doi:10.3969/j.issn.1001-4616.2024.01.010

基于泛化图卷积神经网络的深度文档聚类模型

柴变芳[1],李　政[1],赵晓鹏[2],王荣娟[3]

(1.河北地质大学信息工程学院,河北 石家庄 050031)
(2.河北省财政厅一体化系统运维中心,河北 石家庄 050091)
(3.河北地质职工大学,河北 石家庄 050086)

[**摘要**] 文本分类是自然语言处理中一项重要任务,基于图神经网络的文本分类因其可建模文本间的多种交互成为一种主流方法. 但现有方法大都依赖标签,而真实标签难以获取. 提出一个基于图泛化卷积神经网络的深度文档聚类模型(generalization graph convolutional neural network-deep document clustering,GGCN-DDC),同时实现文本表示学习和无监督文档分类. 该模型首先将每个文档建模为文本图;然后采用泛化卷积层学习更有区分力的文档词特征表示和文档表示;最后通过文档聚类损失和文档图重建损失约束参数学习算法. 在 3 个基准数据集上的实验表明,GGCN-DDC 在多个指标上均优于其他基准算法.

[**关键词**] 图神经网络,深度图聚类,文本分类,文本表示
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2024)01-0082-09

Deep Document Clustering Model Based on Generalization Graph Convolutional Neural Network

Chai Bianfang[1],Li Zheng[1],Zhao Xiaopeng[2],Wang Rongjuan[3]

(1.College of Information Engineering,Hebei GEO University,Shijiazhuang 050031,China)
(2.Integrated system operation and maintenance center,Hebei Provincial Department of Finance,Shijiazhuang 050091,China)
(3.Hebei Vocational College of Geology,Shijiazhuang 050086,China)

Abstract:Text classification is an important task in natural language processing. The method of text classification on graph neural network has become a mainstream method since it can model the interactions among texts. However,most of the existing graph-based classification methods rely on real labels,which are difficult to captain. A deep document clustering model based on graph generalization convolutional neural network(GGCN-DDC)is proposed,which can realize unsupervised text classification while learning text representation. Firstly,the documents are modeled as a text graph. Then generalized convolution layer is used to learn the more distinguishable feature representations of words and the document representations. Finally,The learning algorithm of parameters is constrained by document clustering and reconstructing document graph. Experiments on three benchmark datasets show that GGCN-DDC outperforms other benchmark algorithms on several measures.

Key words:graph neural network,deep graph clustering,text classification,text representation

随着计算机技术、互联网技术的普及,各行各业产生的数据量激增. 文本是这些海量数据中一种重要信息载体,如商品评论、新闻消息、邮件内容、论文摘要、图片文字描述、文学作品等. 对这些文本按照语义进行分类,能够帮助决策者快速了解这些海量信息蕴含的关键知识,进而应用到更精确的后续任务,如客户喜好分析、新闻分类、邮件过滤、情感分类、智能信息检索等. 真实世界的文本数据标签注释困难,而且数据迭代日新月异,对监督分类来说工作量更加巨大[1]. 因此,无监督文本分类在大数据时代有着重要的研究意义.

收稿日期:2023-05-29.
基金项目:河北省高等学校科学技术研究项目(ZD2020175)、河北地质大学 2023 国家预研项目(KY202310).
通讯作者:王荣娟,副教授,主要研究方向:自然语言处理、图挖掘、机器学习等. E-mail:30673253@ qq.om

文本分类的核心主要分为文本表示和分类器训练两部分. 按照文本表示的方法不同可以简单地将文本分类划分为浅层模型和深层模型. 早期的浅层模型主要通过手工特征工程再采用机器学习模型分类的方法[2]. 这一类方法基于先验提取文本特征,可能丢失文本中丰富的语义信息. 随着深度学习的发展,深层模型利用深度学习强大的表示学习能力学习文本的丰富特征,如循环神经网络 RNN、卷积神经网络 CNN. Kim 将 CNN 应用到了文本分类中设计了 TextCNN 模型[3],Liu 等[4]将 RNN 与多任务文本分类相结合,实现了更适合不同长度语句的分类模型 TextRNN. 这些文本分类方法都是针对局部连续的单词序列,忽视了文本全局词共现信息和整体结构信息[5].

将词看作图节点,词间关系作为图连边,可将文本建模为图结构,对文本全局范围内不连续的词语义关系建模. 进而可用图神经网络(graph neural network,GNN)[6]对文本图进行表示学习和分类. Yao 等[7]基于 GCN 提出了 TextGCN,把文档和词建模为图节点,文档和词间关系以及上下文共现词间关系建模为图链接,得到全局图后基于 GCN 进行文本分类. 这种模型可以利用文档间全局信息,但是对内存消耗很大. Dai 等[8]指出这些方法大多忽略了对新文本的可扩展性,而且忽视了文本图质量,提出了一种 GFN 模型,通过动态融合词嵌入来生成文本表示,再利用外部知识(词频共现统计和预训练嵌入)来构建不同级的语料库文本图,通过图卷积后融合这些图表示以相互弥补,从而提升性能. Zhang 等[9]提出了 TextING 模型,将每个文档建模为一个文档图,实现细粒度的文本表示,同时也能实现新文档可扩展学习. 但这些方法都是监督模型,无法适应文档标签难以获取的场景.

Cui 等[10]提出了一种半监督模型 ST-Text-GCN,提取文本中关键词,使得模型能够在有少量标签数据集下,通过自监督学习到足够多的信息,对文本进行分类. Haj-Yahia[11]等利用文档中最相关单词之间的文本相似性和专家标注的反映标签语义的关键字字典,来丰富类别标签,从而在这些知识的基础上,通过文档相关单词和关键字的相似性来对文档进行无监督分类. Schopf[12]等提出的无监督模型 Lbl2Vec,通过计算文档嵌入和预定义标签关键字的余弦相似度来决定文档分类. 这些无监督文档分类方法,要么需要专家知识、要么需要人工定义的关键字或者主题词做指导,并没有做到完全无监督.

近年来有学者提出了一些深度图聚类算法实现无监督图分类,其通过无监督深度学习模型学习图嵌入特征,同时以聚类损失来指导嵌入学习和聚类. Tian[13]等通过堆叠式自动编码器学习图嵌入,在图嵌入上用 K-means 算法获得聚类. Zhang 等[14]提出 AGC 模型,用自适应图卷积根据簇内距离自动选择图卷积层数,来学习指定范围图结构信息并用谱聚类对节点进行划分. Zhu 等[15]使用 Davies-Bouldin 指数(DBI)代替簇内距离作为选择标准来更灵活地选择图信息获取范围. Wang 等[16]提出了 DAEGC 模型,使用图注意力网络学习节点表示,利用聚类损失和图邻接矩阵重建损失微调图嵌入和分类结果. 该方法构建了嵌入学习和聚类的统一框架,使得无监督分类效果更加准确. 深度图聚类方法利用“无监督特征提取+聚类”的思想很好地实现了无监督聚类,但是现有的深度图聚类方法不是直接面向文本,而是针对处理好的链接图数据(如引文网络、蛋白质网络等样本间有边相连的数据). 而真实世界文本语料库之间文档间关系通常并不会显式的给出,这使得它在面向真实文本数据时表现不理想;其次,这些模型特征提取部分借助的图卷积部分无法进行更深层的嵌入学习,如 AGC 采用的自适应方法也只是变相地避开了深层图卷积网络过平滑的问题.

本文提出了一种无监督文档图嵌入学习和分类模型 GGCN-DDC,直接面向文本实现无监督分类. 为了既能得到文档嵌入又能实现文本聚类,模型在继承了 TextING 面向文本表示的可扩展性同时,改进了卷积层无法学习更深层信息的不足,并将文本表示和文档聚类建模为统一框架,提升文本特征提取能力和文档聚类效果. 为了保留 TextING 细粒度文档信息学习能力,提出了适合 TextING 结构的新的重建矩阵损失函数,隐式的学习了无链接文档间的关系,提升了模型的分类性能.

1 深度文档聚类模型及算法

本部分首先介绍深度文档聚类模型 GGCN-DDC 框架,然后介绍模型的各部分如何实现、优化目标及学习算法.

1.1 问题阐述及相关定义

给定一个文档语料库,将每个文档建模为图结构,然后构建图神经网络模型学习文档图中的词、文档

嵌入,并基于文档嵌入实现无监督文档分类. 文档图根据文档所含词之间共现关系构建,每个文档图以其包含的词为图节点,以指定窗口内词间共现关系定义边. 每个文档图记为 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$,其中 $\boldsymbol{V}$ 代表文档的词节点集,$\boldsymbol{E}\subseteq\boldsymbol{V}\times\boldsymbol{V}$ 代表边集,文档图的邻接矩阵 $\boldsymbol{A}=\{a_{ij}\}$,其中 $a_{ij}=1$ 代表节点 i 和 j 之间有边,$a_{ij}=0$ 则表示没有边. $\boldsymbol{D}=\mathrm{diag}\{d_1,d_2,\cdots,d_l,\cdots,d_n\}$ 表示 $\boldsymbol{A}$ 的度矩阵(其中 n 表示文档图中词节点个数). d_l 表示文档中词节点 l 的度. 文档图拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$. 此处文档图间并无连边.

1.2　GGCN-DDC 模型

深度文档聚类模型 GGCN-DDC 框架如图 1 所示. 该模型包括文档图构建及初始化、文档图词节点和图嵌入学习、文档图聚类三部分. 文档图构建及初始化:根据文档语料库构建每个文档的文档图,并对图节点进行词向量初始化. 文档图词节点和图嵌入学习:基于所有文档图,利用图泛化卷积网络实现文档图词节点嵌入,经过注意力机制处理后再池化得文档图嵌入. 文档图聚类:以学到的文档嵌入为输入特征,以聚类损失和重建损失作为约束,联合学习图嵌入和文档聚类模型.

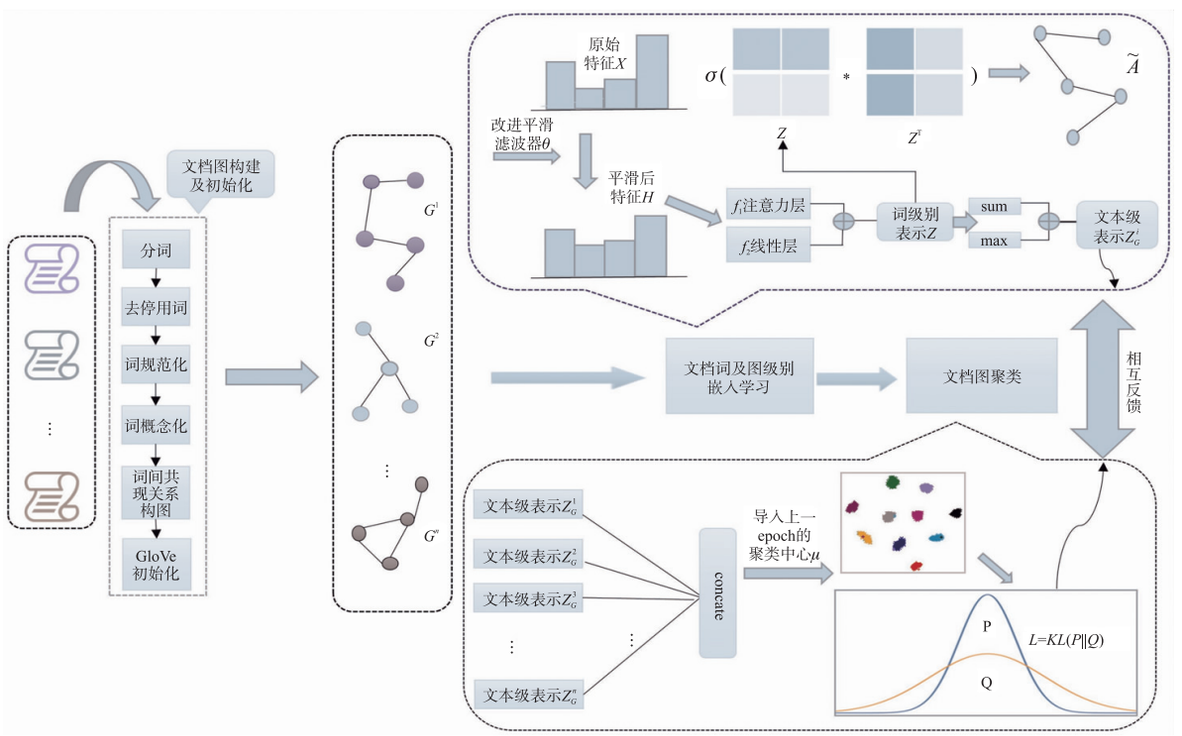

图 1　GGCN-DDC 模型结构

Fig. 1　Model structure of GGCN-DDC

1.3　文档图构建及初始化

文本预处理将语料库中的每个文档构建为文档图. 首先对每个文档进行预处理,包括分词、去停用词、词形规范化、词概念化等,从而获取词典并将文档表示为词序列. 然后根据词序列间共现关系构建文档图,词作为节点,相互之间有共现关系则视为有边,而词间共现关系出现的频率作为边的权重. 将文档图的词节点嵌入初始化,如使用 GloVe[17] 词预训练向量,记为 $\boldsymbol{X}\in\mathbf{R}^{|v|\times d}$,其中 d 是词嵌入的初始维度.

1.4　文档图词节点和图嵌入学习

TextING 模型也能实现文档图词和图级别嵌入学习,但其使用 GRU 提取特征,过于复杂、时间效率很差. 这里采用改进的图卷积实现文本图特征提取,下面介绍改进卷积.

GCN 中图节点嵌入学习部分定义为:

$$\boldsymbol{H}^{(t)}=\boldsymbol{\sigma}(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(t-1)}\boldsymbol{W}^{(t-1)}),\tag{1}$$

式中,$\tilde{\boldsymbol{D}}=\boldsymbol{D}+\boldsymbol{I}$,$\boldsymbol{W}^{(t-1)}$ 是权重矩阵,$\boldsymbol{H}^{(t-1)}$ 是第 $t-1$ 层学习到的特征,$\boldsymbol{H}^{(0)}=\boldsymbol{X}$,$\boldsymbol{\sigma}$ 是激活函数.

以 GCN 的一层图卷积为例:

$$\boldsymbol{Y}=\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X},\tag{2}$$

图特征的拉普拉斯平滑:

$$\tilde{\boldsymbol{Y}}=\boldsymbol{X}-\gamma\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{L}}\boldsymbol{X}=(\boldsymbol{I}-\gamma\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{L}})\boldsymbol{X}, \tag{3}$$

式中,$\tilde{\boldsymbol{L}}=\tilde{\boldsymbol{D}}-\tilde{\boldsymbol{A}}$.

当 $\gamma=1$,且采用对称归一化拉普拉斯算子时,(2)(3)等价,也就是说 GCN 本质是一种特殊形式的拉普拉斯平滑,对应一个低通滤波器. 但此时该滤波器并不是一个最佳的低通滤波,文献[18]研究表明当 γ 为 2/3 时,滤波效果最好,能提取最佳的图特征. 此外,用公式(1)的图节点嵌入学习过程中,基于权重 $\boldsymbol{W}$ 的线性变换直接与滤波器特征过滤联合计算,对特征平滑没有增益,甚至会造成损害. 将线性变换与滤波器滤波过程进行解耦,先用滤波器学习 t 层的特征,然后再进行线性变换. 滤波器特征提取计算如下:

$$\tilde{\boldsymbol{X}}=\boldsymbol{\Theta}^{t}\boldsymbol{X} \tag{4}$$

式中,$\boldsymbol{\Theta}=(\boldsymbol{I}-\gamma\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{L}})$.

经过多层滤波器特征提取后,在输出层之前进行线性变换,计算如下:

$$\boldsymbol{Y}'=\tilde{\boldsymbol{X}}\boldsymbol{W}. \tag{5}$$

公式(1)的节点嵌入学习公式改进为:

$$\boldsymbol{H}^{(l)}=\boldsymbol{\sigma}(\boldsymbol{\Theta}^{l-1}\boldsymbol{X}\boldsymbol{W}), \tag{6}$$

式中,$\boldsymbol{H}^{l}$ 是卷积层最后一层结果.

经过多层卷积后的节点嵌入再进行变换为:

$$\boldsymbol{Z}=\boldsymbol{\sigma}(f_1(\boldsymbol{H}^{l}))\odot\tanh(f_2(\boldsymbol{H}^{l})), \tag{7}$$

式中,f_1 是标准两层多头图注意力网络,f_2 是全连接层. 公式前半部分可以有侧重地学习信息,后半部分则作为一种非线性变换.

对文档图中的词嵌入进行如下池化,得文档嵌入 $\boldsymbol{Z}_G^i$:

$$\boldsymbol{Z}_G=\frac{1}{|\boldsymbol{V}|}\sum_{v\in V}\boldsymbol{Z}+\text{Maxpooling}(\boldsymbol{Z}), \tag{8}$$

最后基于词嵌入重建文档图邻接矩阵 $\widetilde{\boldsymbol{A}}_i$:

$$\widetilde{\boldsymbol{A}}_i=\text{sigmoid}(\boldsymbol{Z}_i^{T}\boldsymbol{Z}_i) \tag{9}$$

式中,$\boldsymbol{Z}_i$ 表示第 i 个文档嵌入.

1.5 文档图聚类

无监督文档分类既可保持文档嵌入的聚类结构,又可约束模型嵌入学习. 利用聚类算法根据 1.4 得到的文本图嵌入预测每个文档类别分布 $\boldsymbol{q}_{iu}$,通过优化其与预分配软标签分布 $\boldsymbol{p}_{iu}$ 的 *KL* 散度损失计算.

$$\boldsymbol{L}_c=\boldsymbol{KL}(\boldsymbol{P}\parallel\boldsymbol{Q})=\sum_i\sum_u p_{iu}\log\frac{\boldsymbol{p}_{iu}}{\boldsymbol{q}_{iu}}, \tag{10}$$

式中,$\boldsymbol{q}_{iu}$表示文档 i 嵌入 $\boldsymbol{z}_G^i$ 和聚类中心 $\boldsymbol{\mu}_u$ 之间的相似性,其计算公式如下:

$$\boldsymbol{q}_{iu}=\frac{(1+\parallel\boldsymbol{z}_G^i-\boldsymbol{\mu}_u\parallel^2)^{-1}}{\sum_k(1+\parallel\boldsymbol{z}_G^i-\boldsymbol{\mu}_k\parallel^2)^{-1}}, \tag{11}$$

$\boldsymbol{p}_{iu}$表示文档 i 分配的软隶属度,其计算公式如下:

$$\boldsymbol{p}_{iu}=\frac{\boldsymbol{q}_{iu}^2/\sum_i\boldsymbol{q}_{iu}}{\sum_k\left(\boldsymbol{q}_{ik}^2/\sum_i\boldsymbol{q}_{ik}\right)}. \tag{12}$$

由公式(12)可知,$\boldsymbol{P}$ 取决于每次迭代时更新的数据分布 $\boldsymbol{Q}$. $\boldsymbol{Q}$ 是不断变化的,每次迭代生成 $\boldsymbol{Q}$ 都更新 $\boldsymbol{P}$ 会导致模型不稳定,难以收敛. 为了避免这种情况,通常设计一个更新间隔 $\boldsymbol{t}$,每 $\boldsymbol{t}$ 次迭代更新一次 $\boldsymbol{P}$,t 的取值会根据不同数据集进行调整.

1.6 GGCN-DDC 模型优化

GGCN-DDC 模型在整个文档上重建损失公式如下:

$$L_r=\sum_{i=1}^{n}\text{loss}(\boldsymbol{A}_i,\widetilde{\boldsymbol{A}}_i), \tag{13}$$

式中,loss 表示二分类交叉熵损失,n 是文档数目.

但是实际上,直接按照这种思路将每个文档图的重建损失累加的效果并不好,而且计算效率低,无法充分利用 GPU 资源. 这里利用 pyg 框架中的数据容器 Dataloader 中的参数 mini-batch-size,重建关于多幅图之间的联合矩阵,能隐式地学习多个文档图之间的关系.

$$L_r = \sum_{n/m} \sum_{m} \text{loss}(\boldsymbol{A}_m, \widetilde{\boldsymbol{A}}_m), \tag{14}$$

式中,mini-batch-size=m 是参与重建联合矩阵的文本数目,$\boldsymbol{A}_m$ 可以直接由初始文档信息得到,利用 pytorch 框架下函数可以通过堆叠 m 个文本图嵌入得到 Z_m,再利用公式(9)即可生成 $\widetilde{\boldsymbol{A}}_m$,loss 函数同上. 这种计算增加了 GPU 一次处理的文档量,提升运算效率.

把文档图嵌入学习和文档图聚类这两部分的损失联合起来,得 GGCN-DDC 总损失函数如下:

$$L = L_r + \lambda L_c, \tag{15}$$

式中,L_r 是文档图联合重建损失,L_c 是文档图聚类损失,$\lambda \geqslant 0$ 是平衡系数.

模型学习算法通过“预训练+微调”的方式实现.

算法 1 GGCN-DDC 模型学习算法

输入:预处理文本图集合 G,更新间隔 t,迭代次数 $epoch=m$,预训练模型参数 $params$;
输出:文档节点嵌入和图嵌入以及文档类别标签.
a)读入预处理后的文档图信息,包括文档图节点初始嵌入和文档中节点连边关系;
b)读入预训练参数 $params$,在无梯度基础上运行一次模型,将所生成的初始图表示,输入 K-means 方法得到初始聚类中心 $\boldsymbol{\mu}_0$;
c)for $i=0$ to m:
d) 运行文档图词节点和图嵌入模块,获得图级别嵌入 $\boldsymbol{Z}_G$;
e) 运行文档图聚类模块,将文档嵌入 $\boldsymbol{Z}$ 连接起来,并结合上一次计算的聚类中心($\boldsymbol{\mu}_{i-1}$),根据公式(11)计算聚类分布 $\boldsymbol{Q}$;
f) If $i\%\ t=0$:
g) 根据公式(12)更新 $\boldsymbol{P}$;
h) 通过梯度下降法优化公式(15);
i)循环结束输出训练好的节点嵌入、图嵌入以及文档分类结果.

2 实验结果与分析

实验处理器采用 Intel(R)Xeon(R)CPU E5-2680 v4@ 2.40GHz,显卡为 RTX 3060(12GB),内存大小是 20GB,主要软件配置为 Python 3.7.13,pytorch 1.11.0.

2.1 数据集及评价指标

使用 3 种本文数据集进行实验[9],如表 1 所示.

表 1 数据集统计

Table 1 Datasets statistics

数据集	文档数	词数	类别数	平均长度
R8	7 674	7 688	8	41.25
3NG	600	12 569	3	118.17
6NG	1 200	16 306	6	115.92
R52	9 100	8 892	52	44.02

R8 和 R52 数据集是路透社新闻数据集的子集,8 个和 52 个类别,前者 5 485 个训练集样本和 2 189 个测试集样本,后者 6 532 个训练集样本和 2 568 个测试集. 20NG 数据集是新闻数据集,包含 18 846 个文档,20 个类别,按 3 类、6 类提取出了 3NG、6NG 两种数据集,每一类选择 200 个样本.

采用准确率(Acc)和标准化互信息指标 NMI[15]为实验结果的评价标准.

2.2 对比模型

据调研目前尚无明确的单纯针对纯文档数据集的聚类模型. 针对改进的表示学习和深度图聚类,设计两类对比模型.

a)无监督嵌入学习+聚类模型:采用目前流行的无监督嵌入学习方法,在单个文本图上学得文档词嵌入,经过与本文相同的池化操作得到文档图嵌入,将这些嵌入用 K-means 聚类.

DeepWalk[19]利用了图结构信息,通过生成随机游走序列来描述图中节点共现关系,再利用 SkipGram 方法最大化共现概率学习嵌入.

TADW[20]是一种矩阵分解的方法,在提取结构信息的基础上还提取了节点内容信息,以学习更好的节点表示.

GAE[21]将 GCN 应用到自编码器中,利用节点内容和节点结构信息来学习节点潜在表示.

b)深度图聚类模型:现有深度图聚类模型都是直接处理链接图数据,因此下面第一个模型将文本数据集处理成单文本图,用深度图聚类模型学习文本图嵌入再聚类;后两种模型构建 KNN 整体图,人工增加链接关系实现统一架构的文本图聚类.

AGC[14]对节点属性和图结构信息联合建模,通过改进 GCN 实现运用高阶图卷积来捕获图的全局结构特征,并且可以对不同的图来自适应地选择合适的阶数.

SDCN[22]针对无链接数据利用文档间余弦相似度制作 KNN 图来手工构造文档间链接关系,使用 GCN 编码后聚类.

DAEGC[16]通过带注意力机制的自编码器学习更科学的嵌入,同时利用自适应聚类方案联合优化嵌入学习和聚类过程. 这里我们参考 SDCN 中的 KNN 图制作方法,增加文档间链接关系,设置一种 DAEGC-K 方法用于本文对比实验.

2.3 超参数分析

文档图词节点学习部分,使用了 3 层改进图泛化卷积,其中卷积输入 300 个神经元,输出 96 个神经元,两层多头图注意力层分别是 64、16 个神经元. 图嵌入学习部分嵌入维度为 96 维. 实验优化算法采用 ADAM 算法,初始化参数部分采用 Xavier 方法,学习率为 0.001,dropout 率设为 0.5,batchsize 默认设置为 4 096,重建矩阵包含文档数目默认为 48,软隶属标签更新间隔默认为 2,训练 epoch = 200. 对于对比模型,除了最终文本嵌入维度选择与本实验相同的 96 维外,其余均按照各模型默认参数.

为了使模型效果达到最优,实验部分对模型关键参数进行对比调整,选择主要的两种超参进行调整,实验如下:

a)更新间隔的调整. 在 1.5 节聚类中心的更新间隔调整如表 2 所示,大多数据上 t 为 2,6NG 上为 5.

表 2 更新间隔调整对比实验

Table 2 Contrast experiments of update-interval

t	R8		3NG		6NG		R52	
	Acc	NMI	Acc	NMI	Acc	NMI	5-char	7-char
1	0.402 5	0.274 2	0.520 0	0.325 3	0.291 3	0.136 8	0.432 9	0.279 6
2	**0.668 2**	**0.393 6**	**0.537 5**	**0.350 1**	0.318 7	0.263 4	0.453 1	**0.318 3**
3	0.650 5	0.339 6	0.525 0	0.342 4	0.323 7	0.264 4	**0.465 5**	0.315 3
5	0.513 9	0.010 9	0.535 0	0.337 2	**0.352 5**	**0.291 7**	0.453 6	0.318 2
6	0.513 9	0.010 9	0.527 5	0.349 5	0.305 0	0.251 5	0.456 6	0.310 0

b)重建矩阵包含文档图数目的调整. 1.6 节重建矩阵包含文档图数目 mbs 调整如表 3 所示,在更新间隔 t 为 2(6NG 设为 5)的条件下,其他参数按默认设置,设置不同的 mbs 取值进行对比实验.

表 3 重建矩阵包含文本图数目对比实验

Table 3 Contrast experiments of MBS

mbs	R8		3NG		6NG		R52	
	Acc	NMI	Acc	NMI	Acc	NMI	5-char	7-char
1	0.513 7	0.116 8	0.518 2	0.304 2	0.351 2	0.291 3	0.473 2	0.242 5
16	0.581 2	0.244 2	0.520 3	0.293 1	0.356 3	0.283 4	**0.567 6**	**0.318 7**
32	**0.709 2**	**0.416 8**	**0.617 6**	**0.391 2**	**0.361 2**	**0.292 3**	0.431 4	0.285 7
48	0.664 1	0.384 7	0.538 4	0.347 1	0.350 8	0.295 8	0.443 2	0.293 3

由表 3 分析可知,在 mbs = 16 和 32 时,模型在各自数据集上均取得了较好的结果,因此对 R8 和 3NG 和 6NG,mbs 设为 32,而对 R52 数据集,mbs 设为 16. 通过观察发现,当 mbs 增大时,实验结果出现先增大

后减小的结果,这是因为重建矩阵包含文档图越多,就相当于迫使模型隐式地学到更多的文档间隐藏信息,促进文档分类. 而当 mbs 超过一定限度时,会使得模型学习受到过多无关文档的噪声影响,从而降低模型结果. 而 R52 数据集由于文档类别多,当包含文档数目变多时,会使模型更容易受到不同类别无关文档的噪声影响,表现在实验结果中,就是当 mbs = 16 时即相对较早的达到了最佳.

2.4 性能对比

为了验证本模型的性能,与 5 种对比模型在上述数据集上进行比较,对比结果如表 4 所示,其中粗体代表最佳结果.

a)TADW 模型在 4 个数据集上均取得与 DeepWalk 模型相当或较好的效果,这是因为 TADW 比 DeepWalk 多关注了图节点内容信息,说明了网络同时捕获内容及结构信息对于嵌入效果提升的有效性.

b)虽然 GAE 可以同时学习内容和结构信息,但是 DeepWalk 的结果比 GAE 要好,这说明基于矩阵分解的优化模型在面向本文这种无链接细粒度文档聚类任务时,要比自编码器这种非线性编码对文档的嵌入学习效果要好.

c)SDCN 在 4 个数据集的结果均比 AGC 要好,这说明了 SDCN 利用 KNN 图形成的文档链接大图,其文档间链接信息对本文这种文档聚类任务的效果有一定提升作用.

d)DAEGC-K 的方法相比两步式模型和 AGC,在 4 个数据集的表现都要好一些,这也说明了人工制作链接关系的作用. 但是它与 SDCN 一样,都比不上本文模型,说明本文这种采用构建联合重建矩阵隐式学习文档间关系比人工添加链接更有效.

e)AGC 的效果也没有比上述无监督嵌入学习+聚类模型更好的表现,说明本文利用现有的表示学习+聚类、深度图聚类+文本复现的两种设计,都不能很好的解决本文这种无链接文档集的无监督分类问题.

表 4 各对比模型在数据集上性能比较

Table 4 Performance comparision different models on datasets

模型/数据集	R8		3NG		6NG		R52	
	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI
DeepWalk	0.278 6	0.103 1	0.435 5	0.036 4	0.234 6	0.037 9	0.233 1	0.057 4
TADW	0.305 1	0.186 7	0.488 7	0.097 6	0.233 1	0.059 1	0.235 0	0.207 5
GAE	0.278 7	0.057 4	0.336 4	0.023 6	0.176 7	0.021 6	0.113 8	0.049 2
AGC	0.200 6	0.083 0	0.393 3	0.017 4	0.224 1	0.016 0	0.087 4	0.158 8
SDCN	0.523 1	0.342 1	0.465 0	0.083 8	0.246 7	0.042 2	0.455 9	0.259 1
DAEGC-K	0.604 2	0.278 9	0.496 7	0.098 0	0.275 8	0.077 1	0.451 3	0.215 3
GGCN-DDC	**0.709 2**	**0.416 8**	**0.617 6**	**0.391 2**	**0.361 2**	**0.292 4**	**0.567 6**	**0.318 7**

本模型在 4 个数据集上表现均优于其他模型,其中,本方法结果比 TADW 效果要好,说明本模型的改进可以克服 b)中提到的问题,使得对于这种单文本学习也能利用到非线性编码复杂信息的优势. 同时本模型的结果也优于 SDCN 和 DAEGC-K,说明了提取细粒度文档信息同时设计联合重建矩阵相对于简单构建链接文档整体图的优越性. 总的来说,本模型结合了两类对比方法的特长,在不使用额外先验知识的同时,利用改进的模型设计同时实现文档图嵌入学习和聚类,取得了比较明显的实验效果.

2.5 消融实验

本文针对 GGCN-DDC 主要的改进部分设计了 2 种变体来进行消融实验,以验证模型各改进组件的有效性,各变体的描述如下:

a)GGCN_1:GGCN-DDC 文本词及图级别嵌入模块改为普通卷积层提取数据,其余部分不变;

b)GGCN_2:GGCN-DDC 文本词及图级别嵌入模块保持不变,重建损失采用公式(13).

从图 2 可以看出,GGCN_1 采用的普通图卷积提取特征,效果不如 GGCN-DDC,这说明改进的泛化图卷积提取特征效果更好. GGCN_2 重建矩阵仅关注单个文档图,它的效果比 GGCN_DDC 要差,这说明了本模型对多个文档间信息的隐式提取可以提升分类性能. GGCN_2 普遍指标要高于 GGCN_1 模型,也说明了获取好的文本表示对分类的重要作用;这些变体的比较说明了本模型提出卷积层和重建损失两部分改进的有效性.

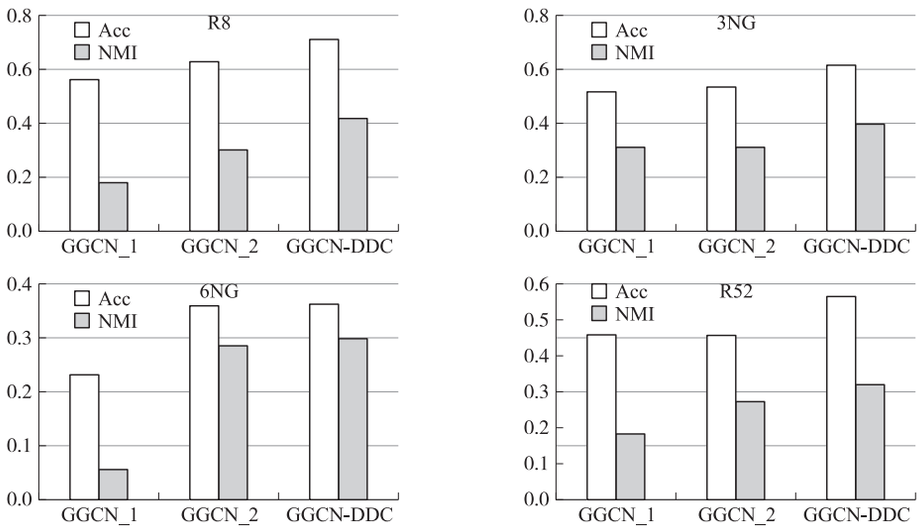

图 2　消融实验对比分析

Fig. 2　Comparative analysis of ablation experiments

3　结论

文本分类有许多场景,但现有方法大都需要标签. 本文构建了一个无监督深度文档聚类模型 GGCN-DDC,模型借鉴了 TextING 细粒度文本特征提取和 DAEGC 深度图聚类的优点,并在特征提取和模型优化方面进行了改进,最终实现适合于无链接文档语料库的无监督分类. 但由于没有标签的约束,仅从数据本身提取特征,而训练语料千差万别,导致学习效果有较大波动. 在未来,将考虑通过多任务、多种数据集的预训练模型或少标注数据来提升 GGCN-DDC 面向多种不同类型数据集的泛化性能.

[参考文献]

[1] KOSAR A,PAUW G D,DAELEMANS W. Unsupervised text classification with neural word embeddings[J]. Computational linguistics in the netherlands journal,2022(12):165-181.

[2] LI Q,PENG H,LI J X,et al. A survey on text classification:from traditional to deep learning[J]. ACM transactions on intelligent systems and technology,2022,13(2):41.

[3] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg,PA:ACL Press,2014:1746-1751.

[4] LIU P F,QIU X P,HUANG X J. Recurrent neural network for text classification with multi-task learning[C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence. Palo Alto,CA:AAAI Press,2016:2873-2879.

[5] 周玄郎,邱卫根,张立臣. 融合文本图卷积和集成学习的文本分类方法[J]. 计算机应用研究,2022,39(9):2621-2625.

[6] KIPF T N,WELLING M. Semi-supervised classification with graph convolutional networks[C]//Proceedings of the 5th International Conference on Learning Representations. (2017-02-22)[2023-03-10]. https://doi.org/10.48550/arXiv.1609.02907.

[7] YAO L,MAO C S,LUO Y. Graph convolutional networks for text classification[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto,CA:AAAI Press,2019:7370-7377.

[8] DAI Y,SHOU L J,GONG M,et al. Graph fusion network for text classification[J]. Knowledge-based systems,2022,236:107659.

[9] ZHANG Y F,YU X L,CUI Z Y,et al. Every document owns its structure:inductive text classification via graph neural networks[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg,PA:ACL Press,2020:334-339.

[10] CUI H Y,WANG G K,LI Y X,et al. Self-training method based on GCN for semi-supervised short text classification[J]. Information sciences,2022,611:18-29.

[11] HAJ-YAHIA Z,SIEG A,LéA A DELERIS. Towards unsupervised text classification leveraging experts and word embeddings [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg,PA:ACL Press,2019:371-379.

[12] SCHOPF T,BRAUN D,MATTHES F. Lbl2Vec:An embedding-based approach for unsupervised document retrieval on predefined topics[J/OL]. (2022-10-12)[2023-3-10]. https://doi.org/10.48550/arXiv.2210.06023.

[13] TIAN F,GAO B,CUI Q,et al. Learning deep representations for graph clustering[C]//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Palo Alto,CA:AAAI Press,2014:1293-1299.

[14] ZHANG X T,LIU H,LI Q M,et al. Attributed graph clustering via adaptive graph convolution[C/OL]. (2019-08-01)[2023-3-10]. https://doi.org/10.24963/ijcai.2019/601.

[15] ZHU D Y,CHEN S D,MA X H,et al. Adaptive graph convolution using heat kernel for attributed graph clustering[J]. Applied sciences,2020,10(2):1473.

[16] WANG CC,PAN S R,HU R Q,et al. Attributed graph clustering:a deep attentional embedding approach[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Palo Alto,CA:AAAI Press,2019:3670-3676.

[17] PENNINGTON J,SOCHER R,MANNING C. Glove:Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg,PA:ACL Press,2014:1532-1543.

[18] CUI G Q,ZHOU J,YANG C,et al. Adaptive graph encoder for attributed graph embedding[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York:ACM Press,2020:976-985.

[19] PEROZZI B,AI-RFOU R,SKIENA S. Deepwalk:online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM Press,2014:701-710.

[20] YANG C,LIU Z Y,ZHAO D L,et al. Network representation learning with rich text information[C]//Proceedings of the 24th International Joint Conference on Artificial Intelligence. Palo Alto,CA:AAAI Press,2015:2111-2117.

[21] KIPF T N,WELLING M. Variational graph auto-encoders[C/OL]//Proc of 30th Conference on Neural Information Processing Systems Workshop on Bayesian Deep Learning. (2016-11-21)[2023-3-10]. https://doi.org/10.48550/arXiv.1611.07308.

[22] BO D Y,WANG X,SHI C,et al. Structural deep clustering network[C]//Proceedings of the Web Conference 2020. New York:ACM Press,2020:1400-1410.

[责任编辑:陆炳新]