

基于改进图卷积网络的学术文献分类方法

袁立宁^{1,2}, 冯文刚¹, 刘钊³

(1. 中国人民公安大学国家安全学院, 北京 100038; 2. 广西警察学院信息技术学院, 广西 南宁 530028;

3. 中国人民公安大学研究生院, 北京 100038)

摘要: 针对现有学术文献分类方法忽略文献数据之间关联信息的问题, 提出一种融合图卷积网络(GCN)和对比学习的文献分类模型对比图卷积网络(CGCN)。首先, 基于文献内容和引用关系定义两类“同质-异质”关联信息, 并将其转换为构建对比损失的自监督信息; 然后, 利用对比损失优化 GCN 特征提取过程, 推动同质文献特征表示彼此接近、异质文献特征彼此远离; 最后, 利用交叉熵损失和 Softmax 函数实现“端到端”的学术文献分类。在 3 个基准学术文献数据集上, CGCN 的文献分类表现优于当前较为先进的基线模型, 特别是 Cora 数据集上 Micro-F1 和 Macro-F1 指标值相较原始的 GCN 模型分别提高 8.29 和 7.91 百分点。CGCN 通过基于“同质-异质”关系构建的对比损失, 增强了模型对文献数据潜在信息的表征能力, 提高了分类的准确性和泛化性, 为学术文献分类研究提供了新思路和新方法。

关键词: 文献分类; 图卷积网络; 对比学习; 同质关系; 异质关系

中图分类号: TP18

文献标志码: A

DOI: 10.19678/j.issn.1000-3428.0252119

Academic Paper Classification Method Based on Improved Graph Convolutional Network

YUAN Lining^{1,2}, FENG Wengang¹, LIU Zhao³

(1. School of National Security, People's Public Security University of China, Beijing 100038, China;

2. School of Information Technology, Guangxi Police College, Nanning 530028, Guangxi, China;

3. Graduate School, People's Public Security University of China, Beijing 100038, China)

【Abstract】 Existing methods used for academic paper classification neglect relational information. To address this issue, this paper proposes a novel classification model that integrates a Graph Convolutional Network (GCN) with contrastive learning, called a Contrastive Graph Convolutional Network (CGCN). First, two distinct types of homogeneous-heterogeneous relational information are defined based on the content and citations of papers, and they are transformed into self-supervised information to construct contrastive loss. Second, contrastive loss is employed to enhance the feature extraction process of the GCN, such that homogeneous papers are extracted to be close to one another, while heterogeneous papers remain distant. Third, cross-entropy loss and the Softmax function are utilized to complete end-to-end academic paper classification. When tested on three benchmark academic datasets, the CGCN outperforms advanced baselines in the classification task. Its Micro-F1 and Macro-F1 scores are higher than those of the original GCN by 8.29 and 7.91 percentage points, respectively, on the Cora dataset. The CGCN enhances the capacity to represent potential information in papers by employing a contrastive loss based on a homogeneous-heterogeneous relationship, thereby improving the prediction accuracy and generalization. This method provides innovative ideas and methods for academic paper classification.

【Key words】 paper classification; Graph Convolutional Network (GCN); contrastive learning; homogeneous relationship; heterogeneous relationship

0 引言

近年来,学术文献数量随着科学研究工作的持续深入快速增长^[1],研究内容日趋复杂,研究主题愈加细化,给学术文献分类任务带来巨大挑战。传统基于

人工的学术文献分类方法依赖专家知识,分类结果往往存在较大的主观认知差异,因此难以满足数量庞大、形式多样且内容繁杂的学术文献分类任务需求^[2]。因此,通过机器学习、深度学习技术实现快速、准确的学术文献自动分类^[3],已成为提升学术文献资

基金项目: 国家重点研发计划(2023YFC3321604);广西哲学社会科学课题(23FTQ005)。

作者简介: 袁立宁(CCF 会员),男,高级工程师、博士研究生,主研方向为机器学习、图神经网络;冯文刚(通信作者),教授、博士;刘钊,副研究员、博士。

收稿日期: 2025-02-11

修回日期: 2025-04-27

E-mail: fengwengang@ppsuc.edu.cn

源管理水平、保证科学研究检索效率的重要方法。

现阶段,利用人工智能技术的学术文献分类方法主要分为两类:一是基于文献内容的分类方法^[4];二是基于文献内容与引用关系的分类方法^[5]。基于文献内容的方法利用模型与算法将学术文献内容转化为包含文献语义信息的特征向量,再通过分类器对文献向量进行分类,例如词袋(Bag-of-Words)模型^[6]、Word2vec^[7]、Transformer^[8]等模型。由于学术文献内容通常存在一词多义、一义多词、主题边界模糊以及学科交叉等情况,仅使用文献内容作为输入的方法通常存在一定的局限性^[9]。基于文献内容与引用关系的方法利用文献之间的引用关系传播和聚合文献内容信息并生成特征向量,再通过分类器对文献向量进行分类,例如图卷积网络(GCN)^[10]、图注意力网络(GAT)^[11]、洛伦兹线性图卷积网络(LLGC)^[12]等模型。

相较于基于文献内容的方法,基于文献内容与引用关系的方法能够同时保留语义和引用信息,对学术文献数据的表征能力相对更强。分析现有研究^[4-5],基于文献内容的方法大多通过增强文献语义特征提取的方式进行改进,基于文献内容与引用关系的方法则通过设计文献特征传递聚合策略进行改进。高效的特征提取和传播方式提升了模型的性能,但在文献分类任务中,上述方法忽略了文献之间语义相似性和引用关系等固有属性,一定程度上限制了模型的表征能力。

为解决上述问题,本文针对学术文献数据中潜在关联信息的表征学习,提出了一种融合图卷积网络和对比较学习^[13]的文献分类方法对比图卷积网络(CGCN),通过对比学习优化图卷积模型的特征提取过程。本文主要贡献如下:1)将学术文献数据中隐式的关联信息定义为两类显式的“同质-异质”关系,并利用显式关系和对比较学习构建对比损失;2)在CGCN的训练过程中引入对比损失,推动同质文献特征表示彼此接近、异质文献特征表示彼此远离;3)在3个开源学术文献数据集上,CGCN的分类表现均优于当前较为先进的基线模型;4)对比损失增强了模型对不同类型学术文献的表征能力,使生成的文献特征向量可区分性更高,不仅提升了模型在学术文献分类中的预测精度,也增强了模型在不同类型文献数据上的泛化能力。

1 相关工作

现阶段,基于深度学习的学术文献分类方法在图书情报学领域得到广泛应用。沈立力等^[14]使用

Transformer的变体BERT^[15]提取文献特征并输出每篇文献对应学科类别的概率,然后根据概率分布设定阈值范围,以减少学术文献分类过程中的人工干预。刘江峰等^[16]对比了不同深度学习模型的性能,验证了Transformer、BERT等深度预训练语言模型在文献分类中的有效性。丁恒等^[9]利用图表示学习方法生成包含文献语义特征和引用关系信息的文献特征向量,并将特征向量应用于文献分类和论文推荐任务。陈帅朴等^[17]将BERT与GAT模型进行融合,利用BERT增强基于文献内容的特征表示,利用GAT提取文献类目对标签的依赖关系,再通过层次多标签分类前馈网络(HMCN-F)模型聚合信息,实现学术文献多层次分类。王静静等^[18]总结了大语言模型^[19]在文献信息处理中的优势,探讨了大语言模型在文献分类任务中的应用前景。罗鹏程等^[20]使用学术文献数据和提示工程对大语言模型进行微调,然后使用微调后的大语言模型预测学术文献一级类目和完整分类号。TANG等^[21]构建了面向引文网络数据的GraphGPT,通过提示工程将引文网络的语义特征、引用关系集成到大语言模型中,从而实现文献主题预测。从实验结果^[20-21]看,大语言模型虽然能够通过提示工程提高学术文献分类结果的准确性,但预测结果相较Transformer、GCN等针对文献分类任务设计的模型仍存在一定差距。

除大语言模型外,现有学术文献分类实践大多通过增强文献语义特征提取、设计高效特征传递聚合策略等方式进行改进,忽略了文献数据本身蕴含的潜在关联信息。对此,本文构建了两类不同的“同质-异质”关系,并利用基于“同质-异质”关系引导的对比损失,优化模型的特征提取过程,从而增强模型对不同类型文献数据的表征能力。

2 模型与算法

学术文献包含文献内容和参考文献两部分,其中,文献内容可以转换为描述语义信息的文本特征,参考文献可以转化为与其他文献之间的引用关系。为了同时保留语义特征和引用关系信息,使用图数据 $G=(X,A)$ 表示学术文献数据。其中,属性矩阵 X 表示所有文献的语义特征, X 中的某一行 x_i 表示文献 i 的文本特征;邻接矩阵 A 表示所有文献之间的引用关系, A 中的某一元素 a_{ij} 表示文献 i 与文献 j 的引用关系(文献 i 与文献 j 之间存在引用则 $a_{ij}=1$,反之 $a_{ij}=0$)。

CGCN以表示文本特征的属性矩阵 X 和表示引用关系的邻接矩阵 A 为输入,利用GCN在不同文献

的一阶邻居之间传递和聚合特征信息,生成保留文献特征和引用关系的低维特征表示。然后,利用对比损失优化文献特征向量表示,推动同质文献(语义特征相似度高或存在引用关系)的特征表示在向量空间中彼此接近,推动异质文献(语义特征相似度低或不存

在引用关系)的特征表示在向量空间中彼此远离。最后,利用 Softmax 函数将文献特征向量转换为不同主题的概率分布,并通过交叉熵损失计算预测文献主题与真实文献主题之间的差异,实现学术文献分类任务。综上,CGCN 的总体框架如图 1 所示。

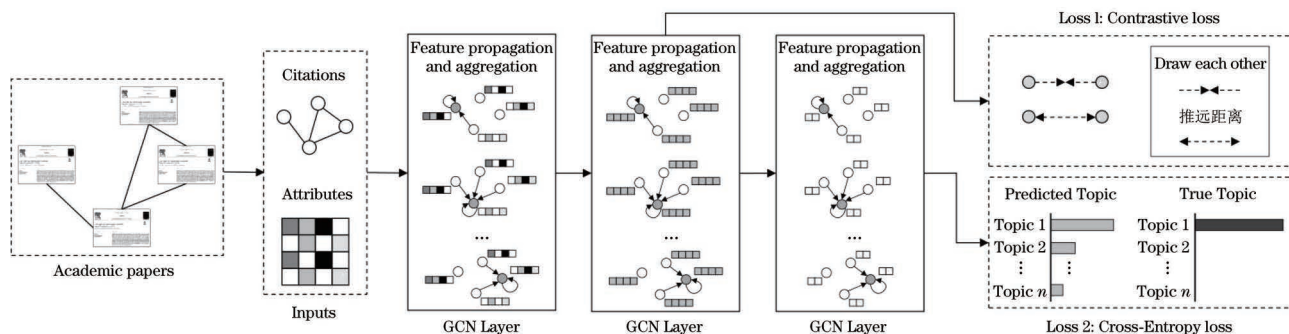


图 1 CGCN 模型整体结构

Fig.1 The overall framework of the CGCN model

2.1 图卷积网络

GCN 通过提取文本特征矩阵 X 和引用关系矩阵 A 中蕴含的潜在信息,生成用于预测文献主题类别的概率分布 $\hat{Y}$ 。对于多层 GCN 模型,文献特征的层间传播公式为:

$$H^{(l+1)} = \sigma(\tilde{A}H^{(l)}W^{(l)}) \quad (1)$$

式中: $H^{(l)}$ 表示第 l 层的文献特征表示($l=0$ 时表示文本特征矩阵 X); $W^{(l)}$ 表示第 l 层的参数矩阵; $\sigma(\cdot)$ 表示激活函数; $\tilde{A}$ 表示对称归一化的引用关系矩阵,用于传递和聚合文献一阶邻居的特征信息。

由于过多的 GCN 层会导致模型出现过平滑^[22],降低图卷积模型的性能,因此 CGCN 使用 3 层 GCN 提取文献数据特征,其前向传播公式为:

$$H^{(1)} = \sigma(\tilde{A}XW^{(0)})$$

$$H^{(2)} = \sigma(\tilde{A}H^{(1)}W^{(1)})$$

$$\hat{Y} = \text{Softmax}[\sigma(\tilde{A}H^{(2)}W^{(2)})] \quad (2)$$

式中: $\sigma(\cdot)$ 表示 ReLU 激活函数; $\hat{Y}$ 表示文献主题类别的概率分布,用于计算学术文献分类任务的交叉熵损失 l_{CE} 。 l_{CE} 计算公式为:

$$l_{CE} = - \sum_{i \in Y} \sum_{m=1}^M (Y_{im} \log_a(\hat{Y}_{im})) \quad (3)$$

式中: Y 表示文献的真实主题类别; M 表示主题类别数; Y_{im} 表示文献 i 真实类别是否为 m (真实类别为 m , $Y_{im}=1$, 否则 $Y_{im}=0$); $\hat{Y}_{im}$ 表示文献 i 属于类别 m 的预测概率。

原始的 GCN 及其变体模型大多通过设计更高效的特征传递和聚合方式,提升文献主题分类等复杂网络分析任务的实验表现,忽略了学术文献数据中潜在的“同质-异质”信息。因此,CGCN 在 GCN

模型的基础上,引入基于文本特征相似性和文献引用关系引导的对比损失函数,推动同质文献特征表示彼此接近、异质文献特征表示彼此远离,增强模型对文献特征信息的表征能力,提升模型的性能。

2.2 文献数据的“同质-异质”关系

在现代数据科学与信息技术领域,数据的“同质性”和“异质性”是两个常见概念^[23],也是影响数据分析和建模过程的重要因素。数据的同质性是指同一数据集中数据样本在特征、类型等方面表现出的一致性 or 相似性,数据的异质性是指同一数据集中数据样本在特征、类型等方面表现出的多样性或差异性。在处理大规模复杂数据时,充分利用数据中蕴含的同质性和异质性信息,能够有效提升模型的预测性能和泛化能力。

在学术文献数据集中,数据样本的同质性和异质性可以通过文献内容和引用关系直接表示。例如,主题相同的学术文献通常使用相似的关键词和专业术语,使得文献内容特征相似度高,反之,主题不同的文献内容特征相似度较低。此外,学术文献通常会引用与研究内容相关的先前研究,使得存在引用关系的文献主题相同的可能性较大,反之,不存在引用关系的文献主题相同的可能性较小。综上,本文将学术文献数据的“同质-异质”关系归纳为两类:一是由文献语义特征相似度确定的数据关系(相似度高为同质文献,相似度低为异质文献);二是由文献引用关系确定的数据关系(存在引用关系为同质文献,无引用关系为异质文献)。

2.3 对比损失

对比损失能够将原始数据中存在的可转移知识

转换为自监督信息,减弱模型训练过程对标签信息的依赖,增强模型对数据中潜在信息的提取和表征能力。其中,最常用的对比损失构建方式是增大同质数据特征表示的相似性和异质数据特征表示的差异性。因此,对比损失的目标函数^[24]可表示为:

$$l_{i,j} = -\log_a \left(\frac{\exp[\text{sim}(\mathbf{z}_i, \mathbf{z}_j)]}{\sum_{k=1}^N \mathbf{1}_{[k \neq i]} \exp[\text{sim}(\mathbf{z}_i, \mathbf{z}_k)]} \right) \quad (4)$$

式中: sim 表示相似度函数; $\mathbf{z}$ 表示数据样本的特征向量; $\mathbf{1}_{[k \neq i]}$ 表示符号函数, 仅当 $k \neq i$ 时取值为 1。本文通过对比损失 $l_{i,j}$ 增大了数据 i 与同质样本 j 特征向量的相似性, 增大了数据 i 与异质样本 k 特征向量的差异性。本文将学术文献数据的全质性和全质性关系转换为构建对比损失的自监督信息, 推动语义特征相似度高或存在引用关系的全质文献特征表示彼此接近、语义特征相似度低或不存在引用关系的异质文献特征表示彼此远离, 从而提升 GCN 模型在学术文献分类任务中的实验性能。

对于学术文献的语义特征相似度, 使用余弦相似度函数和文献特征向量, 计算文献 i 与文献 j 的特征相似度值 S_{ij} :

$$S_{ij} = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|} \quad (5)$$

式中: $\mathbf{x}_i$ 表示文献 i 文本特征向量; $\mathbf{x}_j$ 表示文献 j 文本特征向量。为了有效保留文献的语义特征同质性, 避免低相似度文献的干扰, 每篇文献仅选择相似度值排名前 k 的文献建立同质性关系矩阵 $\mathbf{A}'$, 同时为避免模型训练过程中出现梯度爆炸, $\mathbf{A}'$ 已经过归一化处理。此时, 基于文本特征相似性的文献“同质-异质”关系可定义为:

$$\begin{cases} a'_{ij} \neq 0, \text{文献 } j \text{ 与文献 } i \text{ 为同质关系} \\ a'_{ij} = 0, \text{文献 } j \text{ 与文献 } i \text{ 为异质关系} \end{cases} \quad (6)$$

式中: $a'_{ij} \in \mathbf{A}'$ 。结合式(4), 基于文本特征相似性引导的对比损失表达式为:

$$l_{\text{Feature}}^i = -\log_a \left(\frac{\sum_{j=1}^N \mathbf{1}_{[j \neq i]} a'_{ij} \exp[\text{sim}(\mathbf{h}_i^{(2)}, \mathbf{h}_j^{(2)})]}{\sum_{k=1}^N \mathbf{1}_{[k \neq i]} \exp[\text{sim}(\mathbf{h}_i^{(2)}, \mathbf{h}_k^{(2)})]} \right) \quad (7)$$

通过 l_{Feature}^i 能够拉近向量空间中文献 i 与语义特征相似度高的同质文献的距离, 推远文献 i 与语义特征相似度低的异质文献的距离。

对于学术文献的引用关系, 引用关系矩阵 $\mathbf{A}$ 中元素 a_{ij} 表示文献 i 与文献 j 的引用关系。因此, 基于文献引用关系的文献“同质-异质”关系可定义为:

$$\begin{cases} a_{ij} \neq 0, \text{文献 } j \text{ 与文献 } i \text{ 为同质关系} \\ a_{ij} = 0, \text{文献 } j \text{ 与文献 } i \text{ 为异质关系} \end{cases} \quad (8)$$

此时, 基于文献引用关系引导的对比损失可定义为:

$$l_{\text{Citation}}^i = -\log_a \left(\frac{\sum_{j=1}^N \mathbf{1}_{[j \neq i]} a_{ij} \exp[\text{sim}(\mathbf{h}_i^{(2)}, \mathbf{h}_j^{(2)})]}{\sum_{k=1}^N \mathbf{1}_{[k \neq i]} \exp[\text{sim}(\mathbf{h}_i^{(2)}, \mathbf{h}_k^{(2)})]} \right) \quad (9)$$

最后, 将交叉熵损失 l_{CE} 和对比损失 l_{CL} 进行组合, 构建 CGCN 完整的损失函数:

$$l = l_{\text{CE}} + \alpha l_{\text{CL}} \quad (10)$$

式中: α 表示平衡 l_{CE} 和 l_{CL} 的权重系数。CGCN 通过基于文本特征相似性和文献引用关系引导的对比损失, 在训练过程中保留关于特征和引用的“同质-异质”关系信息, 进而增强模型对学术文献数据的表征能力。

2.4 算法与过程

学术文献分类的根本目标是输出每篇文献对应不同主题的概率分布 $\hat{\mathbf{Y}}$, 并将概率值最大的主题类别作为文献分类的最终结果。因此, 学术文献分类的关键在于寻找合适的映射算法提升文献主题预测结果的准确性。在 CGCN 中, 通过多层神经网络编码学术文献的语义特征矩阵 $\mathbf{X}$ 和引用关系矩阵 $\mathbf{A}$ 中蕴含的特征信息, 生成所有文献的预测主题分布 $\hat{\mathbf{Y}}$ 。在参数优化过程中, 除了使用常见的分类损失函数——交叉熵损失, 还引入了基于语义相似度和引用关系引导的对比损失, 以增强映射模型的表征能力, 提高预测结果的准确性。综上, 使用 CGCN 模型进行学术文献分类的算法过程如算法 1 所示。

算法 1 CGCN

输入 文本特征矩阵 $\mathbf{X}$, 引用关系矩阵 $\mathbf{A}$, 部分文献主题标签 $\mathbf{Y}$, 隐藏层维度 d , 特征相似文献数 k , 权重系数 α

输出 文献主题概率分布 $\hat{\mathbf{Y}}$

1. 计算文本特征矩阵 $\mathbf{X}$ 的余弦相似度值, 选择前 k 个文献数据建立特征同质关系矩阵 $\mathbf{A}'$;
2. 初始化参数矩阵 $\mathbf{W}^{(l)}$, $l=1, 2, 3$;
3. Repeat
4. 根据式(2)计算文献特征表示 $\mathbf{H}^{(2)}$ 和文献主题概率分布 $\hat{\mathbf{Y}}$;
5. 根据式(10)计算损失函数 l ;
6. 计算损失函数的梯度 ∇l , 利用反向传播算法优化参数矩阵 $\mathbf{W}^{(l)}$;
7. Until Convergence

8. 输出文献主题预测结果 $\hat{Y}$ 。

3 实验与分析

3.1 数据集

本文使用 Cora、Citeseer、PubMed 和 ACM 4 个

不同规模和不同主题类型的学术文献数据集^[25], 评估不同方法在学术文献分类任务中的实验表现。在数据集中, 每篇文献由表示文献语义内容的词向量表示, 并保留与其他文献的引用关系。各数据集统计信息如表 1 所示。

表 1 学术文献数据集统计信息

Table 1 Statistical information of academic paper datasets

Datasets	Subject	Word Embeddings	Number of Papers	Number of Citations	Number of Attributes	Number of Topics
Cora	Machine Learning	Bag-of-Words	2 708	5 429	1 433	7
Citeseer	Information Science	Bag-of-Words	3 327	4 732	3 703	6
PubMed	Medical Science	TF-IDF	19 717	44 338	500	3
ACM	Computer Science	Bag-of-Words	3 025	13 128	1 870	3

具体而言, Cora 数据集由机器学习论文组成, 共包含 2 708 篇学术文献以及 5 429 条引用关系, 每篇文献的内容由 1 433 维的词向量表示, 根据内容将文献主题划分为 7 类; Citeseer 数据集由信息科学论文组成, 共包含 3 327 篇学术文献以及 4 732 条引用关系, 每篇文献的内容由 3 703 维的词向量表示, 根据内容将文献主题划分为 6 类; PubMed 数据集由医学论文组成, 共包含 19 717 篇学术文献以及 44 338 条引用关系, 每篇文献的内容由 500 维的词向量表示, 根据内容将文献主题划分为 3 类; ACM 数据集由计算机技术论文组成, 共包含 3 025 篇学术论文以及 13 128 条引用关系, 每篇文献的内容由 1 870 维的词向量表示, 根据内容将文献主题划分为 3 类。

此外, 数据集除规模和研究主题存在差异外, 描述文献内容的词向量构建方式也有所不同。其中, Cora、Citeseer 和 ACM 数据集使用词袋模型, 通过统计文献中每个词出现的频率(去除停用词), 得到描述文献内容的词典, 并生成表示文本特征的 one-hot 词向量; PubMed 数据集使用词频-逆文档频率(TF-IDF)模型^[6], 通过计算文献中每个词汇的重要性, 减少区分性弱的词, 保留区分性强的词(某个词在部分文献中出现频率高, 在其他文献中出现频率低), 得到描述文献内容的词典, 并生成表示文本特征的 one-hot 词向量。通过上述不同规模和类型的文献数据集, 能够有效评估不同算法在文献分类任务中的准确性和泛化性。

3.2 实验设置

为有效评估 CGCN 在文献分类任务中的实验表现, 使用当前较为先进和常用的模型作为基线, 其中包括仅使用文本特征的多层感知机(MLP)^[26]和 Kolmogorov-Arnold 网络(KAN)^[27], 以及同时使用文本特征和引用关系的 Transformer^[28]、GCN、

LLGC、GAT、SuperGAT^[29]和 GKAN (Graph Kolmogorov-Arnold Networks)^[30]。

所有基线按照原始论文中建议的参数进行初始化, 并根据不同数据集上的实验表现进行微调, 以获得最佳性能。同时, 为保证实验的公平性, GCN 及其变体模型均采用与 CGCN 相同的 3 层网络结构, 学习率设置为 0.001, 隐藏层维度设置为 512。CGCN 特有参数文本特征相似文献数 k 和权重系数 α 通过参数搜索的方式确定, k 在 {1, 3, 5, 7, 9, 10, 25, 50, 75, 100, 150} 中搜索, α 在 {1, 5, 10, 15, 20} 中搜索。在 Cora 数据集上, 设置 $k=7$ 、 $\alpha=10$; 在 Citeseer 数据集上, 设置 $k=5$ 、 $\alpha=10$; 在 PubMed 数据集上, 设置 $k=50$ 、 $\alpha=10$; 在 ACM 数据集上, 设置 $k=75$ 、 $\alpha=5$ 。所有模型均使用 Adam 优化器更新模型参数, 同时在训练过程中引入 Dropout, 防止参数过拟合。此外, 各模型采用相同的数据集划分, 每个文献数据集随机抽取 10% 作为训练集, 剩余 90% 作为测试集, 并使用 Micro-F1 和 Macro-F1 指标评估模型实验性能。

3.3 文献分类实验

CGCN 模型与基线模型在 Cora、Citeseer、PubMed 和 ACM 数据集上学术文献分类的实验结果如表 2 所示, 其中粗体表示最优分类结果, 下划线表示次优分类结果, 下同。

从表 2 的实验结果看, 有以下分析:

1) 在 4 个数据集上, CGCN 的学术文献分类表现均优于当前较为先进的基线模型, 特别是 Cora 数据集上 CGCN 的 Micro-F1 和 Macro-F1 指标值相比较表现最优的基线 SuperGAT 分别提升了 3.62 和 3.82 百分点。上述结果表明, 基于文本特征相似性和引用关系引导的对比损失能够增强模型对学术文献数据特征的提取和保留, 提升图卷积模型在学术文献分类任务中的实验性能。

表 2 学术文献分类实验结果

Table 2 Experimental results of academic paper classification

Methods	Cora		Citeseer		PubMed		ACM	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
MLP	62.92	57.15	62.58	52.20	82.05	82.35	85.42	85.32
KAN	55.41	50.23	62.11	53.36	81.70	81.99	86.39	86.38
Transformer	72.23	68.82	72.22	<u>64.96</u>	<u>82.46</u>	<u>82.71</u>	90.43	90.41
GCN	76.25	75.16	71.23	64.69	75.97	75.54	<u>90.97</u>	<u>90.88</u>
LLGC	55.25	40.19	73.22	61.28	61.60	50.49	81.37	81.20
GAT	74.90	73.51	70.49	63.80	77.03	76.52	90.42	90.35
SuperGAT	<u>80.92</u>	<u>79.25</u>	<u>74.74</u>	63.08	<u>75.79</u>	75.58	90.39	90.28
GKAN	79.16	77.99	72.12	60.60	78.08	77.59	90.77	90.67
CGCN	84.54	83.07	74.79	67.97	83.99	83.63	91.28	91.22

2)在 Cora、Citeseer 和 ACM 数据集上,同时使用文献语义特征和引用关系信息的基线模型预测效果较好,但是在 PubMed 数据集上仅使用文献语义特征信息的基线模型预测效果较好。上述结果表明,GCN、GAT 等图深度学习模型在某些数据集上提取文献特征以及融合引用关系信息的能力有限,导致其实验表现弱于 MLP、KAN 等仅使用文献语义特征进行训练的分类器。

3)由于信息传递和聚合方式的不同,GCN 及其变体在不同数据集上的分类表现差异较大,例如 GCN 在 Cora、Citeseer 和 ACM 数据集上表现较

好,但在 PubMed 数据集上表现较差。与之相对,引入对比损失的 CGCN 在 4 个数据集上均展现出良好的分类性能。

上述结果表明,引入基于文本特征相似性和文献引用关系引导的对比损失,不仅提升了模型在文献分类任务中的预测性能,同时增强了模型对不同种类文献数据的泛化能力。

为了进一步验证不同图神经网络模型的泛化能力,分别选取不同训练比率进行文献分类实验,并记录 Micro-F1 值,实验结果如图 2 所示(彩色效果见《计算机工程》官网 HTML 版,下同)。

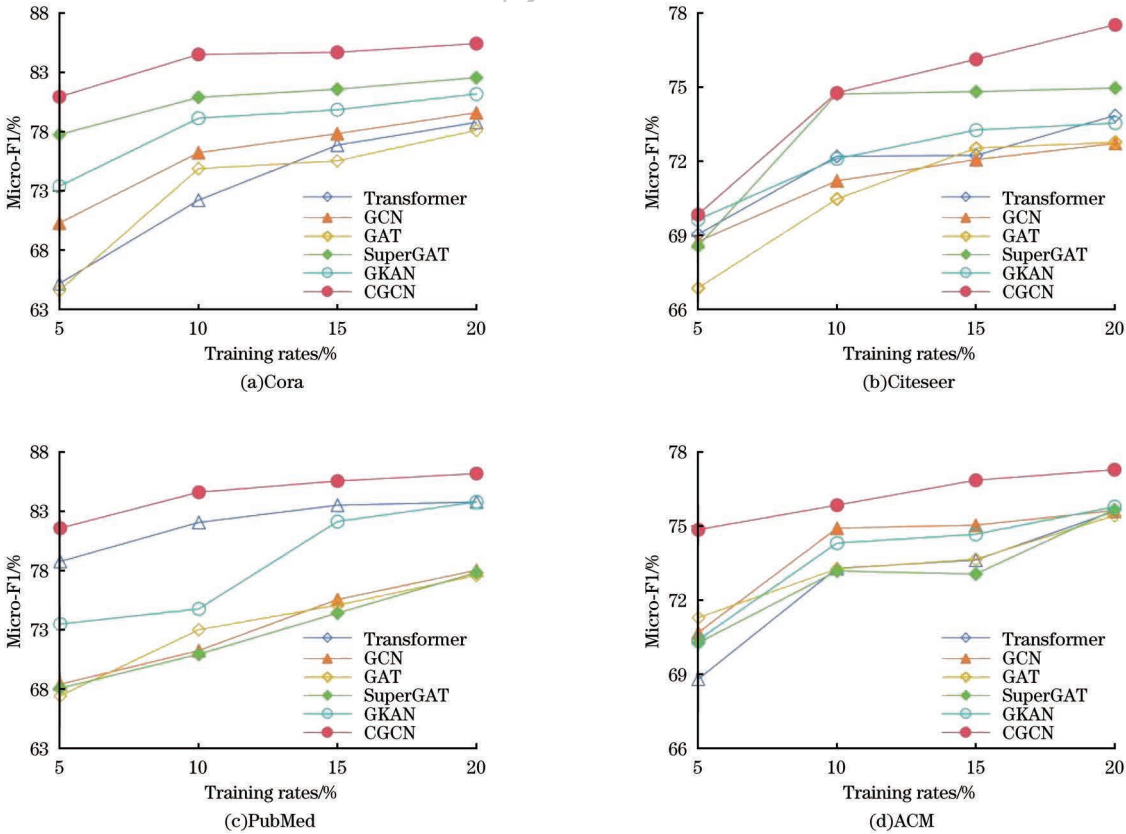


图 2 不同训练比率文献分类实验结果

Fig.2 Experimental results of paper classification with different training rates

从图 2 可以看出,各模型性能均随着训练比率的增加有所提升,但 CGCN 的预测性能始终优于基线模型。在多个数据集上,CGCN 在训练数据仅有 5% 时,相较基线模型提升显著。上述结果表明,引入基于“同质-异质”关系引导的对比损失,增强了 CGCN 从数据中提取和表示文献特征信息的能力,从而提高了模型在文献分类任务中的实验性能。

3.4 消融实验

为了验证基于“同质-异质”关系引导的对比损

失对 CGCN 分类性能的影响,在 4 个学术文献数据集上进行消融实验。其中包括 3 类变体:1)不使用“同质-异质”对比损失的变体 GCN;2)仅使用文本特征(即基于文本特征相似性的文献“同质-异质”关系)引导对比损失的变体 GCN-feature;3)仅使用文献引用关系(即基于文献引用关系的文献“同质-异质”关系)引导对比损失的变体 GCN-citation。为保证实验的公平性,各变体参数均参照 CGCN 进行设置,消融实验结果如图 3 所示。

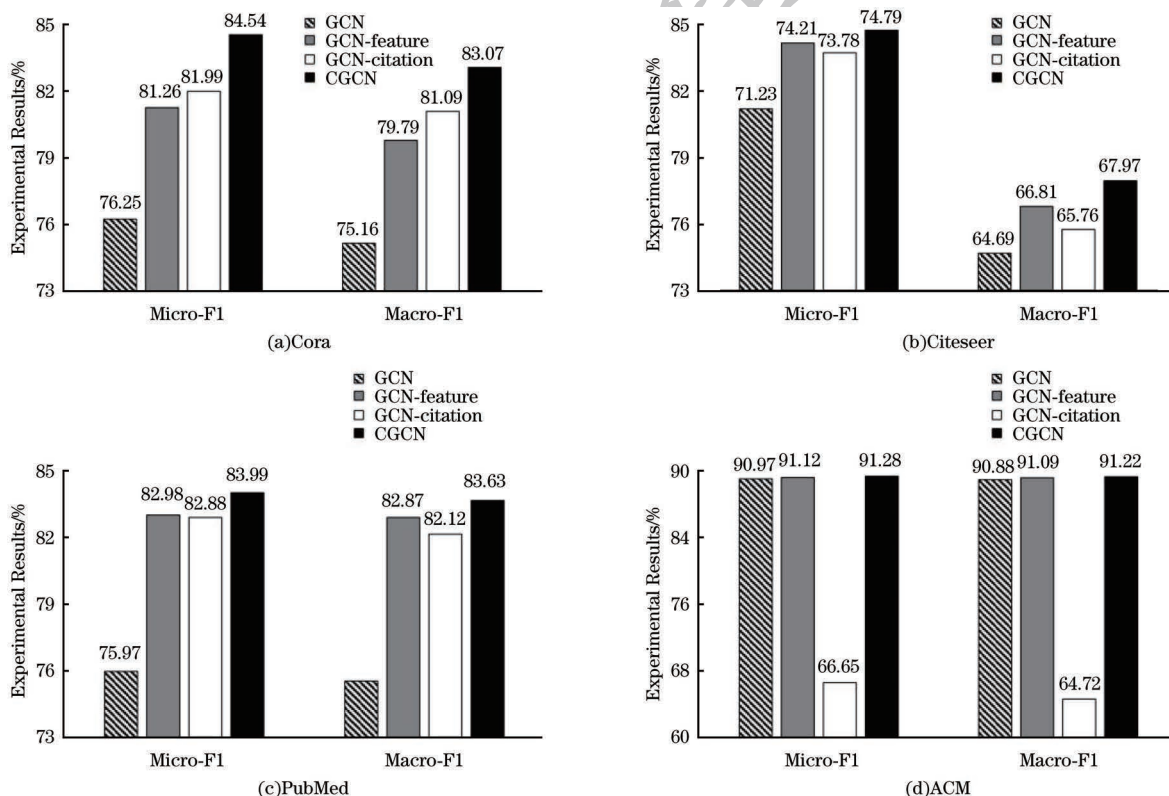


图 3 消融实验结果

Fig.3 Ablation experiment results

在多个数据集上,引入“同质-异质”关系引导对比损失的变体 GCN-feature、GCN-citation 以及 CGCN 的分类表现优于未引入对比损失的方法,表明“同质-异质”关系的引入增强了模型的表征能力和泛化能力,使模型生成的文献特征向量可区分性更高,分类性能更强;在 Cora 数据集上,GCN-feature 的分类表现弱于 GCN-citation,在 Citeseer、PubMed 和 ACM 数据集上,GCN-feature 的分类表现优于 GCN-citation。可见在关注文本语义信息的学术文献数据上,使用文本特征对比损失的 GCN-feature 表现更好;在关注引用关系的学术文献数据上,使用引用关系对比损失的 GCN-citation 表现更好。但在 ACM 数据集上,使用引入“同质-异质”关系引导对比损失的变体 GCN-citation 实验表现显著弱于其他对比模型,甚至低于不使用对比损失的

原始 GCN 方法,表明仅使用基于引用的“同质-异质”关系在特定数据集上泛化能力有限,甚至可能损害模型性能。综合对比 GCN、GCN-feature、GCN-citation 和 CGCN,CGCN 在 4 个数据集上均获得了最佳的分类表现,证明了基于语义特征和引用关系的对比损失能够使模型提取和保留更丰富的原始数据信息,进一步提升图卷积模型在学术文献分类任务中的准确性和泛化性。

3.5 参数实验

为了分析模型参数对 CGCN 实验性能的影响,在 4 个学术文献数据集上进行参数实验。相较学习率、迭代次数和隐藏层维度等常见参数,本节仅对 CGCN 特有参数——文本特征相似文献数 k 和权重系数 α 进行验证,参数实验结果如图 4 和图 5 所示。

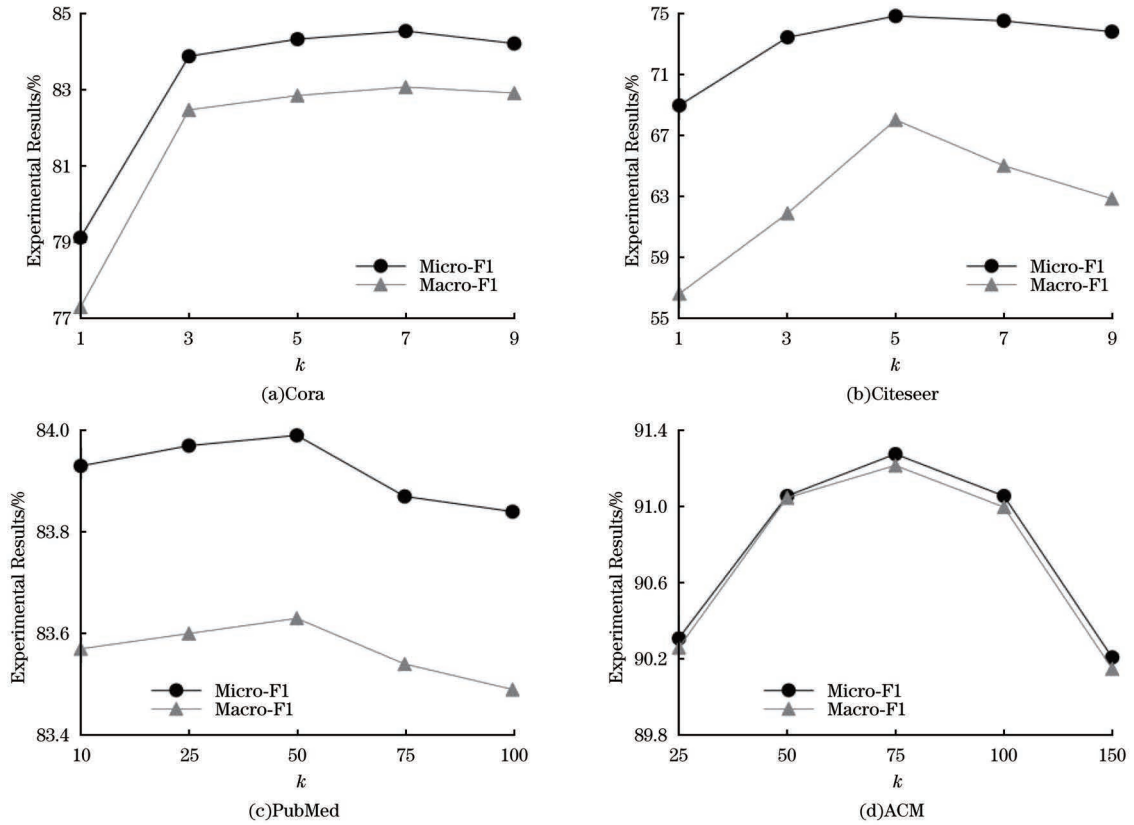


图 4 特征相似文献数对学术文献分类的影响

Fig.4 The impact of number of similar papers on academic paper classification

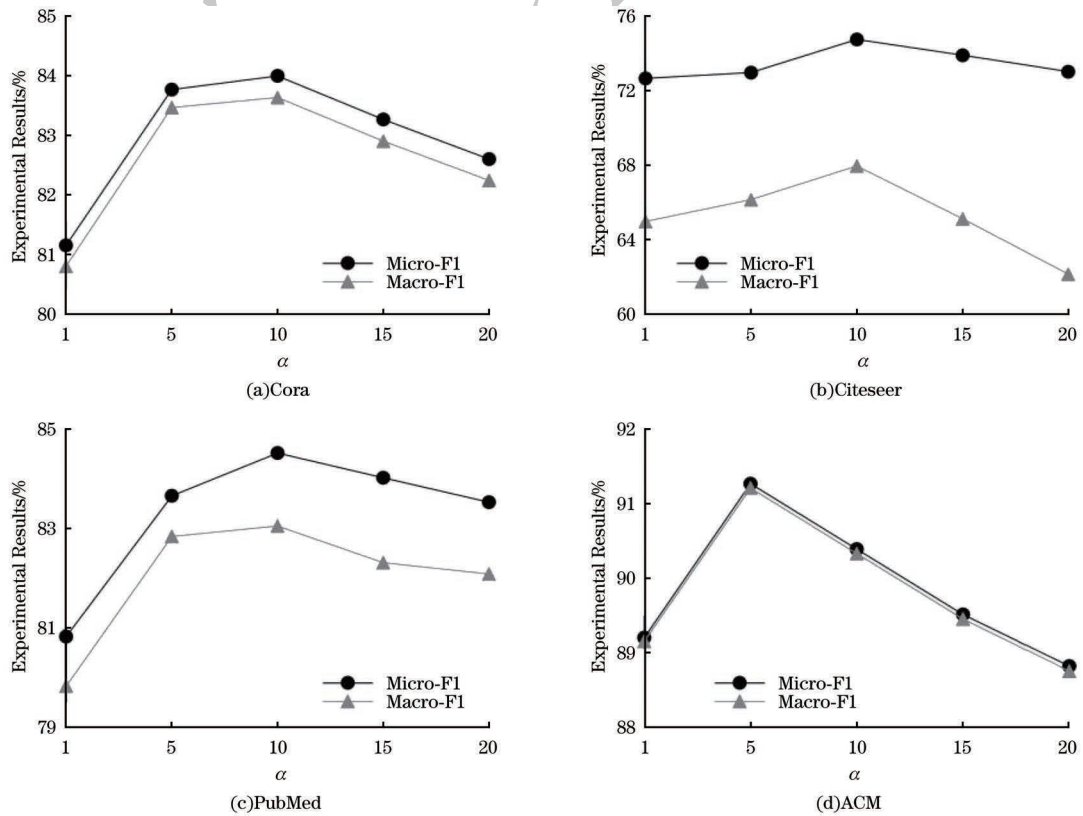


图 5 对比损失权重对学术文献分类的影响

Fig.5 The impact of contrastive loss weight on academic paper classification

为了验证文本特征相似文献数 k 对 CGCN 分类性能的影响,使用不同的 k 值进行参数实验。在

CGCN 中, k 用于调节由文献语义特征相似度确定的同质和异质文献关系。观察图 4 可知,4 个学术

文献数据集上的 Micro-F1 和 Macro-F1 曲线随 k 值的增加呈现先上升后下降的趋势,过小和过大的 k 值均未取得最佳的实验结果。当文本特征相似文献数 k 值较小时,文献之间形成的同质关系较少,无法充分保留文本特征相似度信息;当文本特征相似文献数 k 值较大时,文献之间形成的同质关系较多,甚至与一些低相似度文献形成同质关系,降低了模型对高相似度文献的关注度,同时低相似度同质关系会产生一定的噪声,进而导致模型的分类性能下降。

为了验证权重系数 α 对 CGCN 分类性能的影响,使用不同的 α 值进行参数实验。在 CGCN 中, α 用于调节基于文本特征相似性和文献引用关系的对比损失在模型优化过程中的比重。观察图 5 可知,4 个学术文献数据集上的 Micro-F1 和 Macro-F1 曲线随 α 值的增加呈现先上升后下降的趋势,过小和过大的 α 值均未取得最佳的实验结果。当权重系数

α 值较小时,模型更关注分类损失的优化,即模型在训练过程中对训练集上的分类表现关注度高,对表征文献语义和引用信息关注度低;当权重系数 α 值较大时,模型更关注对比损失的优化,即模型在训练过程中对表征文献语义和引用信息关注度高,对训练集上的分类表现关注度低。从实验结果看,在模型优化过程中应设置合适的权重系数 α 以平衡分类损失和对比损失,获得最佳分类性能。

3.6 对比损失实验

为验证 CGCN 中语义相似性和引用关系两类“同质-异质”关系引导的对比损失的有效性,在 GCN 优化过程中分别引入当前常用的对比损失:一是“局部”拓扑对比损失^[31],变体表示为 GCN-local;二是“局部+全局”拓扑对比损失^[32],变体表示为 GCN-global。为保证实验公平性,保持各模型参数一致,实验结果如表 3 所示。

表 3 对比损失比较实验结果
Table 3 Experimental results of contrastive loss comparison

%

Methods	Cora		Citeseer		PubMed	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
GCN	76.25	75.16	71.23	64.69	75.97	75.54
GCN-local	81.99	81.09	73.78	65.76	82.88	82.12
GCN-global	82.36	81.76	73.92	67.13	83.04	82.79
CGCN	84.54	83.07	74.79	67.97	83.99	83.63

从表 3 的实验结果看,有以下分析:

1)不同对比损失均提升了 GCN 的实验表现,证明了对比学习在文献分类任务中的有效性。

2)相较仅使用局部拓扑的 GCN-local,保留更丰富拓扑信息的 GCN-global 表现更好,证明了高阶结构信息的有效性。

3)对比 GCN-local、GCN-global 和 CGCN,同时使用语义和引用信息的 CGCN 分类表现最好,证明“同质-异质”关系引导的对比损失有效性的同时,表明在使用引用关系的基础上引入语义相似性,能够进一步增强模型对文献数据的表征能力,提升预测性能。

3.7 可视化

为了直观分析 CGCN 模型对学术文献数据的表征能力,使用 t -SNE 算法对 Cora 数据集上基线模型和 CGCN 生成的文献特征向量进行可视化。由于学术文献的特征向量中蕴含了文献语义特征、引用关系等信息,可视化后能够直观反映学术文献数据集的潜在特征以及不同模型的表征能力。学术文献数据的可视化结果如图 6 所示,不同主题的文献被标记为不同颜色。

好的可视化结果通常相同主题的文献彼此接近,不同主题的文献彼此分离。观察图 6 可知,所有模型均能从任意分布的文献数据中提取特征信息,同一主题的文献能够形成一定规模的社区结构,但不同模型的可视化结果差异较大,社区的类内相似性和类间界限区别明显。其中,MLP、Transformer、GCN 和 GAT 的文献分布较为稀疏,不同主题的类间界限不够明显,相反 CGCN 的文献分布更加紧密,不同主题的类间界限更加清晰。可视化实验直观反映了模型提取和保留同一主题文献语义特征和引用关系信息的能力,证明了 CGCN 中引入的对比损失增强了模型对学术文献数据的表征能力,使同一主题的文献特征表示在向量空间中彼此接近、不同主题的文献特征表示在向量空间中彼此远离,生成的文献特征向量可区分性更高,相较现有方法能够更为准确地判别文献所属主题。

3.8 实验结果分析

CGCN 通过引入“同质-异质”关系引导的对比损失,增强模型对文献数据特征提取和信息保留的能力。相较基线模型、使用单一“同质-异质”关系的变体,CGCN 生成的特征向量可区分性更高,对文

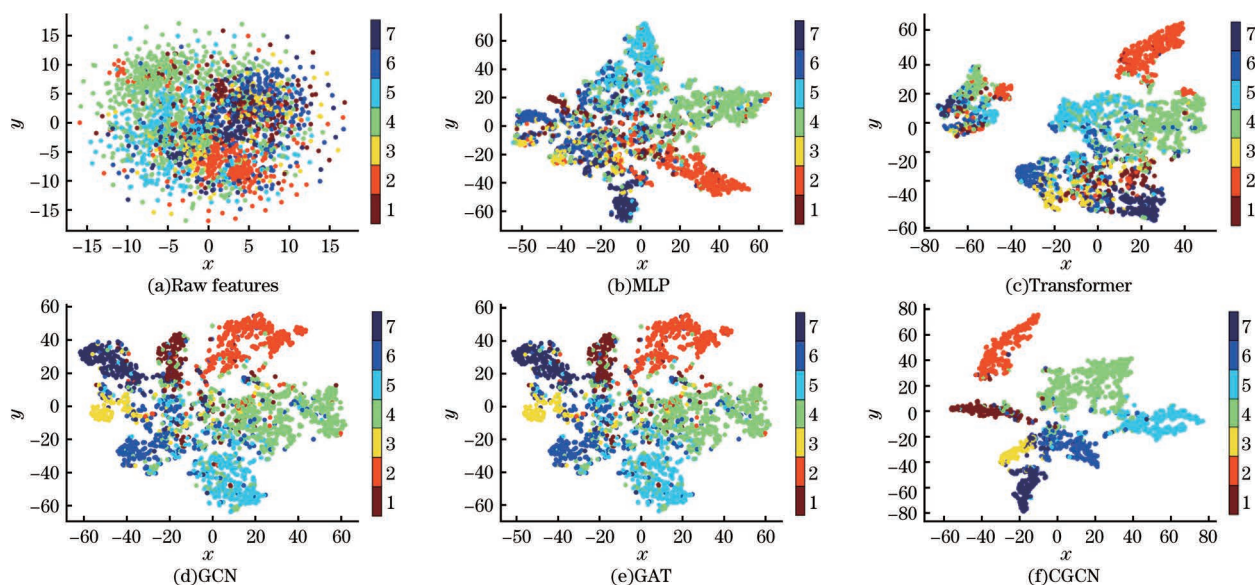


图 6 学术文献的特征向量可视化

Fig. 6 Visualization of academic paper feature vectors

献不同主题类型的判断能力更强,同时 CGCN 能够利用语义特征和引用关系中蕴含的自监督信息优化特征提取过程,增强模型对不同类型学术文献数据的泛化能力。此外,可视化实验直观反映了文献特征向量的可区分性,相较基线模型,两类“同质-异质”关系的引入使 CGCN 形成的社区类内相似性和类间界限更为明显。

4 结束语

学术文献分类是实现文献资源数字化管理、保障学术文献情报库智能化建设的重要基础性工作。针对现有方法忽略文献数据潜在关联信息的问题,本文提出了一种新的文献分类模型 CGCN,通过基于文献“同质-异质”关系构建的对比损失增强模型对文献数据特征信息的保留能力。在 3 个开源学术文献数据集上,CGCN 的分类表现均优于当前较为先进的基线模型,证明对比损失的引入增强了模型对文献数据的表征能力,提高了模型在分类任务中的准确性和泛化性。此外,通过对比损失分析实验可知,同时使用语义特征与引用关系两类“同质-异质”关系的对比损失相较仅使用语义特征或仅使用引用关系的对比损失对文献数据的表征能力更强,在文献分类任务中的预测效果更好。在未来工作中,考虑通过引入高阶特征信息、注意力机制等方式,进一步增强模型的特征提取能力,以实现更高效、更准确的学术文献自动分类。

参考文献

[1] 中国科技论文统计与分析课题组. 2022 年中国科技论文统

计与分析简报[J]. 中国科技期刊研究, 2024, 35(1): 95-103.

Research Group of Statistics and Analysis on Chinese Scientific Papers. A brief report of statistics and analysis on Chinese scientific papers in 2022 [J]. Chinese Journal of Scientific and Technical Periodicals, 2024, 35(1): 95-103. (in Chinese)

[2] 张雨卉. 基于《中国图书馆分类法》的文献自动化深层分类的研究和实现[J]. 图书馆杂志, 2024, 43(3): 61-74.

ZHANG Y H. A study of automated deep classification of literature based on *Chinese library classification* [J]. Library Journal, 2024, 43(3): 61-74. (in Chinese)

[3] 徐江玲, 陈兴荣. 基于层级图标签表示网络的多标签文本分类[J]. 计算机应用研究, 2024, 41(2): 388-392, 407.

XU J L, CHEN X R. Multi-label text classification based on hierarchical graph label representation network [J]. Application Research of Computers, 2024, 41(2): 388-392, 407. (in Chinese)

[4] KHURANA D, KOLI A, KHATTER K, et al. Natural language processing: state of the art, current trends and challenges[J]. Multimedia Tools and Applications, 2023, 82(3): 3713-3744.

[5] WANG K Z, DING Y H, HAN S C. Graph neural networks for text classification: a survey[J]. Artificial Intelligence Review, 2024, 57(8): 190.

[6] AKUMA S, LUBEM T, ADOM I T. Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets [J]. International Journal of Information Technology, 2022, 14(7): 3629-3635.

[7] JOHNSON S J, MURTY M R, NAVAKANTH I. A detailed review on word embedding techniques with emphasis on word2vec[J]. Multimedia Tools and Applications, 2024, 83(13): 37979-38007.

[8] ISLAM S, ELMEKKI H, ELSEBAI A, et al. A comprehensive survey on applications of transformers for deep learning tasks[J]. Expert Systems with Applications, 2024, 241: 122666.

[9] 丁恒, 任卫强, 曹高辉. 基于无监督图神经网络的学术文献表示学习研究[J]. 情报学报, 2022, 41(1): 62-72.

DING H, REN W Q, CAO G H. Using unsupervised graphs of neural networks for constructing learning representations of academic papers [J]. Journal of the China Society for

- Scientific and Technical Information, 2022, 41(1): 62-72. (in Chinese)
- [10] REN Z. VGCN: an enhanced graph convolutional network model for text classification [J]. Journal of Industrial Engineering and Applied Science, 2024, 2(4): 110-115.
- [11] HE L C, BAI L, YANG X, et al. High-order graph attention network[J]. Information Sciences, 2023, 630: 222-234.
- [12] LIANG Q Y, WANG W H, BAO F L, et al. L²GC: lorentzian linear graph convolutional networks for node classification[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italy: ELRA, 2024: 9988-9998.
- [13] 岑科廷, 沈华伟, 曹婧, 等. 图对比学习综述[J]. 中文信息学报, 2023, 37(5): 1-21.
- CEN K T, SHEN H W, CAO Q, et al. A survey on graph contrastive learning [J]. Journal of Chinese Information Processing, 2023, 37(5): 1-21. (in Chinese)
- [14] 沈立力, 姜鹏, 王静. 基于 BERT 模型的中文期刊文献自动分类实践研究[J]. 图书馆杂志, 2022, 41(5): 109-118, 135.
- SHEN L L, JIANG P, WANG J. A study on the automatic classification of Chinese literature in periodicals based on BERT model[J]. Library Journal, 2022, 41(5): 109-118, 135. (in Chinese)
- [15] LIAO W X, LIU Z L, DAI H X, et al. Mask-guided BERT for few-shot text classification[J]. Neurocomputing, 2024, 610: 128576.
- [16] 刘江峰, 林立涛, 刘畅, 等. 深度学习驱动的海量人文社会科学学术文献学科分类研究[J]. 情报理论与实践, 2023, 46(2): 71-81.
- LIU J F, LIN L T, LIU C, et al. Study on the discipline classification of massive humanities and social science academic literature driven by deep learning[J]. Information studies: Theory & Application. 2023, 46(2): 71-81. (in Chinese)
- [17] 陈帅朴, 钱宇星, 钱志强, 等. 多重特征关联和图注意力网络融合的文献分类方法研究——以中文医学文献为例[J]. 情报学报, 2024, 43(4): 470-490.
- CHEN S P, QIAN Y X, QIAN Z Q, et al. Research on literature classification methods based on multiple feature correlation and a graph attention network model: a case study of Chinese medical literature[J]. Journal of the China Society for Scientific and Technical Information, 2024, 43(4): 470-490. (in Chinese)
- [18] 王静静, 叶鹰, 王婉茹. ChatGPT 类 AI-GPT 技术应用对图书馆信息处理的变革探析[J]. 图书馆理论与实践, 2024(1): 122-127, 136.
- WANG J J, YE Y, WANG W R. A prospective analysis on ChatGPT-type AI-GPT technical applications for changing library information processing [J]. Library Theory and Practice, 2024(1): 122-127, 136. (in Chinese)
- [19] 罗文, 王厚峰. 大语言模型评测综述[J]. 中文信息学报, 2024, 38(1): 1-23.
- LUO W, WANG H F. Evaluating large language models: a survey of research progress [J]. Journal of Chinese Information Processing, 2024, 38(1): 1-23. (in Chinese)
- [20] 罗鹏程, 王继民, 聂磊. 基于生成式大语言模型的文献资源自动分类研究[J]. 情报理论与实践, 2024, 47(12): 174-182.
- LUO P C, WANG J M, NIE L. Research on automatic classification of literature resources based on generative large language model [J]. Information Studies (Theory & Application), 2024, 47(12): 174-182. (in Chinese)
- [21] TANG J B, YANG Y H, WEI W, et al. GraphGPT: graph instruction tuning for large language models[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2024: 491-500.
- [22] WU G C, LIN S K, SHAO X X, et al. QPGCN: graph convolutional network with a quadratic polynomial filter for overcoming over-smoothing[J]. Applied Intelligence, 2023, 53(6): 7216-7231.
- [23] WU L R, LIN H T, HU B Z, et al. Beyond homophily and homogeneity assumption: relation-based frequency adaptive graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(6): 8497-8509.
- [24] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//Proceedings of the 37th International Conference on Machine Learning. New York, USA: PMLR, 2020: 1597-1607.
- [25] YANG Z, COHEN W, SALAKHUTDINOV R. Revisiting semi-supervised learning with graph embeddings [C] // Proceedings of the 33rd International Conference on Machine Learning. New York, USA: JMLR, 2016, 48: 40-48.
- [26] YUAN L, JIANG P, HOU W, et al. G-MLP: graph multi-layer perceptron for node classification using contrastive learning[J]. IEEE Access, 2024, 12: 104909-104919.
- [27] LIU Z, WANG Y, VAIDYA S, et al. KAN: Kolmogorov-Arnold Networks[EB/OL]. [2025-01-20]. <https://arxiv.org/abs/2404.19756>.
- [28] LIU C, ZHAN Y B, MA X Q, et al. Gapformer: graph Transformer with graph pooling for node classification[C]//Proceedings of the 32nd International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan Kaufmann, 2023: 2196-2205.
- [29] KIM D, OH A. How to find your friendly neighborhood: graph attention design with self-supervision[EB/OL]. [2025-01-20]. <https://arxiv.org/abs/2204.04879>.
- [30] KIAMARI M, KIAMARI M, KRISHNAMACHARI B. GKAN: Graph Kolmogorov-Arnold Networks [EB/OL]. [2025-01-20]. <https://arxiv.org/abs/2406.06470>.
- [31] ZHANG H R, WU Q T, WANG Y, et al. Localized contrastive learning on graphs [EB/OL]. [2025-03-30]. <https://arxiv.org/abs/2212.04604>.
- [32] TANG Q, ZHAO Y J, WU H, et al. Adversarial cluster-level and global-level graph contrastive learning for node representation[J]. Knowledge-Based Systems, 2023, 279: 110935.

文字编辑 金胡考
栏目编辑 赖玉玲