邻居节点加入边界集 S，即图 3 到图 4 所示的过程。

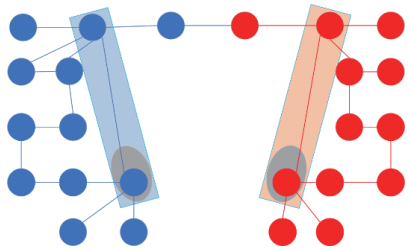

图 3 算法步骤二

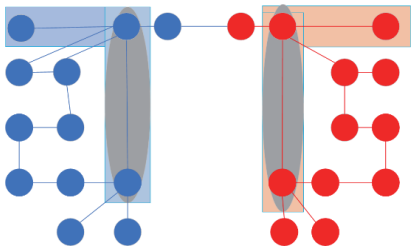

图 4 算法步骤三

接下来，迭代运行步骤二到三，直到划分的社区中的数 E_i 大于图 G 总变数除以划分的社区数量 p，以确保最后发现的社区大小大致相等。

$$|E_i|>\frac{E}{p} \tag{2}$$

为了便于呈现，将算法的运行过程以伪代码的形式呈现在图 5 中。可以看到本文算法复杂度为 $O(E)$，算法的复杂程度只与图中边的规模有关系与图中点的规模无关。

Algorithm 1:

Data: Graph G=(V,E)
Result: Digraph D
$C, S, E_k \leftarrow \emptyset$;
$\delta \leftarrow \alpha m/p$;
while $|E_k < \delta|$ **do**
  IF $S/C = \emptyset$ then
    x is the max degree node in $|V/C|$
  ELSE
    $x \leftarrow \text{argmin}_{v\in S/C}|\text{N(v)}/S|$
  AllocCommmunity(C, S, E_k, x)
end

图 5 算法的运行过程伪代码

2 实验与分析

2.1 实验数据集与评价指标

本文基于邻居扩展的社区发现算法，选用现实中带有真实社区标签的数据集作为基准测试网络。真实数据集的具体信息如表 1 所示。其中，N 为网络中的节点数量；E 为边的数量；k 为网络的平均度；d 为所有节点对之间的平均最短路径长度；cc 为平均聚类系数；ρ 为聚类率；α 为度分布的幂指数（仅当度分布可通过幂律合理拟合时适用）。其中，football 数据集的度分布无法用幂律拟合。

表 1 数据集表格

	N	E	K	d	cc	ρ	α
Karate	34	78	4.59	2.443	0.256	-0.476	1.781
Football	115	613	10.66	2.397	0.407	0.162	—
Polbooks	105	441	8.40	2.841	0.348	-0.128	1.791
Polblogs	1 222	16 717	27.36	2.747	0.226	-0.221	1.415
Cora	2 485	5 609	4.08	5.738	0.117	-0.055	1.645
Citeseers	2110	3 668	3.48	10.527	0.171	-0.024	2.074
Pubmed	19 717	44 327	4.50	2.764	0.060	-0.044	2.227

本文算法与多种经典的社区发现算法都将在表 1 所示的多个数据集上来进行社区发现运算，并将各个算法的运行结果与数据集中真实社区标签结果以 F_1 分数的方式进行对比。由于本文算法是使用 Python 算法实现的，所以其余经典社区算法都以 Python 中 networkx 来进行实现。

其中 F_1 分数是指精确率和召回率的调和平均数，因 F_1 分数具有调和平均数的特性，所以 F_1 分数对极值更敏感，只有在精确率和召回率都具有较高的分数时 F_1 分数才会具有较高的数值，所以 F_1 可以用来很好地平衡精确率与召回率二者的重要性。

所以 F_1 分数常常被当作评估分类模型性能的综合指标，尤其在评价数据不平衡或需要平衡精确率与召回率的场景中具有极高的参考价值。

2.2 实验结果分析

将 NCD 算法与 10 个经典的社区发现算法在这 7 个数据集上进行运算，并将运算结果与数据集中的真实社区标签的结果计算出 F_1 分数，各个算法的时间复杂度与运算后的 F_1 分数展示在表 2 中。

表 2 实验结果

	Karate	Football	Polbooks	Polblogs	Cora	Citeseers	Pubmed	时间复杂度
Spin glass[6]	0.61	0.92	0.62	0.88	0.33	0.22	0.21	NA
GN[7]	0.59	0.84	**0.8**	0.74	0.32	0.23	0.18	$O(N^3)$
GDG[8]	**0.91**	0.6	0.75	0.66	0.29	**0.63**	0.19	$O(\mathrm{d}tN^2)$
Walktrap[9]	0.51	0.88	0.79	0.88	0.29	0.14	0.16	$O(N^2\log N)$
Spectral[10]	0.62	0.54	0.7	0.89	0.32	0.25	0.46	$O(N^2\log N)$
Inferential[11]	0.65	0.87	0.77	0.32	0.31	0.4	0.07	$O(N^2\log N)$
Fastgreedy[12]	0.75	0.56	0.78	0.89	0.39	0.28	0.32	$O(N\log N)$
Infomap[13]	0.76	**0.96**	0.69	0.8	0.07	0.04	0.01	$O(N\log N)$
LPA	0.88	0.79	0.69	**0.91**	0.22	0.11	0.18	$\sim O(E)$
Louvain	0.63	0.87	0.7	0.85	0.32	0.27	0.2	$\sim O(E)$
NCD	0.74	0.4	0.59	0.63	**0.37**	0.36	**0.43**	$O(E)$

在表 2 中，F_1 得分最高的算法以粗体突出显示，其次以下划线突出显示。NCD 算法在拥有最低算法复杂度的同时，性能较好，在 7 个数据集运算结果中，两个数据集的结果排名第一，一个数据集的结果排名第二。在对比中其他算法往往还拥有复杂的计算方法。以 GDG 算法为例，在 GDG 算法中首先基于节点之间的最短路径长度将网络嵌入到向量空间中，然后应用迭代聚类算法，GDG 算法类似于均值移位算法[14]，而迭代聚类算法和均值移位算法为了获得社区的分区在计算上都是非常昂贵的。因此，GDG 算法在 7 个数据集中两个数据集取得最好的性能，而且在 Citeseers 上与其他算法有较大的差距。

而 Louvain 算法由于其良好的性能和高效率而被广泛应用，Louvain 算法与本文算法都是线性时间复杂度为 $O(E)$ 的算法，其中 E 是图中边的数量。相比其他算法复杂度较高的算法，更适合大规模的数据集网络。所以本文算法能够在公平的运行时间下与 Louvain 算法进行更有意义的比较。

此外，本文算法相较于 Louvain 算法在 7 个数据集上在 4 个数据集上有着更优秀的性能，而且在 Cora、Citeseers、Pubmed 共 3 个较大的数据集中也能有较强的精准度。这是因为在更大的社区网络图中，高度节点更有可能成为社区连接中心，本文算法可以通过规避孤立节点作为噪声节点的影响从而在大型的社交网络图中依然保持较高的精准度。

NCD 算法在 Football 数据集中 NCD 算法表现不佳的原因是 Football[15] 数据集是非常同质的，NCD 算法选择高度节点作为扩展节点的原因是基于节点之间的平均路径长度由连接概率决定这个前提来设计的。而同质性较高的 Football 数据集，使得节点间的平均路径长度偏向均等与连接概率相关性下降，所以同样的情况也出现在了平均最短路径长度 $\langle k\rangle$ 最大的 Polblogs 数据集中。

3 总结

本文提出了一种新的基于邻居扩展的社区发现算法 NCD 算法，NCD 算法使用纯粹的局部拓扑信息和广度优先搜索，并在线性时间内运行。还在 Python 中实现了多种经典的社区发现算法并将 NCD 算法与这些经典的社区发现算法在 7 个经典的基准网络数据集中进行了性能比较。

NCD 算法在 7 个数据集基准网络的两个数据集的结果中排名第一，在一个数据集中排名第二，发现 NCD 算法在规模较大的数据集网络中表现更好，并在实验结果分析中，分析了在剩下数据集中表现不佳的原因在于同质性较强的网络数据集会影响 NCD 算法的性能。

参考文献：

[1] FORTUNATO S, HRIC D. Community detection in networks: a user guide[J]. Physics reports, 2016, 659:1-44.

[2] CLAUSET A, MOORE C, NEWMAN M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature,2008,453:98-101.

[3] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008: 10008.

[4] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical review E, 2007, 76(3): 36106.

[5] RAVASZ E, SOMERA A L, MONGRU D A, et al. Hierarchical organization of modularity in metabolic networks[J]. Science, 2002, 297(5586): 1551-1555.

[6] REICHARDT J, BORNHOLDT S. Statistical mechanics of community detection[J]. Physical review e—statistical, non-linear, and soft matter physics, 2006, 74: 16110.

[7] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the national academy of sciences, 2002, 99(12): 7821-7826.

[8] MAHMOOD A, SMALL M, AL-MAADEED S A, et al. Using geodesic space density gradients for network community detection[J]. IEEE transactions on knowledge and data engineering, 2016, 29(4): 921-935.

[9] PONS P, LATAPY M. Computing communities in large networks using random walks[C]//Computer and Information Sciences-ISCIS 2005.Berlin:Springer, 2005:284-293.

[10] NEWMAN M E J. Modularity and community structure in

基于亮度和纹理保持的红外和可见光图像融合

孙卫平[1] 豆俊奇[1] 魏京超[1] 钱程远[2] 许 健[2]
SUN Weiping DOU Junqi WEI Jingchao QIAN Chengyuan XU Jian

摘 要 图像融合是将同一场景中的多幅图像进行融合，产生信息量更大的图像。针对红外图像与可见光图像融合过程中红外目标亮度受损和图像细节较差的问题，文章提出了一种基于亮度和细节保持的融合方法。首先，对可见光图像利用同态滤波和伽马变换进行增强。其次，利用滚动引导滤波将图像分为多个细节层和一个基础层。为了更好地反映图像特征和光照强度，又将基础层划分为显著特征层和亮度层。最后，为进一步提升融合细节，使用一种强度细节和梯度细节结合的融合规则。红外显著区域采用强度细节层，背景区域采用梯度细节层。实验证明，与 GTF、VSM-WLS、CNN 等 10 种方法相比，在主观评价方面，提出的融合方法比其他几种方法亮度更亮，纹理更清晰。在客观评价方面，提出的融合方法在平均梯度、边缘强度等 5 个客观评价指标上均优于其他几种方法，特别在平均梯度、边缘强度和图像清晰度指标上比第二平均大幅提升了近一倍，分别为 72.45%，80.52% 和 78.76%。

关键词 图像融合；多尺度分解；亮度补偿；纹理增强

doi：10.3969/j.issn.1672-9528.2025.11.017

0 引言

在对弹箭等目标进行监视和测量时，通常使用红外（IR）和可见光（VIS）对其进行观测。单个光学传感器由于自身波段的限制，无法完全反映物体的特性。图像融合技术可以有效地解决这一问题。它将同一场景中不同波段的多幅图像结合起来，产生信息量更大的图像 [1]。红外成像捕获目标的热辐射信息，成像效果不易受恶劣天气的影响。然而，红外图像分辨率低，缺乏纹理细节 [2]。相比之下，可见光成像反映了物体的纹理细节，但是容易受到外部因素的影响，如天气、气候和光照 [3]。因此，红外和可见光图像融合可以反映弹箭目标的辐射信息和纹理信息，更有利于后续的目标检测 [4-5]、跟踪 [6] 和目标识别 [7]。

此前，已经提出了大量的 IR 和 VIS 图像的融合方法，主要分为四类，即多尺度变换 [8-9]、稀疏表示 [10-11]、混合模型 [12-13] 和深度学习 [14-15]。基于多尺度变换（MST）的方法通过多级滤波得到不同尺度的子带图像，然后利用一定

1. 青岛兴道新特科技有限公司 山东青岛 266400
2. 中国航天科工集团第六研究院六〇一所
内蒙古呼和浩特 010010

networks[J]. Proceedings of the national academy of sciences, 2006, 103(23): 8577-8582.
[11] PEIXOTO T P. Bayesian stochastic blockmodeling[J]. Advances in network clustering and blockmodeling, 2019: 289-332.
[12] CLAUSET A, NEWMAN M E J, MOORE C. Finding community structure in very large networks[J].Physical review E, 2004, 70: 66111.
[13] 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.
[14] COMANICIU D, MEER P. Mean shift: a robust approach toward feature space analysis[J]. IEEE transactions on pattern analysis and machine intelligence, 2002, 24(5): 603-619.
[15] EVANS T S. Clique graphs and overlapping communities[J]. Journal of statistical mechanics: theory and experiment, 2010, 2010(12): 12037.

【作者简介】
吴帆（1998—），男，贵州贵阳人，硕士研究生，研究方向：计算机应用技术。

（收稿日期：2025-05-21 修回日期：2025-11-12）