计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210100023

一种面向动态科研网络的社区检测算法

蒲 实 赵卫东
复旦大学软件学院 上海 200433
上海市数据科学重点实验室 上海 200433
(spu18@fudan.edu.cn)

摘 要 科研网络是一类动态变化的异构信息网络,科研网络上的社区检测能挖掘出学术主体的所属社区并发现蕴含于科研社区中的洞察。既有的社区检测算法忽略了科研网络的动态特征和科研主体间的特殊关系,未将科研社区内部的紧密程度和社区间的关系纳入社区检测算法中予以优化,对此提出了一种基于动态科研网络表示学习的社区检测算法 DANE-CD。首先基于科研网络自编码器学习科研网络中学术主体的表示向量,然后创新性地在表示学习过程中融入了基于模块度和团队断裂带两个维度的聚类优化,最后基于堆栈自编码器构造了动态科研网络表示学习模型,同时完成了对科研网络的社区检测。在 DBLP 和 HEP-TH 两个真实科研数据集上进行了实验,实验结果显示算法在准确率、归一化互信息和模块度 3 个指标上优于既有科研社区检测算法,可以较好地完成动态科研网络下的社区检测任务。

关键词:科研网络;动态网络;社区检测;异构网络;聚类优化
中图法分类号 TP391

Community Detection Algorithm for Dynamic Academic Network

PU Shi and ZHAO Wei-dong
School of Software,Fudan University,Shanghai 200433,China
Shanghai Key Laboratory of Data Science,Shanghai 200433,China

Abstract Academic network is a kind of dynamic heterogeneous information network. Community detection on the academic network can dig out the communities of academic subjects and discover the insights contained in the community structure. The existing community detection algorithms ignore the dynamics of the academic network and the special relationship between academic subjects and do not optimize the closeness of the academic community and the relationship between academic communities. This paper proposes a community detection algorithm called DANE-CD based on dynamic academic network representation learning. Firstly,an autoencoder is adopted to represent the academic subject in the academic network. Secondly,the clustering optimization based on modularity and team faultlines is innovatively integrated into the representation learning process. Finally,a dynamic academic network representation model is constructed based on the stacked autoencoder,together with the completion of community detection in the dynamic academic network. Extensive experiments on two real-world academic datasets(DBLP and HEP-TH) demonstrate that DANE-CD is superior to the baseline methods and can detect the academic communities effectively.

Keywords Academic network,Dynamic network,Community detection,Heterogeneous network,Clustering optimization

1 引言

科研网络是一种包括了大量学术实体(学者、论文、刊物等)及其间复杂关系(合著关系、引文关系、发表刊载关系等)的大型异构信息网络[1]。对科研网络进行社区检测能发现学术实体的重要性、内在联系以及科研主题的变化[2]。科研网络是动态变化的网络,每年有大量新的学术实体和关系被添加到网络中。在科研网络动态变化的信息中挖掘出有价值的洞察并将其应用于社区检测,不仅能发现社区内学者的动态组织结构和流动情况,更能从动态变化的社区中进一步探寻学术主题的发展方向和演化规律。

社区检测一直是学术研究的热点,早期的研究聚焦于网络结构特征[3-5],后续的工作[6-8]将网络的结构信息与节点的属性信息相融合,混合的特征蕴含更丰富的信息并促进了更合理的社区划分[9]。随着深度学习的发展,基于深度学习的方法显示出其在社区检测任务上的优越性能[10]。一些方法[11-12]采用了不同的深度学习模型对网络进行表示学习,表示学习的结果被应用于聚类算法或社区嵌入算法以获得社区

到稿日期:2021-01-04 返修日期:2021-04-19
基金项目:国家自然科学基金(61671157);教育部哲学社会科学研究重大课题攻关项目(19JZD010)
This work was supported by the National Nature Science Foundation of China(61671157) and Major Project of Philosophy and Social Science Research,Ministry of Education of China(19JZD010).
通信作者:赵卫东(wdzhao@fudan.edu.cn)

检测结果。上述方法仅在静态的网络上进行社区检测，但网络是动态进化的。网络中新的节点或边的加入不仅会影响局部的社区结构，还会导致全局网络的变化[13]。如何捕获网络在动态变化中的信息及其对社区分配的影响有着重要的研究价值。虽然一些研究[13-14]以不同的方法探索了动态网络下的社区检测问题，但是它们忽略了科研网络中节点间的特殊关系、社区内部的聚合程度以及社区间的关系，一些工作将特征提取与社区检测割离，缺乏统一的聚类优化过程。

本文提出了一种基于动态科研网络表示学习的社区检测算法，简称 DANE-CD 算法。具体而言，我们将动态科研网络划分为一系列连续进化的子网络，在每个子网络中，我们首先对科研网络的结构和内容特征进行预训练，然后使用自编码器对融合了网络结构和内容特征的混合特征进行表示学习；我们从模块度和团队断裂带两个维度对表示学习予以聚类优化，并设计了一个联合的损失函数，用于模型训练；我们基于堆栈自编码器将网络表示学习拓展到动态网络的整个生命周期中。DANE-CD 算法最后将获得每个学者的表示向量、社区检测结果以及社区的中心点向量。我们在 DBLP 和 HEP-TH 数据集上进行了实验，结果证明了算法的有效性。

本文第 2 节介绍了相关工作；第 3 节给出了相关概念和问题的定义；第 4 节详细阐述了具体的算法；第 5 节进行了实验验证；最后总结全文并展望未来。

2 相关工作

传统的社区检测算法聚焦于网络的结构信息。GN 算法[3]是一种基于删除边并分裂网络的算法，该方法计算网络中所有边的边介数，删除边介数最高的边，迭代上述操作直到网络中的社区数量达到预设的值。但 GN 算法的复杂度较高，不适用于大型网络。Newman 提出了一种基于模块度的快速社区检测算法[4]，该算法采用了聚集的思想，初始将每个节点视为单独的社区，然后选择两两合并后导致模块度增量最大的两个社区进行合并，重复上述合并过程直至检测到目标数量的社区。模块度度量了社区分配的优劣，成为了社区检测中普遍使用的评价函数。Louvain 算法[5]对模块度计算和合并的过程进行了优化，进一步提升了算法的性能。

基于结构的社区检测算法忽略了节点的特征信息。文献[6]提出了一种混合科研网络结构关系和文本信息的聚类方法，该方法构建了一个融合的距离矩阵，节点对的距离由节点对的最短路径长度和节点对属性的欧氏距离加权融合得到，进而可使用任意的图聚类方法获得社区检测的结果。SToC[7]将节点的属性分为定量属性和类别属性，分别使用欧氏距离和 Jaccard 距离进行度量，并与节点间的拓扑距离进行加权融合，融合后的距离矩阵将应用于聚类算法，以获得社区检测的结果。文献[8]提出了一种考虑社区上下文关系的社区检测算法，将节点的内容和结构特征属性与节点间的上下文相似度融合，在形成的加权网络中迭代使用 Louvain 算法，以获得最佳的社区分配结果。

传统的方法不能自动地提取网络特征，深度学习通过对网络拓扑结构和节点属性的表示学习，可以抽取网络的高维、非结构化的复杂信息，在非监督任务上有优越的性能[10]。ComE[11]将社区检测、社区嵌入和节点嵌入视为一个可以互相促进的闭环任务，并将社区嵌入定义为低维空间中的多元高斯分布，在训练网络的同时优化节点嵌入和社区分配的结果，进而得到融合了高阶社区相似度的节点表示和社区标签。CNRL[12]是一个应用于社区检测和网络表示学习的统一模型，该模型同时检测每个顶点的社区分布，在训练中最大化节点与局部上下文节点的条件概率，以及最大化节点与目标社区的概率，以此获得节点对应的社区标签。

上述方法忽略了网络的动态变化信息，网络中节点和边的变化会影响网络的局部和全局结构关系。DECS[13]是一种基于多目标进化聚类的动态社区检测算法，该算法设计了一个迁移算子来保证在进化过程中节点与其最邻近的节点组合在一起，并使用了一个编码网络结构信息的基因矩阵来拓展可行解的搜索空间，最后结合遗传算法完成动态社区检测任务。dyngraph2vecAERNN[14]是一个动态网络表示学习的模型，该模型从网络中节点归属社区的变化获得启发，在自编码器中引入了长短期记忆网络以学习节点的演化特征，最终获得节点在动态网络演化中的表示。尽管这些相关工作取得了一些进展，但它们都忽略了科研网络中节点间以及社区内外的关系，缺乏将特征提取与社区检测相结合的统一优化过程。

3 问题与定义

本节将形式化定义动态科研网络及其中的社区，以及动态科研网络下的社区检测问题。

定义 1(动态科研网络) 动态科研网络$\mathbb{G}=\{G^1,G^2,\cdots,G^t,\cdots,G^T\}$由一系列连续进化的子网络构成，其中每个子网络$G^t$表示网络$\mathbb{G}$在时间点$t$下的网络快照。子网络$G^t=\{A^t,P^t,V^t,E^t\}$包括了时间点$t$下的网络中的学者节点集合$A^t$、论文节点集合$P^t$、刊物节点集合$V^t$以及 3 种节点间的链接集合$E^t$。

动态科研网络是一个典型的异构信息网络。特别地，动态科研网络中的边和节点在添加到网络中后，不会在未来的变化中消失。

定义 2(社区) 动态科研网络中的一个社区指在动态网络的所有时间步中的一个节点集合，集合内部的节点是紧密相连的，集合内部的节点与集合外部的节点是稀疏相连的。

本文仅研究了科研网络中学者节点的社区检测问题，这是因为学者节点有主观能动性，而论文和刊物节点不能排除学者节点主动建立关系。

定义 3(动态科研网络下的社区检测问题) 给定动态科研网络$\mathbb{G}$，目标是预测$\mathbb{G}$中各个学者节点在不同G^t下对应的社区标签，并由此发现$\mathbb{G}$中各个社区的结构。

4 社区检测方法

本节将详细阐述方法细节，方法的整体框架如图 1 所示。图 1(a)所示的区域对科研网络进行了结构和内容特征的预处理；图 1(b)－图 1(d)所示的 3 个区域展示了本文的 3 个主要模块，分别是科研网络自编码器模块、聚类优化模块和动态网络构建模块。

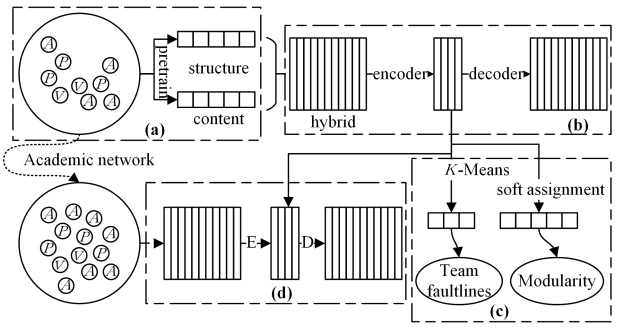

图 1　DANE-CD 算法框架

Fig. 1　Framework of DANE-CD

4.1　科研网络自编码器

科研网络自编码器将高维复杂的科研网络映射到低维的向量空间。自编码器的 encoder 部分将完成自高维网络到低维向量空间的映射任务，decoder 部分将根据低维向量重构整个网络。

由于动态科研网络$\mathbb{G}$是异构的属性网络，因此对于每一个网络快照 G^t，我们首先预训练其结构和内容特征。我们在 G^t 上采用 APA 与 APVPA 两种元路径方式进行随机游走，如图 2 所示，生成的路径将结合 metapath2vec[15] 以提取 G^t 的结构特征，记为 S^t；对于每个学者节点，我们提取其在时间 t 及之前发表的论文的标题和摘要，然后使用 Doc2Vec[16] 预训练出其内容向量，最后获得 G^t 的内容特征，记为 $\boldsymbol{C}^t$。我们将 $\boldsymbol{S}^t$ 和 $\boldsymbol{C}^t$ 融合得到网络的特征矩阵，记为 $\boldsymbol{F}^t=[\boldsymbol{S}^t,\boldsymbol{C}^t]\in\mathbb{R}^{N\times h_1}$。

自编码器的 encoder 定义为从网络融合特征矩阵 F^t 到网络表示矩阵 $\boldsymbol{H}^t\in\mathbb{R}^{N\times h_2}$，如式(1)所示：

$$\boldsymbol{H}^t=\phi(\boldsymbol{F}^t)=\sigma(\boldsymbol{W}^t\boldsymbol{F}^t+\boldsymbol{b}^t) \tag{1}$$

其中，$\boldsymbol{W}^t$ 为权重矩阵，$\boldsymbol{b}^t$ 为偏置，σ 为 sigmoid 函数。

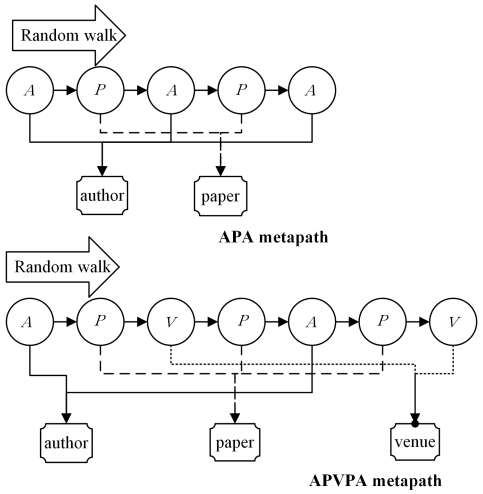

图 2　APA 与 APVPA 元路径

Fig. 2　APA and APVPA meta-paths

自编码器的 decoder 部分将 encoder 的输出重新构造为与 encoder 输入有相同特征维度的矩阵，如式(2)所示：

$$\widetilde{\boldsymbol{F}}^t=\phi(\boldsymbol{W}^t)=\sigma(\widetilde{\boldsymbol{W}}^t\boldsymbol{H}^t+\widetilde{\boldsymbol{b}}^t) \tag{2}$$

其中，$\widetilde{\boldsymbol{W}}^t$ 为权重矩阵，$\widetilde{\boldsymbol{b}}^t$ 为偏置。自编码器的损失定义为网络重构的损失，我们使用最小平方误差来计算该损失，如式(3)所示：

$$L_{ae}^t=\|\widetilde{\boldsymbol{F}}^t-\boldsymbol{F}^t\|_2^2 \tag{3}$$

4.2　聚类优化

科研网络中社区分配的本质是一个聚类的过程，但是在 4.1 节的表示学习中缺乏对聚类目标的优化。因此，本文定义了聚类优化模块从模块度和断裂带两个维度联合优化表示学习的结果。

模块度(modularity)由 Newman 等提出[4]，并拓展到了矩阵维度的表达[17]，是网络中被划分到给定社区的边的比例与对这些边进行随机分配时该比例的数学期望的差值。定义网络的邻接矩阵为 $\boldsymbol{J}$，网络中节点 i 的度数为 d_i，网络中边的数量为 $m=\sum_i d_i/2$，随机分配时节点 i 和 j 间期望的边数为 $d_id_j/2m$；定义矩阵 $\boldsymbol{H}\in\mathbb{R}^{N\times K}$ 表示节点的社区分配矩阵，其中 $\boldsymbol{H}_{ik}=1$ 表示节点 i 属于社区 k；定义矩阵 $\boldsymbol{B}\in\mathbb{R}^{N\times N}$ 为网络的模块度矩阵，其中 $\boldsymbol{B}_{ij}=\boldsymbol{J}_{ij}-d_id_j/2m$，则网络的模块度值的数学表示如式(4)所示：

$$Q=\frac{1}{4m}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{H} \tag{4}$$

最大化式(4)是 NP 难问题，因此拓展矩阵 $\boldsymbol{H}$ 为软分配矩阵[18]，即有 $\boldsymbol{H}_i\in\mathbb{R}$ 且 $\boldsymbol{H}_i^{\mathrm{T}}\boldsymbol{H}_i=N$。由于 4.1 节中得到的表示矩阵 $\boldsymbol{H}^t$ 的特征维度往往远大于社区的数量，因此不能直接使用 $\boldsymbol{H}^t$ 作为本节的社区分配矩阵 $\boldsymbol{H}$。本文定义了一个全连接层 $\widetilde{\boldsymbol{H}}^t$ 来将 $\boldsymbol{H}^t$ 映射到 $\boldsymbol{H}$，最终的模块度优化计算式如式(5)所示：

$$L_{\mathrm{mod}}^t=\frac{1}{4m}Tr(\widetilde{\boldsymbol{H}}^{t\mathrm{T}}\boldsymbol{B}\widetilde{\boldsymbol{H}}^t) \tag{5}$$

其中，Tr 表示矩阵的迹。模块化越高的网络在社区内部有更多的连接，在社区之间有更稀疏的连接，因为我们将最大化式(5)以获得更紧密的社区结构。

团队断裂带(team faultlines)指群体中多种特征因素影响导致群体分化为多个内部同质的子群间的分界线。团队断裂带越显著的群体，子群内部的互相认可就越强，有利于提高群体的自我效能；而子群间有更明显的差异，将有助于跳出“群体思维”，激发创造力[19]。本文创新性地将团队断裂带引入社区检测中，旨在提升社区发现的聚类质量。本文首先在 4.1 节中得到的表示矩阵 $\boldsymbol{H}^t$ 上使用 k-Means 聚类，以获得当前的社区分配结果 $\boldsymbol{\psi}^t$，然后沿用文献[20]中的研究计算当前网络的团队断裂带强度，如式(6)所示：

$$L_{tf}^t=\frac{\sum_{h=1}^{h_2}\sum_{k=1}^{K}|\boldsymbol{\psi}_k^t|(\overline{\boldsymbol{H}_{hk}^t}-\overline{\boldsymbol{H}_h^t})^2}{\sum_{h=1}^{h_2}\sum_{k=1}^{K}\sum_{n=1}^{N}(\boldsymbol{H}_{ihk}^t-\overline{\boldsymbol{H}_h^t})^2} \tag{6}$$

其中，h_2，K，N 分别为特征维度、社区数量和学者节点数量；$|\boldsymbol{\psi}_k^t|$ 表示社区 k 中学者节点的数量；$\boldsymbol{H}_{ihk}^t$ 为节点 i 第 h 维度的特征值；$\overline{\boldsymbol{H}_{hk}^t}$ 为社区 k 中所有节点的第 h 维度的特征均值；$\overline{\boldsymbol{H}_h^t}$ 为网络中所有节点第 h 维度的特征均值。本文将最大化式(6)以优化社区分配结果。

4.3　动态网络模型构造

本文在 4.1 节和 4.2 节中讨论了如何在动态科研网络$\mathbb{G}$的一个子网络 G^t 上对学者节点进行表示学习和聚类优化，但$\mathbb{G}$是动态变化的，本节将讨论如何构造动态网络模型。

本文基于堆栈自编码器(stacked autoencoder)来构建深度自编码器以学习动态网络的信息。具体而言，对于动态科

研网络$\mathbb{G}$的每一个网络快照G^t，首先按照4.1节中的方法预训练其结构和内容特征，并计算融合特征矩阵；对于第一个网络快照G^1，本文沿用4.1节中的encoder部分获得其表示矩阵；对于后续的每一个网络快照，将沿用上一个网络快照中训练完毕的自编码器的参数以获得当前堆栈自编码器前一层的表示矩阵$\boldsymbol{H}^{t-1}$，然后以式(7)获得当前网络快照对应子网络的表示矩阵$\boldsymbol{H}^t$。

$$\boldsymbol{H}^t=\phi(\boldsymbol{H}^{t-1})=\sigma(\boldsymbol{W}^t\boldsymbol{H}^{t-1}+b^t) \tag{7}$$

经过上一个快照的训练，H^{t-1}可视为对输入特征的一次压缩提取。通过持续复用之前训练完毕的网络快照的encoder网络参数，我们可在后续网络训练中复用特征提取过程，捕获网络在动态变化中的信息。

对于动态科研网络$\mathbb{G}$的连续网络快照$\{G^1,G^2,\cdots,G^t,\cdots,G^T\}$，我们将得到每一个快照下的表示矩阵$\{\boldsymbol{H}^1,\boldsymbol{H}^2,\cdots,\boldsymbol{H}^t,\cdots,\boldsymbol{H}^T\}$，最后时间快照下的表示矩阵$\boldsymbol{H}^T$将蕴含$\mathbb{G}$在$T$个时间快照下动态变化的信息。本文最后使用$k$-Means算法在$\boldsymbol{H}^T$上获得最终的社区分配结果$\boldsymbol{\psi}^T$。

4.4 模型的训练

对于动态科研网络$\mathbb{G}$的每一个时间快照G^t，本文设计了一个联合的目标函数来优化网络表示学习的过程，包括自编码器的重构损失、模块度最大化损失和团队断裂带强度最大化损失，如式(8)所示：

$$L^t_{\text{all}}=L^t_{ae}-\alpha\cdot L^t_{\text{mod}}-\beta\cdot L^t_{tf} \tag{8}$$

其中，α,β为损失的权重，各项损失的计算分别如式(3)、式(5)、式(6)所示。本文使用随机梯度下降(SGD)优化算法来训练网络，当训练损失收敛后进入下一个时间快照的训练，依次迭代完成整个动态模型的训练。

5 实验

本节将详细介绍面向第4节提出的DANE-CD算法的实验，包括实验的数据集、预处理方法、对比方法、评估指标、实验的结果及其讨论，以及超参数对实验指标的影响。

5.1 实验设定

本文使用AMiner提供的DBLP-Citation-network V11数据[21-22]和arXiv提供的高能物理理论数据集(HEP-TH)[23]进行实验。

DBLP V11数据集包括了计算机科学文献检索网站DBLP的数据，该版本赋予每一个学者唯一的ID，从根本上避免了其他数据集中存在的命名歧义问题。我们选择了数据挖掘与信息检索(DM-IR)、数据库(DB)、自然语言处理(NLP)和计算机视觉(CV)4个方向的11个会议，具体的会议如表1所列。

表1 DBLP数据集的方向与刊物名

Table 1 Directions and venues of DBLP dataset

Directions	Venues
Data mining andinformation retrieval (DM-IR)	KDD, SIGIR
Database (DB)	SIGMOD, VLDB, ICDE
Natural language processing (NLP)	ACL, EMNLP, NAACL
Computer vision (CV)	CVPR, ICCV, ECCV

由于原始数据存在较多的数据问题，因此本文首先进行数据预处理步骤。通过与DBLP的原始数据进行对比，我们修正了一些论文的发表时间和所属刊物；通过爬取论文原始文稿，本文补充了一部分论文缺失的摘要信息；还剔除了一部分被错误划分到选定刊物的论文。对于每个学者，本文统计其发表论文所属的方向，以包含论文数量最大的方向为其社区标签。本文将数据划分为DBLP-I(2009－2013)和DBLP-II(2014－2018)两个集合，每个集合包括5个时间片快照，在每个数据集合的最后一个时间片快照上进行算法评测。预处理后的DBLP数据集统计如表2所列。

表2 DBLP数据集统计

Table 2 Statistic of DBLP dataset

	DBLP-I	DBLP-II
Snapshots (number of years)	2009－2013(5)	2014－2018(5)
Number of author nodes	23657	35931
Number of paper nodes	15422	20155

HEP-TH数据集包括了arXiv中高能物理理论领域从1992年至2003年的论文，该数据集将学者划分为5个学科方向。首先剔除了HEP-TH数据集中没有作者和刊物的论文；然后通过命名实体消歧为所有学者赋予唯一的ID；类似对DBLP数据集的处理，将HEP-TH数据集划分为HEP-TH-I(1992－1997)和HEP-TH-II(1998－2003)两个集合，每个集合包括6个时间片快照，实验将在最后一个时间片快照上进行评测。预处理后的HEP-TH数据集描述如表3所列。

表3 HEP-TH数据集统计

Table 3 Statistic of HEP-TH dataset

	HEP-TH-I	HEP-TH-II
Snapshots (number of years)	1992－1997(6)	1998－2003(6)
Number of author nodes	6732	6632
Number of paper nodes	9948	8955

本文选择4种方法与本文方法进行对比。Louvain算法[5]是一种基于模块度最大化的快速社区检测算法，仅需网络的结构信息；由于Louvain算法仅能在静态网络上使用，因此本文将在动态科研网络的最后一个时间片快照上基于Louvain算法进行社区检测。metapath2vec[15]是一种基于元路径随机游走的网络表示学习方法，由于metapath2vec仅适用于静态网络，因此同样在动态科研网络的最后一个时间片快照上基于APA和APVPA两种元路径使用metapath2vec生成网络表示向量，最后使用k-Means获得社区分配的结果。本文还对比了动态网络表示学习方法dyngraph2vec-AERNN[14]，并使用k-Means在其动态表示结果上进行社区分配。

本文选择了准确率、归一化互信息和模块度3个指标来评价算法结果。聚类准确率度量了预测社区分配结果与实际结果的差距，计算方法如式(9)所示：

$$Accuracy(\widetilde{\boldsymbol{\psi}},\boldsymbol{\psi})=\frac{\sum_{i=1}^{N}\delta(\widetilde{\boldsymbol{\psi}}_i,map(\boldsymbol{\psi}_i))}{N} \tag{9}$$

其中，$\widetilde{\boldsymbol{\psi}},\boldsymbol{\psi}$分别代表实际社区分配结果和预测结果；$\delta(x,y)$为指示函数，当且仅当$x=y$时值为1，其余情况为0；$map$为保证统计准确的重现分配，一般使用匈牙利算法实现。归一化互信息从信息论的角度度量了社区分配的结果，如式(10)所示：

$$NMI(\widetilde{\boldsymbol{\psi}},\boldsymbol{\psi})=\frac{I(\widetilde{\boldsymbol{\psi}},\boldsymbol{\psi})}{(H(\widetilde{\boldsymbol{\psi}}),H(\boldsymbol{\psi}))/2} \tag{10}$$

其中，I 为互信息计算，H 为熵计算。模块度值 Q 的计算式如式(4)所示，不同于准确率和归一化互信息，模块度旨在通过无监督的方式评估社区分配的结果。

本文使用 PyTorch 实现模型，在多次调参后将预训练中随机游走的窗口大小设为 7，h_2 设为 128，α，β 分别为 0.5 和 0.1。实验机器为 Intel Xeon E5-2620 v4，128 GB 内存；NVIDIA GeForce RTX 2080，8GB 显存。

5.2　实验结果与分析

表 4 与表 5 分别列出了在 DBLP-I 和 DBLP-II 两个数据集上本文提出的 DANE-CD 算法与其他对比方法的实验结果，由于 Louvain 算法的检测结果与目标社区数量不匹配，因此未计算其准确率指标。由实验结果可知，DANE-CD 算法在准确率和归一化互信息两个指标上都达到了最佳，反映出了 DANE-CD 算法在划分社区时的准确性。在模块度指标上，DANE-CD 算法略逊于 Louvain 算法，这是因为 Louvain 算法是以模块度为其优化目标，但仍可见 DANE-CD 在提升社区划分准确性的同时也获得了更紧密的社区结构。

表 4　DBLP-I 数据集上的实验结果

Table 4　Experiment results on DBLP-I dataset

	Accuracy	*NMI*	*Q*
Louvain	—	0.469	**0.545**
metapath2vec	0.898	0.530	0.213
dyngraph2vecAERNN	0.933	0.589	0.248
DANE-CD	**0.942**	**0.719**	0.464

表 5　DBLP-II 数据集上的实验结果

Table 5　Experiment results on DBLP-II dataset

	Accuracy	NMI	*Q*
Louvain	—	0.566	**0.622**
metapath2vec	0.928	0.653	0.318
dyngraph2vecAERNN	0.939	0.687	0.341
DANE-CD	**0.964**	**0.791**	0.554

表 6 与表 7 列出了在 HEP-TH-I 和 HEP-TH-II 两个数据集上的实验结果。实验结果显示，相比 DBLP 数据集，算法的各项指标的表现略有下降，本文将此归因于 HEP-TH 数据集中学者分布较为分散。但 DANE-CD 算法相比其他算法仍有更佳的表现，进一步证实了本文算法的有效性。

表 6　HEP-TH-I 数据集上的实验结果

Table 6　Experiment results on HEP-TH-I dataset

	Accuracy	*NMI*	*Q*
Louvain	—	0.313	**0.498**
metapath2vec	0.853	0.469	0.173
dyngraph2vecAERNN	0.898	0.541	0.264
DANE-CD	**0.934**	**0.662**	0.482

表 7　HEP-TH-II 数据集上的实验结果

Table 7　Experiment results on HEP-TH-II dataset

	Accuracy	*NMI*	*Q*
Louvain	—	0.262	**0.516**
metapath2vec	0.824	0.436	0.152
dyngraph2vecAERNN	0.867	0.478	0.234
DANE-CD	**0.923**	**0.626**	0.473

本文还对 4.4 节中联合的目标函数中的超参数 α，β 进行调整，并在数据规模较大的 DBLP-II 和 HEP-TH-I 数据集上进行实验，结果如图 3 和图 4 所示。

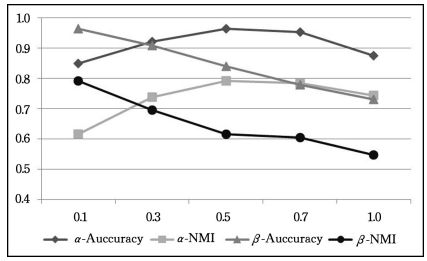

图 3　DBLP-II 数据集上的调参结果

Fig. 3　Parameter adjustments on DBLP-II

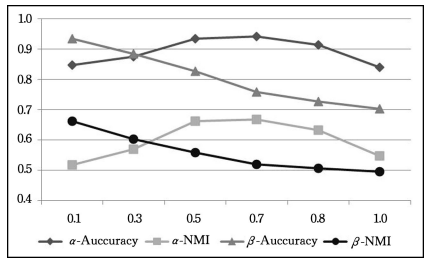

图 4　HEP-TH-I 数据集上的调参结果

Fig. 4　Parameter adjustments on HEP-TH-I

由此可见 α 偏大或偏小时模型的结果都有所下降，β 偏大时模型结果较差。当 α 取 0.5、β 取 0.1 时，算法有较佳的结果。

本文测试了在 DBLP-I 与 HEP-TH-I 这两个数据集上的算法运行时间，结果(10 次运行的均值，单位为 s)如表 8 所列。相比基于网络表示学习的方法，Louvain 算法的运行速度非常快，因其没有特征计算过程。本文方法 DANE-CD 由于捕获了网络动态特征及内容特征，因此运行时间更长，但也取得了最好的检测结果。

表 8　算法运行时间结果

Table 8　Results of running time

	DBLP-I	HEP-TH-I
Louvain	0.503	0.091
metapath2vec	263.78	51.43
dyngraph2vecAERNN	482.17	97.68
DANE-CD	619.62	133.24

结束语　本文提出了一种面向动态科研网络的新颖的社区检测算法 DANE-CD。该算法在提取科研网络独特的结构和内容特征的基础上，通过自编码器获得了科研网络表示学习的结果，并从模块度和团队断裂带两个维度对科研网络表示学习的过程予以优化，最后利用堆栈自编码器完成了对网络动态信息的学习。本文在 DBLP 和 HEP-TH 这两个科研数据集上的实验结果显示了所提算法的有效性。通过分析 DANE-CD 算法在不同时间段下的检测结果发现，科研网络的社区结构及其动态演化情况有助于为学者推荐潜在的合作者，为预测未来研究趋势和学者影响力变化提供了社区维度的信息，具有较好的应用前景。

本文在高能物理理论和计算机科学的部分领域上进行了实验，未来可以在涵盖更多学科的大型网络上进行进一步的实验并优化算法效率。本文仅考虑了非重叠社区场景下的社区检测问题，但学者可能归属于多个科研社区，未来需要考虑重叠社区的场景。跨学科和交叉学科的研究为科研带来了新的方向，如何融合多学科特征将是下一步的研究重点。

参 考 文 献

[1] KONG X, SHI Y, YU S, et al. Academic social networks: Modeling, analysis, mining and applications[J]. Journal of Network and Computer Applications, 2019, 132: 86-103.

[2] CHEN P, REDNER S. Community structure of the physical review citation network[J]. Journal of Informetrics, 2010, 4(3): 278-290.

[3] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821-7826.

[4] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6): 066133.

[5] 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, 2008(10): P10008.

[6] COMBE D, LARGERON C, EGYED-ZSIGMOND E, et al. Combining Relations and Text in Scientific Network Clustering [C]// Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Istanbul, Turkey: IEEE, 2012: 1248-1253.

[7] BARONI A, CONTE A, PATRIGNANI M, et al. Efficiently Clustering Very Large Attributed Graphs [C]// Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Sydney, Australia: ACM, 2017: 369-376.

[8] BHATT S, PADHEE S, SHETH A, et al. Knowledge Graph Enhanced Community Detection and Characterization [C]// Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. Melbourne VIC, Australia: ACM, 2019: 51-59.

[9] CHUNAEV P. Community detection in node-attributed social networks: A survey[J]. Computer Science Review, 2020, 37: 100286.

[10] YU P S, LIU F, XUE S, et al. Deep Learning for Community Detection: Progress, Challenges and Opportunities [C]// Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Yokohama, Japan: IJCAI Organization, 2020, 5: 4981-4987.

[11] CAVALLARI S, ZHENG V W, CAI H, et al. Learning Community Embedding with Community Detection and Node Embedding on Graphs [C]// Proceedings of the 26th ACM on International Conference on Information and Knowledge Management. Singapore: ACM, 2017: 377-386.

[12] TU C, ZENG X, WANG H, et al. A Unified Framework for Community Detection and Network Representation Learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(6): 1051-1065.

[13] LIU F, WU J, XUE S, et al. Detecting the evolving community structure in dynamic social networks[J]. World Wide Web, 2020, 23(2): 715-733.

[14] GOYAL P, CHHETRI S R, CANEDO A. dyngraph2vec: Capturing network dynamics using dynamic graph representation learning[J]. Knowledge-Based Systems, 2020, 187: 104816.

[15] DONG Y, CHAWLA N V, SWAMI A. Metapath2Vec: Scalable Representation Learning for Heterogeneous Networks [C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, NS, Canada: ACM, 2017: 135-144.

[16] LE Q, MIKOLOV T. Distributed Representations of Sentences and Documents [C]// Proceedings of the 31st International Conference on Machine Learning. Beijing, China: PMLR, 2014, 32: 1188-1196.

[17] NEWMAN M E J. Modularity and community structure in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(23): 8577-8582.

[18] YANG L, CAO X, HE D, et al. Modularity Based Community Detection with Deep Learning [C]// Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York, USA: IJCAI/AAAI Press, 2016: 2252-2258.

[19] THATCHER S M B, PATEL P C. Group Faultlines: A Review, Integration, and Guide to Future Research[J]. Journal of Management, 2012, 38(4): 969-1009.

[20] ZHAO W, PU S, JIANG D. A human resource allocation method for business processes using team faultlines[J]. Applied Intelligence, 2020, 50(9): 2887-2900.

[21] TANG J, ZHANG J, YAO L, et al. ArnetMiner: Extraction and Mining of Academic Social Networks [C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, Nevada, USA: ACM, 2008: 990-998.

[22] SINHA A, SHEN Z, SONG Y, et al. An Overview of Microsoft Academic Service (MAS) and Applications [C]// Proceedings of the 24th International Conference on World Wide Web. Florence, Italy: ACM, 2015: 243-246.

[23] ROSSI R A, AHMED N K. The network data repository with interactive graph analytics and visualization [C]// Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Texas, USA: AAAI Press, 2015: 4292-4293.

PU Shi, born in 1997, postgraduate, is a member of China Computer Federation. His main research interests include data mining and recommendation systems.

ZHAO Wei-dong, born in 1971, Ph. D, associate professor. His main research interests include intelligent data analysis and decision support systems.

(责任编辑:李亚辉)