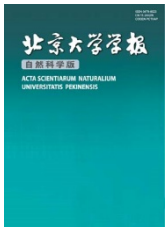

北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
ISSN 0479-8023,CN 11-2442/N

《北京大学学报(自然科学版)》网络首发论文

题目：增强提示学习的少样本文本分类方法
作者：李睿凡，魏志宇，范元涛，叶书勤，张光卫
DOI：10.13209/j.0479-8023.2023.071
收稿日期：2023-05-18
网络首发日期：2023-10-09
引用格式：李睿凡，魏志宇，范元涛，叶书勤，张光卫．增强提示学习的少样本文本分类方法[J/OL]．北京大学学报(自然科学版).
https://doi.org/10.13209/j.0479-8023.2023.071

北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2023.071

增强提示学习的少样本文本分类方法

李睿凡[1,2,3,†]　魏志宇[1]　范元涛[1]　叶书勤[1]　张光卫[2,4]

1. 北京邮电大学人工智能学院, 北京 100876; 2. 教育部信息网络工程研究中心, 北京 100876; 3. 交互技术与体验系统文化和旅游部重点实验室, 北京 100876; 4. 北京邮电大学计算机学院, 北京 100876; † 通信作者, E-mail: rfli@bupt.edu.cn

摘要　针对少样本文本分类任务提出提示学习增强的分类算法(EPL4FTC)。该算法首先将文本分类任务转换成基于自然语言推理的提示学习形式, 实现在利用预训练语言模型先验知识的基础上, 达到隐式数据增强, 并通过两种不同粒度的损失优化。并且, 为捕获下游任务中含有的类别信息, 采用三元组损失联合优化, 同时引入掩码语言模型任务作为正则项, 提升模型泛化能力。在公开的 4 个中文和 3 个英文文本分类数据集上进行了实验评估。实验结果表明, 提出的 EPL4FTC 方法的准确度性能明显优于所对比的基线方法。

关键词　预训练语言模型; 少样本学习; 文本分类; 提示学习; 三元组损失

Enhanced Prompt Learning for Few-shot Text Classification Method

LI Ruifan[1,2,3,†], WEI Zhiyu[1], FAN Yuantao[1], YE Shuqin[1], ZHANG Guangwei[2,4]

1. School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876; 2. Engineering Research Center of Information Networks, Ministry of Education, Beijing 100876; 3. Key Laboratory of Interactive Technology and Experience System, Ministry of Culture and Tourism, Beijing 100876; 4. School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876; † Corresponding author, E-mail: rfli@bupt.edu.cn

Abstract　An enhanced prompt learning method (EPL4FTC) for few-shot text classification task is proposed. This algorithm first converts the text classification task into the form of prompt learning based on natural language inference. Thus, the implicit data enhancement is achieved based on the prior knowledge of pre-training language models and the algorithm is optimized by two losses with different granularities. Moreover, to capture the category information of specific downstream tasks, the triple loss is used for joint optimization. The masked-language model is incorporated as a regularizer to improve the generalization ability. We evaluated our method on four Chinese and three English text classification datasets. The experimental results show that the classification accuracy of the proposed EPL4FTC is significantly better than the other compared baselines.

Key words　pretrained language model; few-shot learning; text classification; prompt learning; triplet loss

文本分类[1]作为自然语言处理领域的典型任务已经得到了广泛的应用, 例如情感分析、新闻推荐和用户画像等场景。这些场景通常可以获取海量的未标注数据, 因而需要大量的人工标注工作。但是, 诸如医疗和安全等一些特殊的工业应用场景通常较难获取大量的数据满足模型的训练。这使得基于数据驱动的深度学习方法在少量数据下较难取得令人满意的效果。为了使机器具有仅通过几个简单样本实现快速学习新事物的能力, 少样本学习[2]的概念被提出。少样本学习的核心目标是面对新的领域任务, 利用先验知识仅通过有限的训练样本快速且准确完成对任务的学习。

国家自然科学基金(62076032)资助
收稿日期: 2023–05–18; 修回日期: 2023–08–30

近年来随着预训练语言模型的发展, 尤其是以 BERT[3]为代表的通用预训练模型的提出, 使得基于预训练和微调的两阶段训练范式逐渐成为新的趋势, 并在大多数的自然语言处理任务上取得了优异的成绩。但是, 在微调阶段, 模型的性能通常取决于任务和有标注训练数据的规模。这使得当模型面对仅含有少量训练样本的下游任务时性能表现往往不佳。针对预训练模型下的少样本学习问题, 基于提示学习的方法提供了一种新颖有效的解决思路。基于提示学习的方法通过调整下游任务的形式与预训练任务形式保持一致, 实现充分发挥预训练模型中语言模型任务本身的优势。同时, 该方法通过减小上下游任务训练方式不一致带来的差异, 达到少样本学习的目的。虽然基于提示学习的方法取得了不错的效果但仍然面临着挑战, 具体包括两点: 1) 在少样本学习的场景中, 容易出现类别的数量远多于单一类别样本量的现象, 使得模型在此类任务上的表现通常较差; 2) 基于提示学习的方法大多依赖预训练语言模型中已经学习到的先验知识, 而较少关注下游具体任务的类别表征信息。

针对上述问题, 提出了一种增强提示学习的少样本文本分类算法(EPL4FTC)。该算法首先将下游任务转换成基于自然语言推理的提示学习形式, 通过任务形式的转换, 达到在有效利用预训练语言模型中已经学习到的先验知识的基础上实现数据的隐式增强, 并通过两种不同粒度的损失进行优化。此外, 为捕获下游具体任务中丰富的类别等表征信息, 该算法通过三元组损失[4]进行联合优化, 同时引入掩码语言模型任务(MLM)作为正则项, 预防过拟合或数据灾难性遗忘发生的风险, 进一步提升模型的泛化能力。

本文贡献主要包括三方面: 1) 提出了一种基于提示学习和三元组损失的少样本文本分类的 EPL4FTC 算法。该算法对仅含有少量实例的文本分类任务能够有效完成文本分类。2) 提出利用提示学习将任务转换成自然语言推理形式, 并且通过三元组度量学习方法实现捕获下游文本分类任务的类别表征, 提升文本分类的准确性。3) 完成在中英文多个数据集上的实验, 结果表明文本分类的准确率有性能的提升, 验证了算法的有效性。

1 相关工作

本文提出的方法主要与基于度量学习的方法和基于提示学习的方法密切相关。

1) 基于度量学习的方法。Koch 等人[5]提出孪生网络模型, 由两个结构相同且部分共享权重的网络构成, 并通过欧式方法计算输入样本对的匹配程度判断是否属于同一类别。Vinyals 等人[6]提出一种匹配网络模型, 该方法通过记忆网络和注意力机制, 实现对以往知识的记忆存储以及快速学习新样本的特征。Snell 等人[7]提出原型网络模型, 将不同类别的平均向量作为类别原型的向量表示, 最后在推理阶段通过计算样本到类别原型向量的距离实现对类别的预测。Sung 等人[8]提出关系网络模型, 该方法通过一个神经网络关系模块, 实现自动学习特征间的距离度量关系表示。Geng 等人[9]提出归纳网络模型, 在关系网络的基础上, 引入动态路由机制实现获取支撑实例的类别向量表示, 最后通过关系模块计算查询实例与支撑实例的关系得分进行分类。随后在 Geng 等人[10]又提出动态记忆归纳网络。通过引入二阶段训练范式, 在第一阶段进行有监督的训练, 实现为第二阶段的训练提供一个具有良好初始化编码器和记忆模块, 同时利用动态路由机制为少样本学习提供更强大的灵活性, 让模型更好、更好地适应训练数据。基于度量学习的方法, 通过采用传统度量方法或深度度量方法等实现对类别的表征表示。但是, 不同的度量方法在不同的具体任务上差异性较大, 无法适应多样的实际问题。其次, 它过于依赖训练数据, 当数据较少时不能很好地学习到类别的映射关系。

2) 基于提示学习的方法。Schick 等人[11–13]提出模式探索训练(PET)方法用于少样本学习。该方法通过定义并添加人工构建的模板, 将文本分类任务转换为完形填空任务。在训练过程中, PET 方法将分类标签转换为标签描述形式, 并使用[MASK]进行替换填入到人工定义的模板当中。通过语言模型还原[MASK]位置的词, 最后使用标签映射策略完成文本分类任务。Liu 等人[14]在 PET 基础上提出 ADAPET 模型, 将模板中需要模型预测的词从有限候选词变成整个词表, 通过扩大其搜索空间增加模型的泛化性能。以及通过正确标签反向预测原文中的字符, 进一步提升模型的性能。Gao 等人[15]提出 LM-BFF 模型。该模型首先通过 T5 模型[16]实现自动化的生成最优模板, 避免了繁杂的人工搜索模板这一过程。接下来将提示示例通过上下文的形式添加到原始输入中, 通过利用更丰富的文本信息完成语言模型的建模工作。Liu 等人[17]提出 P-tuning

模型。它丢弃了提示模板必须是自然语言的假说，采用让语言模型自动学习适合当前任务形式的最佳提示模板形式。在训练过程中，通过使用预训练模型词表中未使用的字符去学习模板的连续表示形式，并且 P-tuning 通过只学习更新模板对应的参数，从而极大减小了模型需要学习的参数量。Wang 等人[18]提出了 EFL 模型，与将文本分类任务转换为完形填空任务形式不同的是，EFL 方法是将文本分类任务转换为文本蕴含任务形式。在训练过程中，EFL 对于每一个原始输入，都会与正确的标签描述生成新的正例，以及与其余候选标签随机生成若干新的负例。通过上述数据构造方式，实现原始输入与正确的标签描述模板构成蕴含关系；否则为非蕴含关系。Jiang 等人[19]提出两种不同的模板集成方法。首先是概率平均的集成方法，通过训练集选择出若干个性能最好的提示模板，然后在推理阶段中使用候选若干个提示模板的概率平均作为最终预测结果。Hu 等人[20]提出一种知识型的提示学习调优方法，使用外部知识库扩展了标签词空间，提高了标签词的覆盖率，在零样本和少样本文本分类任务上证明了有知识调优的有效性。Min[21]等人提出一种用于语言模型提示的噪声通道方法，该方法证明了使用计算给定标签输入通道的噪声通道可以显著优于直接计算标签概率的方法。Zhang[22]等人提出了同时使用基学习器和元学习器的提示学习方法，证明了度量学习可以帮助提示学习的方法更快收敛。基于提示学习的方法在大规模无监督语料训练的预训练语言模型基础上发展起来，旨在减小预训练任务和下游任务形式之间的巨大差异，使下游任务形式尽可能与预训练任务形式保持一致。

2 EPL4FTC 算法

EPL4FTC 模型结构由基于自然语言推理的提示学习模块和度量优化模块两部分组成，且两个模块共享编码层的参数。其中，基于自然语言推理的提示学习模块通过掩码语言模型头层计算输入句子中的[MASK]位置处的概率，并通过单句级和句群级两种不同粒度的损失方法进行模型优化。度量优化模块随机对训练样本进行抽样，通过共享编码层编码后，使用三元组损失计算锚点与正负例之间的损失，最后对两个模块联合学习。

2.1 基于自然语言推理的提示学习模块

如图 1 所示，基于自然语言推理的提示学习模块负责将文本分类任务转换为基于自然语言推理形式的完型填空任务。具体地，对于原始输入文本将真实标签通过模板映射转化为自然语言推理形式。其中推理词使用预训练语言模型中[MASK]字符替代，通过建模上下文间的关系，推理出[MASK]位置上真实的推理词。

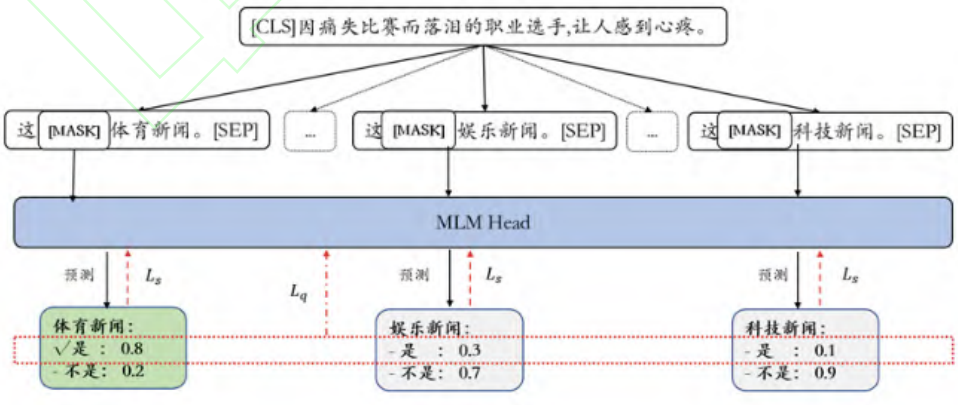

图 1　基于自然语言推理的提示学习结构图

Fig. 1　Structure diagram of prompt learning based on natural language reasoning

下面给出基于自然语言推理的提示学习方法的形式化表达。对于给定的输入文本x,所对应的真实标签为l，以及需要推理判断的标签描述d，通过函数f_{prompt},实现将输入x转换为基于提示学习新的输入形式x':

$$x' = f_{prompt}(x, z, d) \tag{1}$$

其中，z表示通过verbalize映射将真实标签与需要推理判断的标签描述的关系转为逻辑推理词，可表示为:

$$z = \text{verbalize}(l, d) \tag{2}$$

定义模板的一般形式为: $[x']$="$[x]$,$[z]$$[d]$。"。对于原始输入$x$，将其填充到$[x]$，需要推理判断的标签描述$d$填充到$[d]$中；接下来，将输入$x$的真实标签描述$l$，与当前填充需要推理判断的标签描述$d$，通过映射函数verbalize$(l, d)$获得当前输入的逻辑推理词$z$。其中，$[z]$将被预训练语言模型中[MASK]字符进行替代，以及逻辑推理关系词z将作为$[z]$的真实标签参与模型的优化。在推理阶段，对输入x和所有的标签描述d，通过映射函数f_{prompt}，转化为基于提示学习x'的形式。最后通过计算$[z]$处的自然语言推理词概率，选取预测为蕴含关系最大概率的标签描述d所对应的真实标签，作为最后的预测结果。

当采用自然语言形式的逻辑推理词时，使用自然语言中的"是"表示蕴含推理关系，"不是"表示非蕴含推理关系。进一步，为了让语言模型学到更通用自然语言推理的表示，对于推理词采用连续式的提示模板形式。也即，使用词表中未使用过的字符"[U1]"代表蕴含推理关系，"[U2]"代表非蕴含推理关系。

针对单样本输入形式和通过数据增强形式扩增负样本形成样例集合的形式，设计了两种不同粒度损失函数优化建模效果。

1) 单句级损失函数。如图 2 左所示，对于每一个通过f_{prompt}映射函数构成新的输入实例，需要模型完成建模上下文信息并预测推理出[MASK]位置处的真实推理词，并通过交叉熵进行优化。在给定输入x情况下，定义[MASK]处推理词z的概率分布如下所示:

$$q(l|x) = \frac{e^{s(z|x)}}{\sum_{z \in Z} e^{s(z|x)}} \tag{3}$$

其中，Z表示候选推理词集合，$s(z|x) = MLM(\text{z}|f_{prompt}(x))$表示在[MASK]位置处对候选推理词集合的语言模型得分。最后，通过交叉熵损失计算单句级损失，如下所示:

$$\mathcal{L}_s = CE(q(l|x), z) \tag{4}$$

2) 句群级损失函数。单句级损失函数仅考虑对实例本身进行优化，没有考虑同一组正负样本间的关系。所以定义句群级损失函数，实现对一组中的正负样本间关系进行优化，如图 2 右所示。具体地，在对输入的实例进行数据构造时，通过输入实例与所对应的类别生成一个正例。将输入实例与其他的类别进行数据构造生成 n-1 个负例。这最终为每一条输入样本获得 n 个实例样本。最后，采用交叉熵损失在句群级进行优化。

$$\mathcal{L}_q = CE(g(s(z|x)), I_{entail}) \tag{5}$$

其中，I_{entail}表示在当前样例组中真实标签为蕴含关系的位置索引，$g(s(z|x))$表示语言模型对[MASK]位置处的推理词在蕴含关系上的预测得分。最后，基于自然语言推理的提示学习模块的损失函数定义如下:

$$\mathcal{L}_p = (1 - \alpha) \cdot \mathcal{L}_s + \alpha \cdot \mathcal{L}_q \tag{6}$$

其中，α为可调节的超参数。

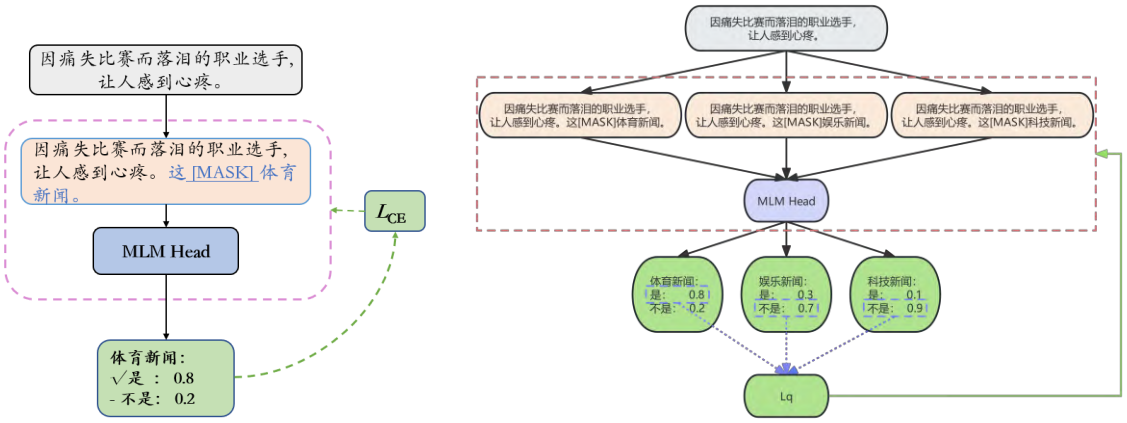

图 2　单句级与句群级的优化图

Fig. 2　Sentence-Individual Level and Sentence-Group level optimization

2.2 度量优化模块

提示学习利用预训练语言模型在预训练任务中学习到的先验知识，在具体下游任务中可以取得一个良好的性能。但对文本分类任务而言，类别特征的表示也至关重要。通过度量学习实现将原始语义空间的实例映射到目标任务上语义空间的表示，使实例在目标任务上的语义空间的表示具有更强的区分能力。

度量优化模块的目标是使在语义空间中属于同一类别的实例的距离更接近，而不同类别的实例的距离远离。通过三元组损失函数进行有监督的度量学习，使模型可以更好学习不同类别间的距离关系信息[23]。此外，使用带间隔的损失函数可以提升模型的泛化性能，如公式 7 所示。具体地，构造三元组数据时，在某个类别中选定一个实例作为锚点，同类别的实例作为正例，其他类别的实例作为负例。

$$\mathcal{L}_{tml} = \sum_{m=1}^{M} max(0, d(A_m, P_m) - d(A_m, N_m) + \Delta) \tag{7}$$

其中，$d(A_m, P_m)$表示锚点与正例间的距离，$d(A_m, N_m)$表示锚点与负例间的距离，Δ表示设定的间隔值。

此外，少样本学习场景中用于训练的数据量通常十分有限。为了缓解灾难性遗忘的问题，使用掩码语言模型优化目标作为正则项进行建模。所以度量优化模型的损失函数表示为:

$$\mathcal{L}_{aux} = (1 - \beta)\mathcal{L}_{tml} + \beta\mathcal{L}_{mlm} \tag{8}$$

其中，$\mathcal{L}_{mlm}$表示语言模型损失，β表示相应的权重参数。

最后，整体的损失函数由提示学习损失$\mathcal{L}_p$和度量优化损失$\mathcal{L}_{aux}$的加权构成，具体如下:

$$\mathcal{L}_{total} = (1 - \gamma)\mathcal{L}_p + \gamma\mathcal{L}_{aux} \tag{9}$$

其中，$\mathcal{L}_p$表示基于自然语言推理的提示学习模块的损失，$\mathcal{L}_{aux}$表示度量优化模块的损失，γ表示权重参数。

2.3 模型训练与推理

EPL4FTC 算法将文本分类任务转化成自然语言推理任务，即转化后的任务是一个二分类任务。因此，当一个原始分类任务包括 *N* 个类别时，该算法需要进行 *N* 次推理，最后选择推理概率最大所对应的标签类别作为最终预测结果。所以，1) 在模型训练过程中，为提升模型的泛化性能同时降低模型训练的成本，通过负采样的方式对下游任务进行训练。对于一个包含多个类别的分类任务，将每一个实例与之对应的类别作为正例，同时随机选择 *K* 个其他类别与当前实例构成负例。以上数据构造方式不但能够提升模型的性能，而且相比使用全部类别作为负例进一步缩短了训练模型所需的时间。2) 在模型推理阶段，EPL4FTC 算法仅使用基于自然语言推理的提示学习模块。具体地，对于包含 *N* 个标签的文本分类任务，对每一个实例生成包含自然语言推理提示模板的 *N* 条新的输入实例。通过模型预测出每一个实例中[MASK]位置处所蕴含推理词的概率，在 *N* 个预测结果中选择预测最大概率所对应的标签作为当前原始输入实例的预测结果。

3 实验与结果

3.1 实验数据集

1) 中文数据集。中文数据集使用少样本评测数据集 FewCLUE[24]中的文本分类任务对应的数据集。本文在 4 个不同领域的评测数据集上进行实验。其中，EPRSTMT 为电商评论情感分析任务，是典型的包含正向和负向情感的二分类任务。CSLDCP 是科学文献学科领域的长文本多分类任务，包含了 67 个类别。TNEWS 是新闻标题的短文本分类任务，包含了教育、娱乐和文化等 15 个类别。最后，IFLYTEK 是根据 APP 应用的长文本主题描述信息对超过 100 多个应用类别进行分类的任务。2) 英文数据集。本文采用 3 个英文文本分类数据集 AG News、TREC 以及 Yelp Review 进行评测。其中，AG News[25]是学术新闻搜索引擎从多个新闻来源中搜集超过了 100 万篇新闻文章构成的数据集。它包含 4 类新闻主题，分别是世界、体育、商业和科技。TREC[26]数据集包含 6 个一级标签和 47 个二级标签。Yelp Review 数据集来自 Yelp 的用户评论。它的标签是用户对商品的星级打分，共分为 5 级。用于评测的英文数据集将从以上数据集中抽样获得。将每一个原始英文数据集中随机抽取 8 个、16 个和 32 个实例形成对多个不同规模的数据集用于训练，测试集为默认。

3.2 基线方法

采用的基线方法包括: 1)基于微调方法[3]: 在预训练语言模型的基础上，通过为模型添加任务相关的分类器，达到使模型可以处理具体的下游任务的目的。2)Zero-shot 方法[27]: 基于 Roberta 等自编码预训练语言

模型, 通过 MLM 进行推理评测。3)Zero-shot(GPT)方法[17]: 基于 GPT 自回归预训练语言模型[28], 通过从左至右的语言模型进行推理评测。4)PET 方法[12]: 通过添加人工自定义模板, 将下游任务转化成完成填空形式的任务, 然后在候选标签列表中选择合适的标签。5)ADAPET 方法[29]: 对模板搜索正确答案时从有限候选词变成整个词表, 扩大了模型的搜索空间。此外, 对正确标签反向预测原文中的词, 实现模型性能的提升。6)LM-BFF 方法[15]: 通过自动化生成的离散化自然语言作为提示模板, 同时通过采样的形式将实例以上下文的方式添加到每一个输入中。7)P-tuningR 方法[17]: 区别于自然语言形式的提示模板, 采用 Roberta 作为预训练语言模型, 实现让模型自动学习到最佳的连续式的非自然语言提示模板。8)EFL 方法[25]: 通过添加人工自定义模板, 将下游任务转化成蕴含任务形式, 并添加额外的二分类器, 实现对下游任务的微调。

3.3 实现细节与评测指标

实验在配有 CUDA 环境的 Linux 操作系统下进行, 并配置了两块 GTX 1080Ti 显卡。代码使用基于 PyTorch[30]框架的 HuggingFace 工具包实现。对于中文数据集的评测, 采用 12 层网络结构的中文 RoBERTa-wwm-ext[31]预训练模型。对于英文数据集的评测, 采用 12 层结构的 BERT-BASE 预训练模型。模型参数设置如下: 学习率为10^{-5}, 超参α设置为 0.7, β为 0.01, γ为 0.02, 三元损失间隔Δ为 0.15, 并且使用 AdamW[32]优化器进行模型参数的优化。依据之前的研究, 在少样本学习问题中通常使用准确率(Accuracy)作为评测指标。它表示模型预测正确的样本数量占所有的样本数量的比例。

3.4 实验结果

1) 中文数据集的实验结果。实验结果如表 1 所示。可以看到, 对于基于微调的方法, 在小样本学习场景中模型性能通过表现不佳。而对于采用基于提示学习的方法, 通过使用 PET、LM-BFF、EFL、P-tuningR 方法、以及 EPL4FTC 算法, 在小样本学习场景中模型的准确率都有大幅提高, 显示出提示学习方法具有强大的潜能。通过对比 EPL4FTC 算法与其他基于提示学习的方法(PET、ADAPET、LM-BFF、EFL 和 P-tuning 等), 可以看出 EPL4FTC 算法在 EPRSTMT、CSLDCP 和 TNEWS 等数据集上取得了优异的成绩。此外, 在 IFLYTEK 数据集上也取得了与其他现有方法同等效果的性能。而且, EPL4FTC 算法在中文文本分类任务的平均准确率性能上取得了最高的成绩。与转换为完形填空任务形式的 PET 和 ADAPET 等方法相比, EPL4FTC 算法在利用预训练模型中已经学习到的通用知识基础上, 引入下游任务的类别信息实现更好的建模效果, 并在任务的平均准确率上高出 3.9%。与转化为文本蕴含任务的 EFL 方法相比, EPL4FTC 算法没有引入额外需要学习的大规模参数, 并且与预训练语言模型任务保持一致, 有效减小上下游任务间的差异性, 最终在任务的平均准确率上高出 4.2%。与使用自动构建模板或是非自然语言形式模板的 LM-BFF 和 P-tuning 方法相比, EPL4FTC 算法无需繁琐的模板构建形式, 并且在任务的平均准确率上高出 1.6%。

表 1　中文少样本数据集实验结果

Table 1　Experimental results of Chinese few-shot datasets

方法	EPRSTMT	CSLDCP	TNEWS	IFLYTEK	MEAN
Human	90.0	68.0	71.0	66.0	73.8
Fine-tuning	65.4	35.5	49.0	32.8	45.7
Zero-shot	85.2	12.6	25.3	27.7	37.7
Zero-shot (GPT)	57.5	26.2	37.0	19.0	35.0
PET	**85.6**	51.7	53.7	35.0	56.5
ADAPET	85.1	43.5	46.4	36.6	52.9
LM-BFF	84.6	54.4	53.0	46.1	59.5
EFL	85.0	45.0	52.1	42.7	56.2
P-tuningR	80.6	54.8	52.2	**48.0**	58.8
Our EPL4FTC	85.3	**55.1**	**54.6**	46.4	**60.4**

2) 英文数据集的实验结果。本文英文数据集的训练集中每一类别包含不同规模的实例数量(K=8, 16 和

32)。实验结果如表 2 所示。从结果可以看出, 对于不同的实例数量, 基于微调的方法、PET、ADAPET、EFL、P-tuning 以及 EPL4FTC 算法, 都表现出随着实例数量的不断增多, 模型的准确率都有着明显的提升。这表明在基于深度模型的少样本学习场景中, 训练数据的规模对模型性能有着较大影响。其次, 在实例数 K=8 时, 虽然 PET、ADAPET、EFL 和 P-tuning 等基于提示学习的方法比基于微调的方法模型的准确率有很大提升, 但 EPL4FTC 算法却表现出更加出众的性能, 其模型准确率远高于其他方法。这表明在给定较少实例的情况下, EPL4FTC 算法能够有效地对下游任务进行建模, 也进一步说明了该算法的有效性。进一步, 随着实例数的增加(K=16, 32), 虽然其他基于提示学习方法的性能也有所提升, 但 EPL4FTC 算法的模型准确率相比于其他方法依然保持较高水平。即使在实例数 K=32 的情况下, EPL4FTC 算法也与现有模型在性能上保持在同一水平, 并在任务的平均准确率性能上保持最佳。

表 2　英文少样本数据集实验结果

Table 2　Experimental results of English few-shot dataset

方法	AG News	TREC	Yelp Review	MEAN
Few-shot (K=8)				
Fine-tuning	52.5	29.2	18.7	33.5
PET	76.0	38.8	25.0	46.6
ADAPET	78.8	21.6	24.2	41.5
EFL	78.3	41.6	21.2	47.0
P-tuningR	68.3	31.8	22.2	40.8
Our EPL4FTC	**79.5**	**55.8**	**26.1**	**53.8**
Few-shot (K=16)				
Fine-tuning	66.4	46.2	26.5	46.4
PET	83.3	62.4	30.3	58.7
ADAPET	83.6	62.7	31.7	59.3
EFL	84.6	55.4	30.0	56.7
P-tuningR	83.2	66.4	30.0	59.9
Our EPL4FTC	**84.9**	**68.6**	**32.3**	**61.9**
Few-shot (K=32)				
Fine-tuning	80.8	66.8	32.9	60.2
PET	84.7	80.3	39.0	68.0
ADAPET	85.8	80.1	31.5	65.8
EFL	86.0	81.0	35.9	67.6
P-tuningR	86.1	**81.8**	39.2	68.4
Our EPL4FTC	**86.2**	80.8	**40.0**	**69.0**

3.5 组件有效性分析

1) 度量优化模块有效性分析。

首先对比基于度量学习不同的损失优化方法。在基于度量学习的不同的损失优化方法对比实验中, 对比了欧式距离和余弦相似度作为度量方法的二元交叉熵损失对比损失以及三元组损失的优化方法。具体地, 在使用二元交叉熵损失作为损失优化的实验中, 采用欧式距离作为度量方法。由于其度量值域范围是$[0,+\infty)$, 为便于计算二元交叉熵损失, 将其映射到值域空间$[0,1)$的范围, 如下所示:

$$f(r) = tanh(\frac{1}{r+\epsilon}) \tag{10}$$

其中，ϵ超参数为避免分母为 0。其次，在使用基于余弦相似度的度量方法中，其值域范围是$[-1,+1]$。同理，将其映射到值域空间$[0,1]$的范围，具体如公式 11 所示。

$$f(r) = \frac{1}{2}(r+1) \tag{11}$$

实验结果如表 3 所示。从实验结果看出，使用欧式距离或余弦相似度作为度量方法的二元交叉熵损失优化方法的性能较差，而对比损失和三元组损失优化方法性能上有较大提升。这得益于这两个损失优化方法中引入间隔的策略，使模型有了一定的容错空间，进而提升了模型泛化性能。进一步，三元组损失通过同时对比三组不同的实例，使得它可以同时获取更多的信息帮助模型优化，从而提升模型性能。与对比损失相比，三元组损失在不同任务的平均准确率性能有 1.4%的提升。

接下来，本文对采用三元组损失的度量优化模块进行了消融实验。实验结果如表 4 所示。从实验结果可以看出，当移除完整的度量优化模块后模型的准确率有明显的性能下降，在中文数据集中平均下降了 1.6%而在英文数据集中平均下降了 3.2%。这验证了度量优化模块的有效性。该模块通过学习下游任务中的类别信息，实现了对模型性能的提升。进一步，在度量优化模块中将 MLM 损失作为三元组损失的正则项引入。为了验证 MLM 正则项的有效性，本文仅保留了三元组损失并移除了 MLM 损失正则项。从实验结果可以看出，当移除 MLM 正则项后，模型的准确率性能在大部分任务上都有明显的下降，在中文 CSLDCP 任务下降了 2.1%，而在英文 TREC 任务上下降 5.8%。这也说明了引入 MLM 损失作为正则项对模型的性能提升的有效性。

表 3　中文(左)英文(右)数据集上不同损失优化实验结果

Table 3　Experimental Results of different loss optimization methods on Chinese (Left) and English (Right) datasets

方法	EPRSTMT	CSLDCP	TNEWS	IFLYTEK	MEAN
BCE Loss (CS)	77.6	54.3	54.1	45.2	57.8
BCE Loss (ED)	80.6	**55.5**	52.8	45.1	58.5
Contrastive Loss	83.2	54.5	53.9	44.4	59.0
Triplet Loss	**85.3**	55.1	**54.6**	**46.4**	**60.4**

方法	AG News	TREC	Yelp Review	MEAN
BCE Loss (CS)	72.2	53.4	25.3	50.3
BCE Loss (ED)	72.1	51.6	25.1	49.6
Contrastive Loss	72.9	54.0	**26.6**	51.2
Triplet Loss	**79.5**	**55.8**	26.1	**53.8**

表 4　中文(左)英文(右)数据集上度量优化模块消融实验结果

Table 4　Ablation results of metric optimization module on Chinese (Left) and English (Right) datasets

方法	EPRSTMT	CSLDCP	TNEWS	IFLYTEK	MEAN
Our(-Triplet Loss -MLM)	83.4	54.6	54.0	43.3	58.8
Our (-MLM)	85.2	53.0	54.2	45.7	59.5
Our Loss	85.3	55.1	54.6	46.4	60.4

方法	AG News	TREC	Yelp Review	MEAN
Our(-Triplet Loss -MLM)	75.5	54.4	21.7	50.6
Our (-MLM)	78.8	50.0	26.3	51.7
Our Loss	79.5	55.8	26.1	53.8

2) 句群级损失有效性分析。基于自然语言推理的提示学习模块中，通过句群级损失实现对一组正负实

例间的优化。为了确定该损失优化方法的有效性，对其进行消融实验。实验结果如表 5 所示。在中文数据集上的实验结果可以发现，句群级损失优化方法对不同的任务都有明显的性能提升，特别是对于 IFLYTEK 任务有 3%的性能提升。对英文数据集在实例数 K=8 上进行实验。结果相比于中文数据集，英文数据集中句群级损失对模型的准确率具有更大的提升作用。在 AG News 数据集上显著提升了 38.6%；在 Yelp Review 数据集中也有 5.9%的提升。该实验表明针对组内优化句群级损失方法的有效性。它通过对比组内正负间的实例，可以学习到更好的知识表示。

表 5　中文(左)英文(右)数据集上句群级损失有效性分析

Table 5　Validity Analysis of Sentence-Group Level Loss in Chinese (Left) and English (Right) Datasets

方法	EPRSTMT	CSLDCP	TNEWS	IFLYTEK
EPL4FTC	85.3	55.1	54.6	46.4
w/o L_{group}	84.9	54.2	54.1	43.4

方法	AG News	TREC	Yelp Review
EPL4FTC	79.5	55.8	26.1
w/o L_{group}	40.9	39.6	20.2

3.6 提示模板分析

1) 推理词形式性能分析。EPL4FTC 算法将文本分类任务转换为基于自然语言推理形式的完型填空任务，同时受到 P-tuning 方法的启发，推理词不仅可以是自然语言形式，也可以是非自然语言形式。为此，本文将对这两种形式的推理词进行性能评估。实验结果如表 6 所示。从中文、英文数据集的实验结果可以看出，非自然语言形式的推理词较为稳定，模型的性能较好。具体地，对于形式简单、数据区分度高的任务，如 EPRSTMT 和 TREC 等任务，自然语言形式的推理词表现较为出众。而对于类别数较多、复杂的任务，如 TNEWS、IFLYTEK 和 CSLDCP 等任务，非自然语言形式的推理词具备更好的性能。这是由于它可以从具体任务中自主学习到更适合当前模板的推理词形式，而不受自然语言形式的限制。也就是说，非自然语言形式的推理词可以从众多的上下文信息中学习到推理词的连续化表达形式，从而有效避免了单一推理词的影响。

表 6　中文(左)英文(右)数据集推理词形式性能比较

Table 6 Performance Comparison of Inference Word Form in Chinese and English Datasets

方法	EPRSTMT	CSLDCP	TNEWS	IFLYTEK	MEAN
自然语言推理词	86.0	54.3	53.5	45.4	59.8
非自然语言推理词	85.3	55.1	54.6	46.4	60.4

方法	AG News	TREC	Yelp Review	MEAN
自然语言推理词	77.3	59.0	26.1	54.1
非自然语言推理词	79.5	55.8	26.1	53.8

2) 提示模板性能分析。手工设计的提示模板会对模型的效果产生一定的波动，本文将评估手工设计的模板对模型性能产生的影响。实验结果如表 7 所示。从结果可以看出模型的性能会受到提示模板较大的影响。具体地，在中文 TNEWS 和英文 TREC 任务上对模板采用了前缀式与后缀式的形式进行评测。相比之下，在中文数据集上模型的性能差异性相对较小，最大与最小的差值为 1.1%。而在英文数据上，模型的性能则表现出较大的差异性，最大与最小的差值为 6.4%。这表明提示模板对模型的准确率影响与具体的下游任务有较大的关系。通过优化模板的形式可以较大程度提升模型的性能。

表 7　准确率受提示模板的影响

Table 7　Accuracy with different prompt templates

模板	准确率	模板	准确率
TNEWS Classification		TREC Question Classification (K=8)	
下面<MASK><desc>新闻: <text>。	54.2	This <MASK> the <desc> question: <text>.	50.0
<text>, <desc>新闻？<MASK>。	54.6	<text>, it <MASK> about <desc> question.	56.4
<text>，这<MASK><desc>新闻。	53.9	<text>. it <MASK> <desc> question.	55.8

3.7 可视化分析

为评估在引入度量优化模块后所提出的模型获得任务类别信息的有效性，将通过 t-SNE[33]方法对中文 TNEWS 数据集通过随机采样进行可视化分析。首先，为了验证模型的编码层是否有效学习到任务中的类别信息，将预训练模型中编码层 CLS 位的输出作为当前整个实例的向量化表示，如图 3 左展示了编码后的分布情况。可以看出实例类别的分布依然保持着与 pooler 层相似的分布情况。对于简单的新闻类别，诸如股票、娱乐、电竞和汽车等依旧保持着较为紧凑的聚集现象。而对于文化和故事等较为抽象或是涵盖范围较广的新闻类别，虽分布别较为分散但也有着一定的区域性。这表明 CLS 作为整个句子的编码表示已经学习到了一定的实例类别信息。通过对实例类别的分布可视化分析，表明度量优化模块可以为整个模型提供更多额外的类别知识等信息。

在度量优化模块中采用三元组损失优化类别间的距离。具体地，本文将预训练模型中 pooler 层的输出通过度量优化模块进行度量学习。图 3 右展示了实例经该模块编码输出后的向量分布情况。可以看出同一类别的实例间都较为紧凑，同时不同类别间也存在着较为明显的间隔距离。这说明模型至少在 pooler 层中已经学习到了非常好的类别表示。

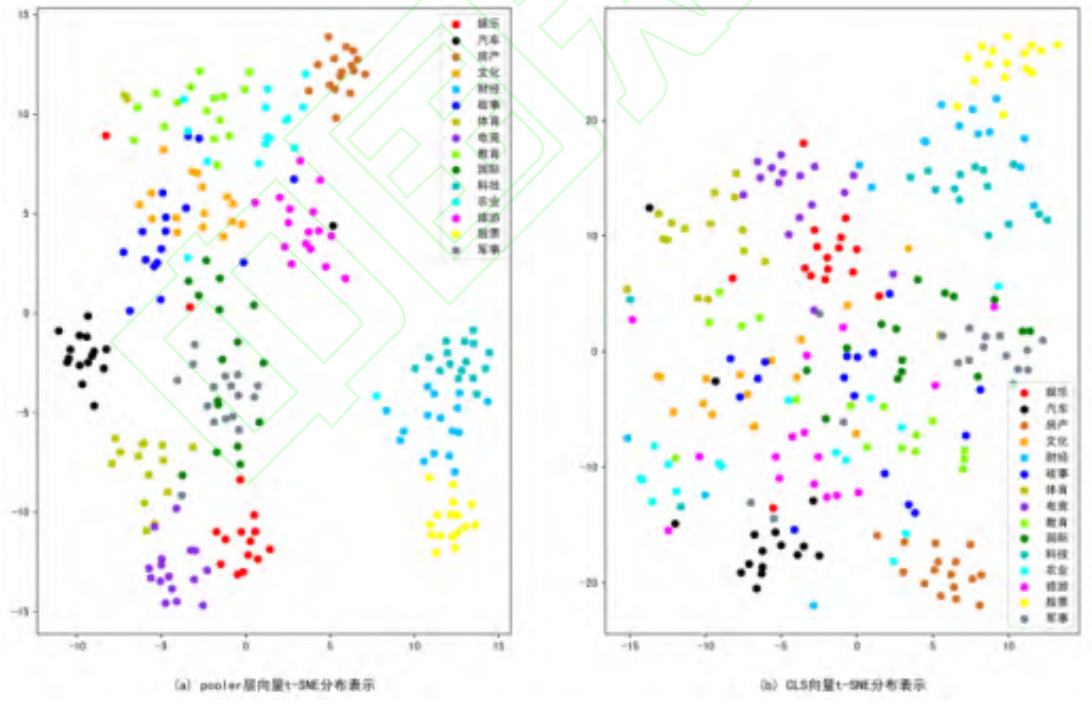

图 3　实例向量 t-SNE 分布可视化

Fig. 3　visualization on instance vectorization with t-SNE distribution

4 结论

提出了一种利用三元组损失增强提示学习的少样本文本分类算法(EPL4FTC)。该算法通过提示学习激活预训练语言模型中已学习到的通用知识，并且通过句子和句群两种粒度的三元组损失优化，实现捕获下

游具体任务的类别信息。同时，引入掩码语言模型任务的训练目标最为正则项，提升模型的泛化性能。论文在公开的多个中文和英文数据集上进行实验评估，同时执行了大量的剖析实验。结果表明了该算法的有效性。未来工作将尝试将该方法应用于其他主题的少样本任务场景。再者，对于除中英文之外语种的少样本文本分类研究也是一个有趣的问题。

参考文献

[1] Minaee S, Kalchbrenner N, Cambria E, et al. Deep learning--based text classification: a comprehensive review. ACM Computing Surveys, 2021, 54(3): 1-40

[2] Wang Y, Yao Q, Kwok J T, et al. Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys, 53(3): 1-34, 2020

[3] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C] NAACL, pp. 4171-4186, 2019

[4] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]//CVPR, 2015: 815-823

[5] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition[C]//ICML deep learning workshop. Cambridge, MA: JMLR. 2015

[6] Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning[C]. NIPS, pp. 3637–3645, 2016

[7] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. NIPS, 2017, 4080–4090

[8] Sung F, Yang Y, Zhang L, et al. Learning to compare: Relation network for few-shot learning[C]//CVPR, 2018: 1199-1208

[9] Geng R, Li B, Li Y, et al. Induction Networks for Few-Shot Text Classification[C]// EMNLP-IJCNLP, 2020: 3904-3913

[10] Geng R, Li B, Li Y, et al. Dynamic Memory Induction Networks for Few-Shot Text Classification[C]// ACL, 2020: 1087-1094

[11] Schick T, Schütze H. Few-shot text generation with pattern-exploiting training, EMNLP, pp. 390-402, 2022

[12] Schick T, Schütze H. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference[C]// EACL, 2021: 255-269

[13] Schick T, Schütze H. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners[C]//NAACL, 2339-2352, 2021

[14] Liu H, Tam D, Muqeeth M, et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning[C]. NeuIPS, 35: 1950-1965, 2022

[15] Gao T, Fisch A, Chen D. Making Pre-trained Language Models Better Few-shot Learners[C]//ACL. 2021: 3816-3830

[16] Raffel C, Shazeer N, Roberts A, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 2020, 21: 1-67

[17] Liu X, Zheng Y, Du Z, et al. GPT understands, too. arXiv preprint arXiv:2103.10385, 2021

[18] Wang S, Fang H, Khabsa M, et al. Entailment as few-shot learner. arXiv preprint arXiv:2104.14690, 2021

[19] Jiang Z, Xu F F, Araki J, et al. How can we know what language models know? TACL, 2020, 8: 423-438

[20] Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification. [C]// ACL, 2225–2240, 2022

[21] Min, S., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. Noisy Channel Language Model Prompting for Few-Shot Text Classification. ACL. 5316–5330, 2022

[22] Zhang H, Zhang X, Huang H, et al. Prompt-Based Meta-Learning for Few-shot Text Classification[C]//EMNLP. 2022: 1342-1357

[23] Weinberger K Q, Saul L K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. Journal of Machine Learning Research, 2009, 10(1): 207-244

[24] Xu L, Lu X, Yuan C, et al. Fewclue: A Chinese few-shot learning evaluation benchmark. arXiv preprint: 2107.07498, 2022

[25] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification[C]. NIPS, 2015, 28

[26] Li X, Roth D. Learning question classifiers [C]//COLING 2002, 556-562

[27] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized Bert pretraining approach. arXiv preprint:1907.11692, 2019

[28] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. https://openai.com/research/language-unsupervised, 2018

[29] Tam, D., Menon, R.R., Bansal, M., Srivastava, S., & Raffel, C. Improving and Simplifying Pattern Exploiting Training. EMNLP, pp. 4980–4991, 2021

[30] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library[C]. NeuIPS, 2019, 32

[31] Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for Chinese Bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514

[32] Loshchilov I, Hutter F. Decoupled weight decay regularization[C]. ICLR, https://openreview.net/forum?id=Bkg6RiCqY7 , 2019

[33] Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, 9(11)