

基于节点属性与拓扑结构融合的增强型社群检测方法

胡子成

南京中新赛克软件有限责任公司

摘要: 本文提出了考虑节点属性与拓扑结构的增强型群检测方法。在现有群检测方法不能对复杂网络进行分析的前提下, 考虑节点属性与拓扑结构相结合增强群检测方法的准确性与稳定性。本文先介绍了群检测的意义与作用, 指出群检测方法目前存在的问题, 然后介绍了节点属性与拓扑结构的理论定义及其在群检测中的意义与作用, 接着对增强型群检测方法进行理论叙述, 即提出考虑节点属性与拓扑结构的群检测方法的必要原因, 介绍方法的具体细节: 包括对节点属性的计算与预处理过程, 对拓扑结构的陈述、计算与预处理过程; 最后根据节点属性与拓扑结构设计群检测算法, 提出节点属性与拓扑结构相结合的算法融合和算法优化方案。

关键词: 节点属性; 拓扑结构; 社群检测; 融合方法; 算法优化

一、引言

1.1 社群检测的重要性与应用背景

在当前的社会网络和信息系统, 对于社群检测技术的需求正在变得越来越重要。社群检测可以帮助我们更好地理解网络结构的规律, 并且在各种领域中都有广泛的应用价值。无论是在社交媒体中的用户聚类, 还是在生物信息学中对蛋白质网络的检测, 无论是网络安全中的异常检测, 还是推荐系统中的个性化推荐, 社群检测都是不可或缺的。通过识别网络中的社群结构, 我们可以更好地进行信息传播、资源管理和决策制定, 从而推动各个领域的发展和进步。

1.2 现有社群检测方法的局限性分析

现有的社群检测算法也取得了不错的效果, 但是现有基于社群检测算法检测复杂网络的实现过程仍然有很多值得完善的地方。比如缺乏考虑到各个方面, 现在的算法大部分都是只考虑了网络的部分内容, 比如只考虑了网络的拓扑结构或者只考虑了网络的节点属性内容, 对网络信息的全面化考虑不足让很多方法在检测异质网络中的社群时都是不准确, 如果涉及属性也不确定, 就更差了。针对动态网络适用性差, 现有的算法大部分不能适用于动态网络的结构变化, 对于以社交网络为代表的动态时变网络抓不住。计算量太大, 现有算法大部分都存在计算复杂性大的问题, 使得算法在大规模网络中的时效性比较差, 不利于走向真正工业应用。

二、节点属性与拓扑结构的理论基础

2.1 节点属性的定义及其在社群检测中的作用

节点属性是指网络中节点的自身属性信息, 可以是节点的年龄、性别、职业等等; 也可以是节点的其它相关信息, 比如活跃度、影响力等等。对于检测社群来说是不可缺少的, 因为节点属性能提供更多信息, 有利于对节点以及社

区进行检测。通过对节点属性的检测, 就可以了解到拓扑结构中隐藏的深层社群信息, 并以此来增强社区检测的准确可靠性。另外, 还可以利用节点属性来辅助对拓扑结构的检测, 为社区检测提供更多维度。

2.2 拓扑结构的定义及其在社群检测中的作用

拓扑结构是指网络中节点与节点之间的连接关系以及网络的整体布局结构, 反映了网络中节点相互连接的方式和网络的组织形式。拓扑结构在社群检测中具有重要的作用, 一方面, 拓扑结构可以直观地展现网络中节点的聚集状态, 通过观察节点之间的紧密程度, 可以初步判断可能存在的社群边界, 为社群检测提供基本的结构框架。另一方面, 拓扑结构所蕴含的网络特征, 如度分布、聚类系数、路径长度等, 可以从不同角度描绘社群的结构特征, 帮助我们更好地理解社群的内在组织规律, 从而提高社群检测的准确性和有效性。

三、增强型社群检测方法的提出

3.1 融合节点属性与拓扑结构的必要性

在社群检测领域, 单一依赖节点属性或拓扑结构进行检测往往存在一定局限性。节点属性虽能提供丰富的个体特征信息, 但可能忽略节点在网络中的位置及连接关系对社群划分的影响; 而仅考虑拓扑结构, 又可能无法充分挖掘节点自身的属性特征对社群形成的内在驱动。因此, 融合节点属性与拓扑结构成为一种必要且有效的手段。通过将两者有机结合, 可以综合利用节点属性所蕴含的丰富语义信息和拓扑结构所反映的网络连接模式, 从多个维度全面刻画社群的特征, 从而更准确地识别社群边界、揭示社群内在组织规律, 提高社群检测的精度和可靠性。

3.2 增强型社群检测方法的理论框架

增强型社群检测方法的理论框架主要基于多源信息融合

的思想,将节点属性所蕴含的语义信息与拓扑结构所反映的网络连接模式进行深度整合。具体而言,该框架包含三个核心层次:首先,在数据融合层,通过特征提取与降维技术,将高维的节点属性数据与复杂的网络拓扑数据转化为统一的可比较特征空间;其次,在模型构建层,设计基于图神经网络与属性注意力机制的混合模型,实现节点属性与拓扑结构的动态交互与协同学习;最后,在决策优化层,引入多目标优化策略,平衡社群内聚性与分离性指标,通过迭代优化算法提升社群检测的鲁棒性与可解释性。这一框架不仅克服了单一信息源的局限性,更通过跨模态信息互补机制显著提升了复杂网络环境下的社群检测性能。

四、方法实现

4.1 节点属性的量化与预处理

节点属性量化是将属性从非数值型转化为计算机能够识别的数值型,用于表示属性间的相对关系,例如基于独热编码、标签编码等将分类属性转化为数值型。标准化处理是对属性进行缩放,消除属性之间量纲的影响,保证模型训练过程的稳定性。特征选择是筛选出对社群检测结果影响较大的属性,通过统计方法、模型重要性排序等方式选择,减少数据维度,提高模型训练效率。此外,在进行模型训练和使用之前,数据预处理中还需要注意数据清洗,处理缺失值、异常值,确保数据质量,避免影响后续模型训练结果。

4.2 拓扑结构的建模与分析方法

拓扑结构的建模是社群检测的另一个关键步骤,通过将复杂的网络结构转化为数学模型,便于后续的分析 and 处理。这包括图的表示、邻接矩阵的构建、网络特征的提取等。其中,图的表示方法展示了网络中节点之间的连接关系,为后续分析提供了基础。邻接矩阵通过矩阵形式揭示了网络中节点之间的相互作用强度,为数学运算和模型训练提供了输入。另外,网络特征的提取也是拓扑结构建模的重要内容,包括度中心性、接近中心性、介数中心性等。这些指标提供了网络结构中节点的不同衡量维度,为社群检测提供了支持。在完成了拓扑结构的建模后,还需要使用图论、网络科学等方面的理论和方法对模型进行分析,挖掘网络中的社群结构和发展规律。

具体来说,图论中的连通分析,能够找出连通的子图,这些子图是可能的社区人员,网络科学中的模块度最大化,能够度量社区结构的紧密程度,最大化得到模块度,然后找到最优社区划分。随机游走谱聚类也可以运用在拓扑结构分析中,从不同的方向挖掘了网络中的社区信息,提高社区检

测的准确度和鲁棒度,这些方法在实际使用中可以根据具体的需求选择使用,也可以多种方法结合使用得到更准确全面的社区结果。

五、算法设计与优化

5.1 基于节点属性的社群检测算法

这类算法主要根据节点固有的属性信息进行社群检测。首先,收集节点的所有属性信息,节点的属性可能包括年龄、性别、职业、兴趣爱好等等。然后,通过合适的相似度度量方法(比如余弦相似度、欧式距离等)计算节点之间的属性相似度,根据计算得到的相似度矩阵,使用合适的聚类算法(比如 K-means、层次聚类等)将对属性相似度高的节点归为同一个社群。另外,还可以考虑使用机器学习中的分类算法(比如支持向量机、决策树等),对有社群标签的节点进行训练得到分类模型,然后使用训练得到的模型对新社群标签未知的节点进行分类。

在实际的应用中,基于节点属性的社区检测算法简单易懂,因为直接利用节点的属性信息,可以挖掘出该社区中节点的共同属性,但通常忽略了顶点间的拓扑结构信息,对社区识别的效果不是太理想,而且顶点属性信息不完备或者存在噪声的情况下,会进一步影响社区检测效果。为了提升社区检测的效果,可以将顶点的属性信息与拓扑结构信息结合起来进行社区检测。

在具体实现基于节点属性的特征检测算法时,要考虑处理好节点属性的数据类型、数据维数问题。不同的属性数据类型可能有差异,例如数值型、类别型等,需要采用不同方法预处理,统一数据类型,以便正确进行相似度运算。另外有时属性数据的维数可能很高,需要降低数据的维数,例如可以用主成分分析的方法对属性进行降维,提高运算效率。

社群检测效果不仅与在相似度测度法与聚类算法选择有关,还与算法中参数的选取相关,如何选取 K 值(基于 K-means 的算法)、距离阈值(层次聚类的算法)等,均需要根据社群检测任务与数据本身特点来定。使用机器学习分类算法时,机器学习的训练与验证也同样重要的是需要注重训练和验证的数据多样性、有效性以及验证方法的合理性和严格性。

5.2 基于拓扑结构的社群检测算法

这类算法关注于网络上每个节点的连接关系,即拓扑结构能够划分得到准确的社群,首先要构建网络的拓扑结构模型,明确节点之间的连接关系,即通过收集网络节点之间的关系来进行构建,例如社交网络中的好友关系、通信网络中的通讯关系等。然后利用图学中的相关概念和指标,例如度

中心性、介数中心性、接近性等来分析节点在拓扑结构中的重要程度和影响力。最后应用现有的图聚类算法，如谱聚类算法、模块度优化算法等等，将具有强关系联结的节点聚集在同一类人群中，再结合复杂网络中的社区发现理论，利用随机游走和标签传播等手段进一步发现网络中的社群结构。

在实际应用中，基于拓扑结构的社群检测算法还需考虑网络的动态变化特性。由于现实中的网络往往是随时间不断演变的，节点的连接关系会频繁地增加或减少，这就要求算法能够适应这种动态性，实时更新网络的拓扑结构模型，并快速调整社群的划分结果。为了实现这一目标，可以采用增量式计算的方法，仅对网络中发生变化的部分进行局部分分析和处理，从而避免对整个网络进行重新计算，提高算法的响应速度和计算效率。同时，还可以引入时间衰减因子等机制，对历史交互数据进行加权处理，以反映节点连接关系的时效性，使社群检测结果更加符合实际情况。

在处理大规模网络时，基于拓扑结构的社群检测算法还需关注算法的可扩展性和计算复杂度。由于网络中的节点数量可能非常庞大，传统的图聚类算法可能会面临计算资源消耗过大、运行时间过长等问题。因此，需要设计高效的算法来降低计算复杂度，提高算法的可扩展性。例如，可以采用分布式算法将大规模网络的社区发现算法任务划分成若干个子任务分配到不同算力节点上，实现大规模网络的社区发现并行计算，提高算法的运行效率。还可以采用近似算法或启发式算法，降低算法的计算复杂度，在保证社区发现的质量同时提升算法运行效率。

5.3 融合策略与算法优化

在融合方法上必须要充分考虑到节点属性和拓扑结构对社群检测的重要程度，建立起有效的运算权重分配策略。例如能够按照对不同应用领域中节点属性和拓扑结构的重要程度来给予融合操作不同的权重分配，这样能够使得融合后的社群检测结果与社群的实际情况能够更加相适应。此外，还可以将融合操作进行层次性的分配，首先对节点属性和拓扑结构分别进行社群检测，然后再将检测结果进行融合操作，以此来提升社群检测的精度。

在算法改进方面，可以针对融合算法可能出现计算量大、收敛速度低等问题来进行优化处理。例如，使用并行计算技术来提高算法的计算速度，通过分布式计算框架将算法的计算任务分配到多个计算节点上并行计算；使用启发式算法来优化算法的搜索过程，帮助算法更快地搜索到最优解；使用机器学习中的一些优化算法来优化算法的参数，如梯度下降

法、遗传算法等，从而提高算法的性能和鲁棒性。

此外，还可以在融合方法中加入反馈策略，根据融合结果是否精确，对融合过程中使用的权重机制进行调整，让算法能够在不同情况下自动适应复杂社区结构。除此之外，在算法的优化过程中，还可以加入剪枝方案，对检测社区贡献不大的边或点进行剔除，减少算法重复计算，提高算法效率。还可以针对大规模网络，采用近似算法、采样算法，在不影响算法精度的情况下降低算法的时间复杂度。

结语

综上所述，基于节点属性和拓扑结构相结合的强化社群检测通过对“节点自身因素”和“网络拓扑关系”的整合，为社群检测领域呈现了一种更为精确、高效和灵活的解决方案。通过突破传统社群检测算法在复杂网络结构上的局限性，并通过算法设计与优化的不断迭代，为社群检测提供了良好的检测性能和稳定性。随着网络数据的进一步发展和复杂性提升，该方法有望在社交网络分析、生物信息学、推荐系统等领域发挥更大的作用，引领着应用领域的研究和业务向更高质量迈进。

参考文献

- [1] 李贺;刘嘉宇;沈旺;时倩如;解梦凡. 社交参与视角下超图增强的学习趣缘社群群体检测研究[J]. 情报学报. 2024 , 43 (12):1425-1439.
- [2] 夏飞;赵新建;王恺祺;陈石. 基于图卷积网络与社群发现的异常检测方法[J]. 软件导刊. 2024 , 23 (12):58-65.
- [3] 张东林;徐建. 基于多视图证据融合的社交水军检测. 计算机应用研究, 2024(10):2939-2946.
- [4] 李宁;梁永全;张琪. 一种基于时序邻居序列的游离水军群组检测方法[J]. 计算机应用研究, 2023(03):776-785.
- [5] 张琪;纪淑娟;张文鹏;曹宁;李宁. 考虑结构与行为特征的水军群组检测算法[J]. 计算机应用研究, 2022(05):1374-1379.