

基于BERT-Prototypical模型的小样本科技文献分类方法

白文清, 崔彩霞

(太原师范学院 计算机科学与技术学院, 山西 晋中 030619)

摘要: 由于学科的不断细化和学科间发展速度的不均衡, 个别学科可用于分类训练的数据极少, 为科技文献分类工作带来了一定困难。为此, 针对科技文献长尾问题严重且传统文本分类方法已经无法取得更好分类效果的问题, 提出一种基于BERT-Prototypical模型的小样本科技文献分类方法。该模型以迁移学习中的原型网络为基础, 首先借助BERT预训练模型深入挖掘科技文献文本间的关系以获得更好的特征表示; 然后将编码后的文本特征输入到原型网络中, 通过优化原型网络的编码方式和参数设置提高科技文献分类效果。实验结果表明, 在5-way 20-shot任务中, BERT-Prototypical模型的分类准确率达到95.6%; 在样本有限的5-way 5-shot任务中, BERT-Prototypical模型的分类准确率可达78.4%, 相较对照模型的分类效果有所提升。

关键词: 科技文献分类; 小样本学习; 原型网络; BERT模型; 不平衡数据

DOI: 10.11907/rjdk.241517

开放科学(资源服务)标识码(OSID):

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2025)004-0042-06



Few-shot Scientific and Technological Literature Classification Method Based on BERT-Prototypical Model

BAI Wenqing, CUI Caixia

(College of Computer Science and Technology, Taiyuan Normal University, Jinzhong 030619, China)

Abstract: Due to the continuous refinement of disciplines and the uneven development speed between disciplines, there is very little data available for classification training in individual disciplines, which brings certain difficulties to the classification of scientific literature. To address the serious problem of long tail in scientific literature and the inability of traditional text classification methods to achieve better classification results, a small sample scientific literature classification method based on BERT Prototypical model is proposed. This model is based on the prototype network in transfer learning, and first uses the BERT pre trained model to deeply explore the relationships between scientific literature texts to obtain better feature representations; Then input the encoded text features into the prototype network, and improve the classification performance of scientific literature by optimizing the encoding method and parameter settings of the prototype network. The experimental results show that in the 5-way 20 shot task, the classification accuracy of the BERT Prototypical model reaches 95.6%; In the 5-way 5-shot task with limited samples, the classification accuracy of the BERT Prototypical model can reach 78.4%, which is improved compared to the control model.

Key Words: scientific and technological literature classification; few-shot learning; prototype networks; BERT model; imbalanced data

0 引言

近年来,随着科研水平的快速提高,各种研究成果丰富

多彩,大量科技文献也随之产生。然而,不同学科之间发展速度的不均衡导致相应科技文献的产出数量严重失衡。中国科学技术信息研究所于2023年9月发布的《中国科技论文统计报告(2)——2023中国卓越科技论文产出状况报

收稿日期: 2024-06-10

扫描二维码阅读全文:



基金项目: 山西省基础 Research 计划(自由探索)项目(20210302123334)

作者简介: 白文清(1999-),女,CCF会员,太原师范学院计算机科学与技术学院硕士研究生,研究方向为机器学习、数据挖掘;崔彩霞(1974-),女,博士,太原师范学院计算机科学与技术学院副教授、硕士生导师,研究方向为机器学习、数据挖掘。本文通讯作者: 崔彩霞。

告》显示,2022年临床医学学科共发表卓越论文82 104篇,占统计范围内39个学科卓越论文发表总量的15%;计算机技术学科发表卓越论文31 727篇,占比6%;而天文学、核科学技术等学科的卓越论文数量均不足1 000篇,安全科学技术学科更是仅发表卓越论文617篇,占比近乎为0。这种文献总量大但个别学科文献数量极少的情况为文献分类管理工作带来了极大挑战。

科技文献数据规模庞大,内容高度专业化和领域化,分类难度较大,且存在部分类别文献数量极少、长尾效应明显的情况,传统文本分类方法已无法取得较好的分类效果。作为一种针对数据量较少的文本分类问题而设计的算法,小样本文本分类算法可以更有效地利用有限的资源解决科技文献分类中的数据稀疏、标注困难等问题,提高分类准确性和效率。对稀缺文献的分类与挖掘可为相关科研工作者提供更全面的信息服务,推动科技文献资源的合理利用与传播。因此,研究小样本文本分类方法具有积极意义。

1 相关研究

1.1 小样本学习

小样本学习的概念最早起源于计算机视觉领域,并在图像分类领域中受到广泛关注,现已有很多性能优异的算法模型。然而在自然语言处理领域,小样本学习的发展还较为缓慢。最初,小样本学习的研究方法多基于模型微调。例如,Howard等^[1]曾在2018年提出一个通用微调语言模型ULMFit,通过模型层次(特征独特性越高,所在层次越高)和同层迭代次数两个维度改变学习速率来微调语言模型;Nakamura等^[2]提出一种新的微调方法,借助自适应梯度优化器,通过在再训练过程中使用更低的学习率,或者必要时调整整个网络实现模型微调。为解决模型微调中可能产生的过拟合问题,Finn等^[3]率先提出使用元学习模型(Model-Agnostic Meta-Learning, MAML)进行半监督学习,利用无标签数据调整嵌入函数的参数,采用带标签数据调整分类器参数。这种基于数据增强的方法一定程度上解决了模型微调可能带来的过拟合问题,并提高了样本的多样性。随后,Wang等^[4]通过将元学习与数据生成相结合生成虚拟数据来解决小样本问题;Hou等^[5]采用交叉注意力网络使所提取到的特征更具鉴别性。基于增强样本特征空间的思路,Liu等^[6]提出特征迁移网络,用于描述物体姿态变化引起的运动轨迹变化;Chen等^[7]采用双向网络对图像特征进行增强;Shen等^[8]通过将固定的注意力机制替换成不确定的注意力机制来优化分类网络。

迁移学习作为一种新兴的学习框架,也逐渐被应用到小样本学习中,并取得不错表现。例如,Koch等^[9]提出使用孪生神经网络进行单样本图像识别;Wang等^[10]采用回归网络解决小样本问题;Vinyals等^[11]采用匹配网络解决单样本学习问题;Snell等^[12]采用原型网络(Prototypical Networks)解决小样本问题;Ren等^[13]在原型网络的基础上

使用半监督学习的思想改善分类边界;Sung等^[14]提出一种基于关系网络的通用小样本学习框架;在关系网络的基础上,Zhang等^[15]提出一种基于度量学习的小样本学习算法,与关系网络只对特征提取网络最后输出的特征图进行相关性计算不同,该算法对特征提取网络中间阶段的输出结果均进行了相关性计算。此外,还有协方差度量网络、深度最近邻神经网络、基于记忆增强的神经网络、基于注意力机制的未知任务元学习方法都被用来解决自然语言处理领域的小样本学习问题^[16]。

1.2 科技文献分类

科技文献分类是指将学术文献分类到指定类别,属于典型的文本分类问题^[17]。然而,科技文献中包含大量专业术语,存在众多跨学科文献,如何准确理解这些术语的含义、全面考虑文本内容,以便将其归入正确类别,这需要利用自然语言处理技术对文本内容进行深入挖掘。纵观科技文献分类领域相关研究,可将分类方法分为基于规则的方法、基于统计学习的方法和基于深度学习的方法3类。在最初的基于规则的分类方法中,周丽红等^[18]针对科技文献的结构特性,通过关联规则将其划分为综述型、理论型和应用型。这种分类方法实现简单、可解释性强,但需要文献分类专家整理大量规则,且无法直接迁移到其他分类。随后,基于统计学习的方法开始兴起。例如,何浩等^[19]提出N-Gram语言模型,采用纯统计分析的方法实现中文文献的自动分类;王方等^[20]提出一种基于向量空间模型的科技文献自动分类方法,首先构建分类代表文本,计算单文本与分类代表文本之间的相似度,然后利用训练集计算出分类归属度,从而判断某一文本属于某种类别的可能性。相较于基于规则的方法,基于统计学习的方法分类速度更快、准确率更高,但仍需要专家设计特征以及采用大量数据训练。近年来,基于深度学习的文本分类方法发展迅速,弥补了这一不足。例如,谢红玲等^[21]基于TextCNN进行文献分类;雷兵等^[22]首先基于卷积神经网络生成词向量矩阵,然后经过卷积层、池化层与Softmax层进行文献分类;张智雄等^[23]借助大规模预训练语言模型BERT(Bidirectional Encoder Representations from Transformers)构建多层分类器集群来实现精准分类。针对文献分类层次性的特点,赵旻等^[24]使用多层分类器实现中文医学文献的分类;张晓丹^[25]利用文献中的词和句子等信息构建关系图,基于图神经网络实现文本分类。然而,上述方法均需要大量标注数据的支持,即每个文献类别都需要有充足的训练数据,因此并不适用于文献分类中的低资源场景。为此,黄金凤等^[26]通过引入和改进多种数据增强策略提升小样本类别的数据多样性与分类鲁棒性,从而实现科技文献的不平衡分类。

综上所述,为解决科技文献分类问题,众多学者提出了创新性方法,但未在小样本分类算法方面有深入研究。为此,本文将小样本分类算法应用于科技文献分类领域,利用基于迁移学习的原型网络改进现有科技文献分类方法,提出一种基于BERT-Prototypical模型的小样本科技文

献分类方法。为从有限的样本数据中挖掘更多有效信息,提高分类方法的准确性和泛化能力,该方法引入BERT预训练模型优化原型网络的编码方式,深入挖掘文本的上下文关系,以获取更丰富的特征信息;然后,将这些特征输入到以任务为单位训练更新的原型模型中(任务包含支持集和查询集)。在训练过程中,通过计算查询集中样本与支持集中同类样本得到的原型向量之间的距离判断样本类别,最终完成科技文献的分类任务。

2 分类模型构建

2.1 BERT 模型

2018年,Google推出预训练语言模型BERT,其核心理念是通过大规模的无监督预训练来学习通用的文本表征,并可在特定任务上进行微调以提升性能。BERT的预训练过程包括MLM(Masked Language Model)和NSP(Next Sentence Prediction)两个关键任务阶段。在MLM中,模型会随机mask掉输入文本中的部分单词,然后尝试预测这些被mask掉的单词,从而促使模型在句子上下文中考虑缺失信息,更好地学习单词之间的关联。NSP则旨在判断两个句子是否相邻,有助于BERT学习句子之间的逻辑关系,提升对文本语境的理解水平。

BERT模型由多层双向Transformer编码器堆叠而成,同时结合Self-Attention机制和前馈神经网络。这种设计使其能够同时捕捉文本的局部和全局信息,实现更强大的文本表征学习功能。BERT模型在自然语言处理任务中取得了巨大成功。

2.2 原型网络

原型网络是一种用于解决监督学习问题的神经网络架构,其核心思想是假设存在一个嵌入空间,在这个空间中,每个类别的点都围绕着该类别的原型(类别中心点)聚集。首先,使用深度神经网络将输入映射为向量,对于同属一个类别的样本,求得其向量平均值作为该类别的原型;然后,通过持续训练模型和最小化损失函数,使得同一类别内的样本距离更为接近,不同类别的样本更为远离,从而更新嵌入函数的参数;最后,通过找到最近的类原型对嵌入的查询点执行分类操作。

原型网络通过学习少量样本来捕捉类别之间的相似性,从而展现出对新类别的良好泛化能力。因此,其常被应用于小样本学习和零样本学习等任务,传统深度学习方法在这些任务中通常表现不佳。

2.3 BERT-Prototypical 模型

本文提出的BERT-Prototypical模型的分类过程包括数据预处理、BERT模型编码、原型网络分类3个阶段。具体结构如图1所示。

2.3.1 数据预处理阶段

对科技文献数据集CNKI_Chinese_paper_dataset中的原始数据进行清洗、扩充和筛选,保留论文的标题和关键

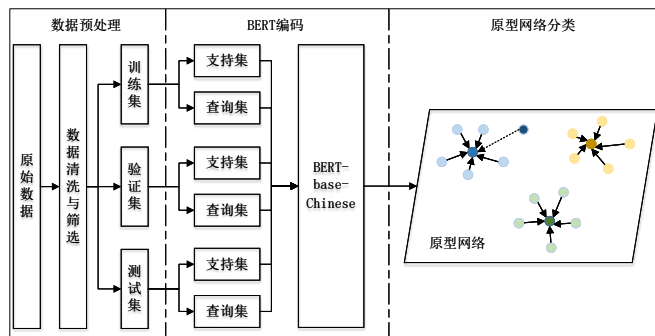


Fig. 1 Structure of the BERT-Prototypical model

图1 BERT-Prototypical 模型结构

字作为样本数据,整合同类别数据,去除数据中的缺失值、重复值和特殊符号,补充极个别类的样本数量以满足实验需求^[27]。将处理后的实验数据按照6:2:2的比例划分为训练集、验证集和测试集。

2.3.2 BERT 模型编码阶段

在每个元任务中,分别从训练集、验证集和测试集中以实验设置的n-way k-shot参数采样数据并划分相应的支持集和查询集。借助从HuggingFace资源库中下载的BERT-base-Chinese预训练模型对划分后的数据进行编码。

BERT模型在编码时首先将文本转换成词向量,然后经过多层Transformer(每一层都包括自注意力机制和前馈神经网络)的处理得到对文本的深层次表示,实现子词级别的文本特征编码。具体来说,BERT模型首先对输入文本进行分词,将字转换为固定维的向量表示,得到一个词汇表中的token序列,同时在序列开头添加一个特殊的[CLS]标记作为分类任务的输入表示,在每个句子的末尾添加[SEP]标记用于区分两个句子的向量表示;然后,引入段嵌入指定该句子来自于哪个句段,引入位置嵌入来指定序列中单词的位置,以帮助模型理解单词的顺序;最后,将经过分词和位置嵌入处理后的token序列输入到双向Transformer模型中进行编码。编码过程如图2所示。

2.3.3 原型网络分类阶段

基于编码后的样本数据,在训练时首先通过支持集中的样本计算各类别原型向量;然后计算查询集中样本数据与各原型之间的距离,通过距离大小判断数据所属类别。其中,训练集用于训练模型超参数,并根据验证集的表现进行调整。使用测试集评估模型在未见数据上的表现,以检验其泛化能力。原型网络分类过程如图3所示。

分类判别计算过程为:设数据集 $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$,其中 x_i 表示样本, y_i 表示样本对应的标签(共有 c 个类别);支持集 S 由类别 k 中采样的 n 个样本构成,查询集 Q 由去掉构成支持集样本的类别 k 中的 n 个样本构成。

首先计算各类别的原型向量 $proto_k$,取支持集特征空间中同类样本的均值作为该类别的原型。计算公式为:

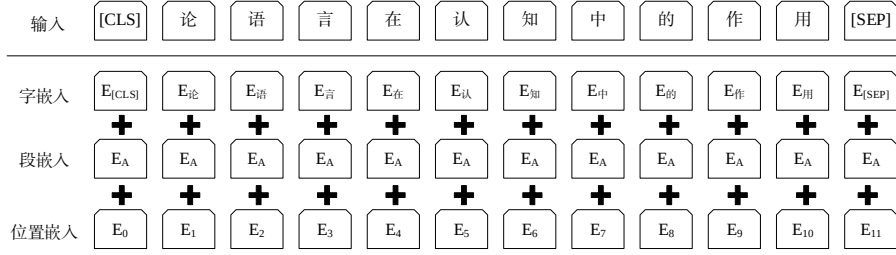


Fig. 2 BERT model coding process

图2 BERT模型编码过程

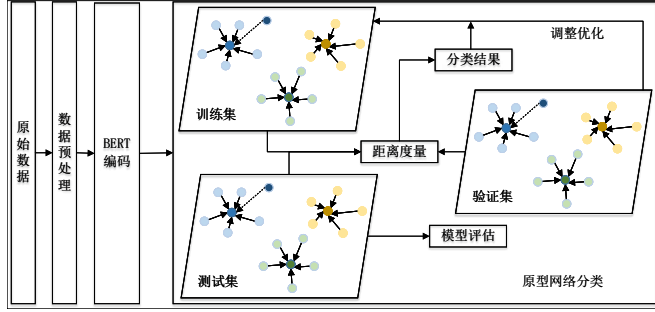


Fig. 3 Classification process of prototype network

图3 原型网络分类过程

$$proto_k = \frac{1}{|s_k|} \sum_{((x_i, y_i) \in s_k)} f_{Bert}(x_i) \quad (1)$$

式中: s_k 表示支持集 S 中标记为类别 k 的一组样本; $f_{Bert}(x_i)$ 表示编码映射函数。

然后针对查询集中的样本,采用欧几里得距离计算出其与各原型向量的距离 $d(f_{Bert}(x_i), proto_k)$ 。计算公式为:

$$d(f_{Bert}(x_i), proto_k) = \sqrt{\sum_{i=1}^n \sum_{k=1}^c (f_{Bert}(x_i) - proto_k)^2} \quad (2)$$

最后通过 Softmax 函数计算出当前样本属于每个分类的概率。计算公式为:

$$p_\phi(y = k|x) = \frac{\exp(-d(f_{Bert}(x), proto_k))}{\sum_{k'} \exp(-d(f_{Bert}(x), proto_{k'}))} \quad (3)$$

在训练过程中,通过随机梯度下降最小化负对数概率损失函数。计算公式为:

$$J(\phi) = -\log(p_\phi(y = k|x)) \quad (4)$$

3 实验方法与结果分析

3.1 实验环境与参数设置

实验硬件配置: GPU 为 NVIDIA Corporation Device 2204, 显存容量为 24 G, 内存容量为 251 G; 实验软件环境: Python3.9, CUDA12.0, Pytorch2.0.1。

模型采用 Adam 作为优化算法, 学习率为 $10e-3$, 衰减步长为 1, epoch 设为 600, Episode 设为 10。在每个 Episode 中各选取每类 5、10、20 个样本构成支持集和查询集, 以便更好地训练和评估模型。为避免过拟合, patience 设为 60。

3.2 数据集选取与预处理

使用 CNKI_Chinese_paper_dataset 数据集。该数据集

广泛搜集爬取了 24 000 多篇发表于中国知网的中文论文信息, 包含期刊名、作者、标题、关键词、摘要、中国图书分类号、学科领域、出版年月等 8 个字段。由于文献分类号的设置是基于《中国图书馆分类法》, 由论文作者或期刊编辑确定, 具有较高的类别标记准确性, 论文的数据标签依据其文献分类号确定。

为了优化数据质量, 本文对基础数据进行简单筛选和清洗。首先去除论文期刊名、作者等信息, 保留标题和关键字作为样本数据; 然后整合同类数据并剔除样本数据中的特殊字符和重复内容。最终筛选得到 22 500 条论文数据, 涵盖了 21 个不同类别的样本。具体类别分布情况如图 4 所示, 其中 A 表示马克思主义、列宁主义、毛泽东思想、邓小平理论; B 表示哲学、宗教; C 表示社科总论; D 表示政治、法律; E 表示军事; F 表示经济; G 表示文化、科学、体育; H 表示语言、文字; I 表示文学; J 表示艺术; K 表示历史、地理; N 表示自科总论; O 表示数理科学与化学; P 表示天文、地球科学; Q 表示生物科学; R 表示医药、卫生; S 表示农业科学; T 表示工业技术; U 表示交通运输; V 表示航空航天; X 表示环境科学、安全科学。然而, CNKI_Chinese_paper_dataset 数据集的论文数据存在严重的不平衡现象, 个别类别样本数据甚至不足以满足 5-way 20-shot 实验任务的需求。因此, 本研究对数据集中极个别类别进行一定量的扩充, 以确保所有类别的样本数量均不少于 50 条。后续实验基于清洗和处理后的 CNKI_Chinese_paper_dataset 数据集进行, 最终实验数据总量为 22 651 条, 按照 6:2:2 的比例划分训练集、验证集和测试集。

3.3 评价指标

由于实验针对的是不平衡数据集下的多分类任务, 为避免因不平衡数据而导致某些类别被忽视, 本文采用宏平均 (macro-average) 下的准确率 (accuracy)、精确率 (Precision) 和 F1 值作为评价指标。准确率指分类正确的样本数占总样本数的比例; 精确率指预测为正类别的样本中真正为正类别的比例; F1 值综合了精确率和召回率的性能, 取调和平均值, 从而更全面地评估模型的分类效果。计算公式为:

$$m_acc = \frac{1}{n} \sum_{i=1}^n \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i} \quad (5)$$

$$m_Pre = \frac{1}{n} \sum_{i=1}^n \frac{TP_i}{TP_i + FP_i} \quad (6)$$

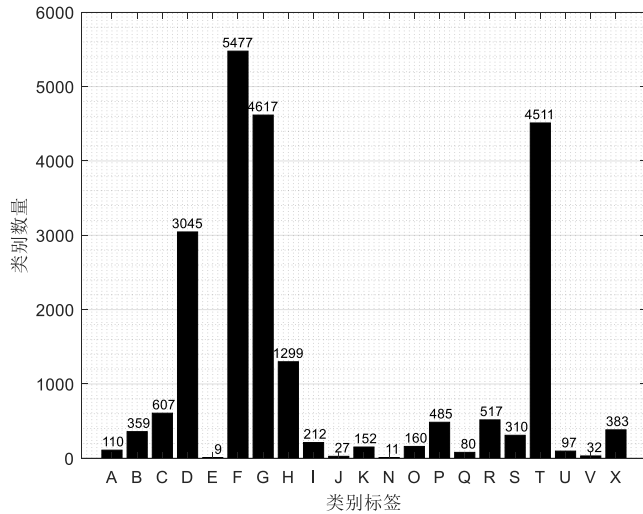


Fig. 4 Category distribution of CNKI_Chinese_paper_dataset

图4 CNKI_Chinese_paper_dataset数据集类别分布

$$m_F1 = \frac{2TP_i}{2TP_i + FP_i + FN_i} \quad (7)$$

式中:对于每个类别(类别标签为*i*), TP_i 表示模型正确预测为类别*i*的样本数量; TN_i 表示模型正确预测为非类别*i*的样本数量; FP_i 表示模型错误预测为类别*i*的其他类别样本数量; FN_i 表示模型错误预测为其他类别而实际类别为*i*的样本数量。

3.4 实验结果分析

3.4.1 比较实验

为证实 BERT-Prototypical 模型对小样本文献的分类效果,选取以下 4 种方法进行比较:①匹配网络。使用卷积神经网络提取支持集和目标样本的特征,计算目标样本的特征向量与支持集中每个样本特征向量之间的相似度,基于计算出的相似度,使用注意力机制为支持集中的每个样本分配一个权重,根据加权后的支持集标签预测目标样本的类别;②关系网络。使用卷积神经网络对输入样本进行特征提取,在特征空间中计算输入样本之间的关系评分以表示其相似性,其中特征提取器和距离度量均包含可训练的参数,可以通过监督学习得到;③TextCNN。使用卷积操作提取不同尺寸的 *n*-gram 特征,并通过最大池化(Max Pooling)操作选择最显著的特征,然后拼接成一维向量,通过全连接层进行分类;④IPMT(Intermediate Prototype Mining Transformer)^[28]。引入中间原型缓解支持集和查询集样本之间的类别信息差距。在每个 IPMT 层中,通过将支持特征和查询特征中的对象信息传播到原型中获得中间原型,然后采用中间原型激活查询特征映射,通过迭代执

行逐步得到改进的中间原型和查询特征,最后利用优化后的查询特征进行精确的分类预测。

为保持实验条件的一致性,减少不同训练样本数量对分类结果的影响,所有小样本方法均以 5-way 5-shot 为基准进行实验,所有卷积操作均采用 4 个卷积块。实验结果如表 1 所示。可以看出,在保持一致实验条件的情况下, BERT-Prototypical 模型在各评价指标上的表现均明显优于匹配网络和关系网络这两种经典小样本模型。相较于以 5-way 5-shot 作为基准实验的匹配网络和关系网络, TextCNN 模型表现更佳,这是由于处理后的文献数据中短文本数量较多, TextCNN 可通过卷积操作捕获短文本中的特征,训练中可利用的样本数据更为充足。 IPMT 模型在一定程度上弥补了支持集和查询集之间的类别信息差距,取得了比 TextCNN 更好的准确率,但仍不及 BERT-Prototypical 模型。 BERT-Prototypical 模型通过引入 BERT 编码模块挖掘文本的上下文关系,获取了更丰富的特征信息,弥补了其他小样本模型在挖掘文本上下文信息方面的不足,在各评价指标方面均表现最佳。

Table 1 Comparative experimental results

表1 比较实验结果

模型	m_acc	m_Pre	m_F1
匹配网络	0.485 6	0.404 6	0.414 5
关系网络	0.537 2	0.454 5	0.471 1
TextCNN	0.614 2	0.574 2	0.587 9
IPMT	0.683 9	0.708 1	0.671 6
BERT-Prototypical	0.784 0	0.797 6	0.777 5

3.4.2 消融实验

为验证 BERT-Prototypical 模型各模块的有效性,本文进行了消融实验。同时,为证明所提模型在不同训练样本数量下均具有较高的分类准确率,在与原型网络(Prototypical 模型)、BERT 模型的比较实验中使用 5-way *k*-shot (*k*=5, 10, 20)作为元任务。实验结果如表 2 所示。可以看出,随着元任务中训练样本数量的增加, BERT 模型、Prototypical 模型和 BERT-Prototypical 模型分类效果均有所提高,这表明随着支持集规模的扩大,模型性能得到相应提升。然而,在数据集极度不平衡且元任务训练样本较少的情况下,即使 BERT 模型可以获得更有效的向量表示,但其分类准确率仅为 35.49%,针对小样本问题提出的 Prototypical 模型亦只取得 50.00% 的准确率。 BERT-Prototypical 模型综合了 BERT 模型和 Prototypical 网络的优势,在训练样本数量较少的情况下依然取得 78.40% 的分类准确率,证明了模型中各模块的有效性。

Table 2 Ablation test results

表2 消融实验结果

模型	5-way 5-shot			5-way 10-shot			5-way 20-shot		
	m_acc	m_Pre	m_F1	m_acc	m_Pre	m_F1	m_acc	m_Pre	m_F1
BERT	0.354 9	0.290 9	0.276 1	0.475 7	0.349 7	0.374 7	0.648 4	0.524 5	0.576 7
Prototypical	0.500 0	0.465 2	0.460 4	0.610 0	0.636 8	0.591 2	0.793 3	0.796 6	0.772 3
BERT-Prototypical	0.784 0	0.797 6	0.777 5	0.830 0	0.859 2	0.815 8	0.956 0	0.964 3	0.955 3

4 结语

为解决科技文献分类的长尾问题,本文创新性地利用基于迁移学习的原型网络改进现有文本分类方法,克服数据稀疏性和样本不平衡性带来的挑战,同时引入BERT预训练模型作为原型网络的补充,克服其无法深入挖掘文本上下文关系,从而获取更丰富特征信息的问题。所提模型在科技文献分类任务中具有良好的准确性和泛化能力,在训练样本有限的情况下仍表现良好。目前本文仅根据中图分类号中的大类对科技文献进行了分类,未来会针对大类中的子类进行更深入的研究,辨析子类间的相互关系,实现更精准的文献分类。

参考文献:

- [1] HOWARD J, RUDER S. Universal language model fine-tuning for text classification[DB/OL]. <https://arxiv.org/abs/1801.06146>.
- [2] NAKAMURA A, HARADA T. Revisiting fine-tuning for few-shot learning[DB/OL]. <https://arxiv.org/abs/1910.00216>.
- [3] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//International Conference on Machine Learning, 2017: 1126-1135.
- [4] WANG Y X, GIRSHICK R, HEBERT M, et al. Low-shot learning from imaginary data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7278-7286.
- [5] HOU R, CHANG H, MA B, et al. Cross attention network for few-shot classification[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019: 4003-4014.
- [6] LIU B, WANG X, DIXIT M, et al. Feature space transfer for data augmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 9090-9098.
- [7] CHEN Z, FU Y, ZHANG Y, et al. Semantic feature augmentation in few-shot learning[DB/OL]. <https://arxiv.org/abs/1804.05298>.
- [8] SHEN W, SHI Z, SUN J. Learning from adversarial features for few-shot classification[DB/OL]. <https://arxiv.org/abs/1903.10225>.
- [9] KOCH G, ZEMEL R, SALAKHUTDINOV R. Siamese neural networks for one-shot image recognition[C]//Lille: 32nd International Conference on Machine Learning, 2015.
- [10] WANG Y X, HEBERT M. Learning to learn: model regression networks for easy small sample learning[C]//14th European Conference on Computer Vision, 2016: 616-634.
- [11] VINIYALS O, BLUNDELL C, LILLICRAP T, et al. Matching networks for one shot learning[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016: 3637-3645.
- [12] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 4080-4090.
- [13] REN M, TRIANTAFILLOU E, RAVI S, et al. Meta-learning for semi-supervised few-shot classification [DB/OL]. <https://arxiv.org/abs/1803.00676>.
- [14] SUNG F, YANG Y, ZHANG L, et al. Learning to compare: relation network for few-shot learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1199-1208.
- [15] ZHANG X, SUNG F, QIANG Y, et al. Deep comparison: relation columns for few-shot learning[DB/OL]. <https://arxiv.org/abs/1811.07100>.
- [16] ZHAO K L, JIN X L, WANG Y Z. Survey on few-shot learning[J]. Journal of Software, 2020, 32(2): 349-369.
- [17] 赵凯琳, 靳小龙, 王元卓. 小样本学习研究综述[J]. 软件学报, 2020, 32(2): 349-369.
- [18] PANG G S, JIANG S Y. A summary of research on automatic text classification technologies [J]. Information Studies: Theory & Application, 2012, 35(2): 123-128.
- [19] 鹿观松, 蒋盛益. 文本自动分类技术研究综述[J]. 情报理论与实践, 2012, 35(2): 123-128.
- [20] ZHOU L H, LIU K. Scientific literature classification based on associated rules[J]. Library and Information Service, 2012, 56(4): 12-16, 119.
- [21] 周丽红, 刘勘. 基于关联规则的科技文献分类研究[J]. 图书情报工作, 2012, 56(4): 12-16, 119.
- [22] HE H, YANG H T. Approach of chinese document automatic classification based on the frequency of n-gram[J]. Journal of the China Society for Scientific and Technical Information, 2002(4): 421-427.
- [23] 何浩, 杨海荣. 一种基于N-Gram技术的中文文献自动分类方法[J]. 情报学报, 2002(4): 421-427.
- [24] WANG F, RUAN M H, ZHU H G, et al. Research on vector space model-based automatic classification of sci-tech document [J]. Information Research, 2013(12): 1-3, 8.
- [25] 王方, 阮梅花, 朱海刚, 等. 基于向量空间模型的科技文献自动分类研究[J]. 情报探索, 2013(12): 1-3, 8.
- [26] XIE H L, FENG G H, HE W L. Research on semantic classification of scientific and technical literature based on deep learning[J]. Information Studies: Theory & Application, 2018, 41(11): 149-154.
- [27] 谢红玲, 奉国和, 何伟林. 基于深度学习的科技文献语义分类研究[J]. 情报理论与实践, 2018, 41(11): 149-154.
- [28] LEI B, LIU X, ZHONG Z. Research on fine-grain classification method of academic literature based on bibliographies [J]. Library and Information Service, 2021, 65(14): 128-137.
- [29] 雷兵, 刘小, 钟镇. 基于题录信息的领域学术文献细粒度分类方法研究[J]. 图书情报工作, 2021, 65(14): 128-137.
- [30] ZHANG Z X, ZHAO Y, LIU H. Construction of a practical application-oriented automatic classification engine for scientific literature [J]. Journal of Library Science in China, 2022, 48(4): 104-115.
- [31] 张智雄, 赵旸, 刘欢. 构建面向实际应用的科技文献自动分类引擎[J]. 中国图书馆学报, 2022, 48(4): 104-115.
- [32] ZHAO Y, ZHANG Z X, LIU H. A research on automatic classification of Chinese medical literature based on hierarchical classification [J]. Research on Library Science, 2021(21): 49-55, 61.
- [33] 赵旸, 张智雄, 刘欢. 基于层次分类法的中文医学文献分类研究[J]. 图书馆学研究, 2021(21): 49-55, 61.
- [34] ZHANG X D. The application of improved graph convolutional neural network in big data classification of scientific and technological documents [J]. Journal of Information, 2021, 40(1): 184-188.
- [35] 张晓丹. 改进的图神经网络文本分类模型应用研究——以NSTL科技期刊文献分类为例[J]. 情报杂志, 2021, 40(1): 184-188.
- [36] HUANG J F, GAO Y, XU T, et al. Research of imbalanced classification for technical documents based on few-shot data augmentation [J]. Frontiers of Science and Technology of Engineering Management, 2022, 41(3): 23-30.
- [37] 黄金凤, 高岩, 徐童, 等. 基于小样本数据增强的科技文档不平衡分类研究[J]. 工程管理科技前沿, 2022, 41(3): 23-30.
- [38] JIANG Y T, HU R F. Influence of NLP on other fields based on data mining of CNKI Chinese papers [J]. Journal of Information, 2021, 40(12): 169-176.
- [39] 蒋彦廷, 胡初奋. 自然语言处理在其他学科领域的影响考察——基于CNKI的中文文献挖掘[J]. 情报杂志, 2021, 40(12): 169-176.
- [40] LIU Y W, LIU N, YAO X W, et al. Intermediate prototype mining transformer for few-shot semantic segmentation [J]. Advances in Neural Information Processing Systems, 2022, 35: 38020-38031.

(责任编辑:尹晨茹)