基于动态聚类与标签空间映射的上下文学习模板构建方法

张琦　金醒男　裴誉　杜永萍
北京工业大学，北京
{ZQi74091,peiyu}@emails.bjut.edu.cn, jinxingnan@outlook.com, ypdu@bjut.edu.cn

摘要

面向大语言模型提供自然语言指令，可生成预期输出，体现了其上下文学习能力。上下文学习的性能与上下文模板质量密切相关，现有的工作通常使用单一的选择算法进行模板构建，无法充分激发上下文学习能力。本文提出基于动态聚类与标签空间映射的上下文学习模板构建方法，动态选择相关示例，进一步提出聚类筛选方法，实现不同语义簇中示例多样化的选择。设计基于损失函数的排序选择方法，评估模板学习正确标签空间映射分布的能力，排序形成最终模板。在自然语言推理等任务中的实验结果表明，本文提出的方法使两个不同的大语言模型准确率最高分别提升3.2%和8.9%。

关键词： 大语言模型；上下文学习；模板构建；动态聚类

In-Context Learning Demonstration Construction Method based on Dynamic Clustering and Label Space Mapping

Qi Zhang　Xingnan Jin　Yu Pei　Yongping Du
Beijing University of Technology，Beijing, China
{ZQi74091,peiyu}@emails.bjut.edu.cn, jinxingnan@outlook.com, ypdu@bjut.edu.cn

Abstract

Providing natural language instructions to large language models to generate the expected output reflects the In-Context Learning capability. The performance of In-Context Learning is closely related to the quality of demonstration. Existing methods typically use a single selection algorithm for demonstration construction, which fail to fully exploit the In-Context Learning capability. In this paper, we propose an In-Context Learning demonstration construction method based on dynamic clustering and label space mapping, which dynamically selects relevant samples. Then we propose a clustering filtering method to achieve diversified selection of samples within different semantic clusters. Furthermore, we design a ranking selection approach based on loss functions to assess the demonstration's ability of learning correct mapping distributions in label spaces, and form the final demonstration by sorting samples. Experimental results of different tasks such as natural language inference show that our method improves the accuracy of two large language models by up to 3.2% and 8.9%, respectively.

Keywords: Large Language Model , In-Context Learning , Demonstration Construction , Dynamic Clustering

基金项目：国家重点研发计划(2023YFB3308004），国家自热科学基金（92267107）

1 引言

大语言模型（Large Language Model, LLM）具有庞大的参数规模，在多种任务中展现出强大性能。当参数规模达到一定程度时，LLM表现出了预训练语言模型中没有表现出来的能力，其中，上下文学习（In-Context Learning, ICL）是具有代表性的能力之一(Zhao et al., 2023)。上下文学习是自然语言处理中一种新的范式，LLM只需根据由几个示例组成的上下文模板进行预测，而无需进行任何参数的更新，具有良好的可解释性和可行性(Brown et al., 2020)，因此通过研究上下文学习来提升LLM的能力已经成为新的热点。

现有对上下文学习的研究工作，专注于使用不同的评价指标作为度量，从训练集中选择若干示例，或通过不同的检索方法在大量数据中搜索示例，来构建上下文模板(Dong et al., 2022)。然而，不同测试样例与模板的相关程度不同，使用同一种度量方式构建模板不一定适用于所有样例。此外，上下文模板不仅对示例的要求较高，同时也受示例之间顺序关系的影响，如果不同时考虑这些因素，则会导致上下文学习难以达到理想效果。本文提出的方法相比于已有工作KATE(Liu et al., 2022)，能够根据不同的输入自动调整示例选择，有利于选取优质示例，同时比Levy et al.(2023)的工作使用更少的示例达到理想效果。此外，本文工作能够充分利用训练集中的标签空间映射信息，以此作为标准来判断不同排序方式蕴含正确信息分布的能力，不需要高质量的检索集，相比于Wu et al.(2023)的工作具有更好的泛化性。

本文工作主要贡献如下：

（1）本文提出了动态选择方法，通过对传统Top-K算法进行改进，动态选取与不同测试样例相关的示例，有效剔除无关示例，增加模板与问题的相关性；

（2）本文提出了聚类筛选的方法，采用聚类算法对动态选择后的示例进行划分，并在每个划分后的类中选择示例组成模板，在保证上下文模板示例相关性的同时，增加示例的多样性；

（3）本文针对筛选出的上下文示例，采用交叉熵作为损失函数，选择包含正确标签空间映射的示例排序方式进行排序，提升模板对测试样例标签的预测准确率。

2 相关工作

2.1 大语言模型与上下文学习

大型语言模型通常包含数千亿参数，是基于Transformer的语言模型，如GPT-3(Brown et al., 2020)、PaLM(Chowdhery et al., 2023)、LLaMA(Touvron et al., 2023)。由于模型规模庞大，计算资源有限，Hoffmann et al.(2022)使用缩放技术进行更加高效的计算资源分配；Rasley et al.(2020)提出了Deepspeed框架，支持分布式训练算法来学习LLM的网络参数。除此之外，提示学习与语言模型的结合，也能够实现在较低资源条件下完成自然语言处理任务。言佳润、鲜于波(2023)提出了基于微调与提示学习的大模型算法，在小样本或零样本数据集中，将GPT与提示学习相结合，较好地完成辩论关系识别任务。穆建媛et al.(2023)基于提示学习，进行中文短文本分类，使模型在少样本情况下取得了性能提升。

以上方法需对模型进行微调，当模型规模变大时，仍需要较多计算资源。上下文学习是LLM出现的一种新兴能力，它只需向大语言模型提供少量示例，即可让模型从中学习到上下文相关知识，并在不更新参数的情况下进行正确推理，目前，上下文学习已被广泛应用于传统自然语言处理任务(Kim et al., 2022)，如机器翻译(Zhu et al., 2023)、信息抽取(Wan et al., 2023)，以及数学推理问题(Wei et al., 2022)。Sun et al.(2022)表明，ICL凭借无需参数更新的学习框架大大降低了对计算资源的要求。此外，已有研究工作表明(Liu et al., 2022)，由于ICL的上下文模板采用自然语言形式，具有较强的可解释性，并且仅调整上下文模板即可将上下文知识与LLM融合。

为了进一步挖掘LLM的上下文学习能力，研究人员进行了大量的工作。Min et al.(2022)提出通过MetaICL来减小预训练和下游任务之间的差距，提高LLM的少样本学习能力。Wei et al.(2023)提出了符号微调技术，利用“输入-标签对”微调LLM，以此促使模型充分挖掘并利用上下文信息中蕴含的语义信息。然而，ICL的性能过度依赖上下文模板的质量。Zhao et al.(2021)研究表明，上下文示例的选择、排列顺序等因素在很大程度上影响着ICL的表现，这激励着研究者进一步探索如何设计一组高质量的上下文学习模板。此外，由于LLM本质上是基于大规模文本语料库训练的文本生成器，因此在分类任务上表现不佳。本文面向NLP领域的七个不同的分类任务数据集，提出基于动态聚类与标签空间映射的上下文学习模板构建方法，增强了LLM的上下文学习能力，提升LLM在分类任务中的表现。

2.2 上下文学习模板设计

上下文学习的示例模板为LLM提供丰富的语义信息，对上下文学习起着至关重要的作用。通常ICL即是从训练集中选取少量示例，将其改写为自然语言形式，形成上下文示例。随后，将待预测的测试样例与上下文示例连接，形成上下文模板，并将整个模板作为LLM的输入(Dong et al., 2022)。LLM从模板中学习数据与标签之间的映射关系，以及其中蕴含的语义信息，对测试数据进行正确预测。

在给定训练集的情况下，模板设计的重点包括示例子集的选择，以及如何对它们进行合理排序。互信息被用作选择度量(Sorensen et al., 2022), 其优点是不需要标记的样本和特定的最小二乘模型。受到使用检索模块增强神经网络的启发，KATE策略(Liu et al., 2022)被提出，选择与测试样例的语义信息相似的示例构建上下文学习模板，取得了较好的性能。此外，Gonen et al.(2023)发现，提示模板的质量与模型对其所涵盖语言的熟悉程度相关联，并发现模型对模板的困惑度越低，执行任务的能力就越强。Levy et al.(2023)考虑到示例的多样性，提出了一种选择不同上下文示例的方法，旨在覆盖输出内容中的所有结构，鼓励模型从这些结构中学习。在基于检索的方法中，Rubin et al.(2022)提出了一个二阶段的检索方法来选择示例，使用标记数据训练一个检索器，用于检索示例并构建模板。为了解决先前工作难以在各种不同任务之间迁移的问题，Li et al.(2023)提出多任务统一检索器，充分融合各类任务相关知识，使模型取得了优秀的性能表现。本文工作优化传统Top-k算法，根据不同输入样例动态选择上下文示例，增加示例与测试样例之间的相关性，并使用聚类策略提升示例的多样性。

除示例的选择外，对选定的示例进行合理排序也是模板设计的重点任务之一。研究表明，不同的语言模型均存在对示例排序敏感这一问题(Lu et al., 2022)。现有工作通常通过计算示例与测试样例之间的负欧几里得距离或余弦相似度，并以此进行排序(Liu et al., 2022)。熵值同样被用作度量指标，可分为全局熵和局部熵度量，并证实熵度量和ICL性能之间存在正相关关系(Lu et al., 2022)。Wu et al.(2023)受Solomonoff(1964)的一般推理理论和信息论的最小描述长度原则的启发，设计了一种基于最小描述长度的排序算法。本文基于输入数据与标签之间的映射关系，设计损失函数进行度量，为LLM提供充分的标签映射信息，提升预测准确率。

3 方法

上下文学习模板设计旨在通过对训练集中标注数据进行筛选、排序等操作，将其与测试样例相连接构建示例模板。具体可描述为：假设有上下文标注示例$\mathcal{D}$,$\mathcal{D} = [D_1, D_2, \ldots\ldots, D_i, \ldots\ldots D_j]$，其中$D_i = T(x_i, y_i)$，T代表将数据转换为自然语言的操作，$x_i$为第i个样本的输入，$y_i$是$x_i$的标签。对于给定测试样本$x'$，使用LLM预测对应标签$y'$可表示为：

$$y' = argmax_{y \in Y} \frac{P(y|\mathcal{D} \oplus x')}{\sum_{y \in Y} P(y|\mathcal{D} \oplus x')} \tag{1}$$

基于动态聚类与标签空间映射的上下文学习模板构建方法整体结构如图1所示。首先采用动态搜索方法，对测试样例进行实例级的上下文示例的选择，进一步设计聚类算法，采用两种不同的样本抽取方式择优筛选，确定最终的上下文示例集，最后，根据上下文模板中蕴含的标签空间映射关系分布进行示例排序，增强模板的有效性。

3.1 动态选择

传统Top-K算法可能会导致选取的示例与测试样例的语义关系不匹配，或遗漏掉语义关系相近的优质示例。有研究表明，可通过设置参数对距离进行筛选，减少预测误差(张少辉et al., 2022)。基于该发现，本方法改进传统Top-K算法，根据测试样例与其他示例之间的欧式距离来动态选取候选示例。欧式距离的计算如公式2所示：

$$L_2(x_i, x_j) = \sqrt{\left(\sum_{j=1}^{n} |x_i^{(1)} - x_j^{(2)}|^2\right)} \tag{2}$$

本方法首先使用all-mpnet-base-v2模型中的编码器，将数据中的句子和段落映射到768维的密集向量空间，形成大小为(64,768)的多维矩阵，并创建Faiss索引，使用内积计算，得到同样

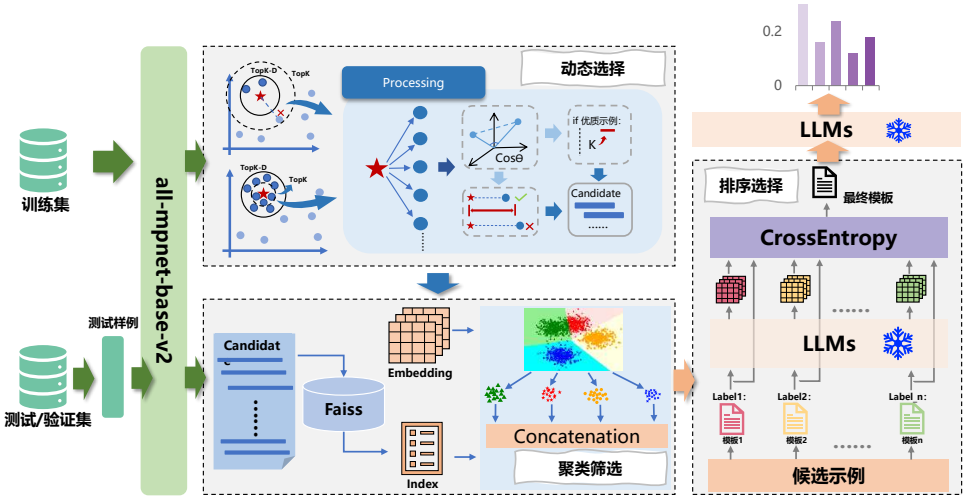

Figure 1: 基于动态聚类与标签空间映射的上下文模板构建方法

为768维的索引，将嵌入向量添加到Faiss索引中，完成样本的编码操作。随后获取测试样例x_i和其他样本x_j之间的欧氏距离，并由小到大进行排序。再根据K值筛选上下文示例，根据示例与测试样例之间的距离设置距离阈值，作为判断示例质量的依据。为了提升算法的准确度，将阈值分为优质阈值和无关阈值。当该组示例中有超过半数的数据与测试样例之间的距离小于无关阈值，则认为测试数据周围的相似样本稀疏，而当示例中距离最远的数据仍然大于优质阈值，则认为测试数据周围的相似样本稠密。

当测试样本周围的相似示例稠密，则动态扩大K值的范围；当测试样本周围的相似示例稀疏，则剔除与测试样本之间距离大于无关阈值的样本。为了防止剔除无关样本后候选样本数量过少，本方法随时动态监测上下文示例的数量，并设置上下文模板示例数量的上限和下限参数。根据欧氏距离选取数量等于上限参数的相关样本，从而保证候选示例与测试样本的相关性。

3.2 聚类筛选

动态选择模块对上下文示例进行了初步筛选，保证了模板中示例之间的相关性。然而，若直接对动态选择模块输出的候选示例进行抽取并排列，搜索空间为指数级，这将消耗大量的时间和计算资源。此外，初步筛选后的示例集中蕴含丰富多样的上下文语义信息，若从中随机抽取少量示例作为模板，存在知识信息冗余、提示角度片面等问题。例如：在Trec主题分类数据集中，当输入数据为“What class of animals makes up more than two-thirds of known species?”时，一个良好的上下文模板应该从“animals”、“species”、“ratio”等角度考虑问题。为了最大限度地挖掘初筛后示例集的信息，该模块使用KMeans聚类算法，计算动态选择后的候选示例在映射空间中的距离关系并进行分簇。经过聚类筛选后的上下文示例被分为若干簇，每个簇内部的示例具有相似的语义信息，而不同簇中示例的语义信息差别较大。从每个簇中进行示例抽取构建模板，充分地利用上下文示例中不同簇的语义信息，促使大语言模型从多角度进行综合决策。在对每个簇进行示例抽样时，为了同时考虑上下文示例的相关性和多样性，采用随机抽取和顺序抽取两种不同方法，如图2所示。

随机抽取：从每个上下文示例簇中随机抽取少量示例组成上下文模板的候选示例集，该方法保证了示例多样性。

顺序抽取：基于每个簇中示例与测试样本之间在嵌入空间的距离，依次从每个簇中顺序抽取相关度高的示例组成候选示例集，保证示例的高相关性。

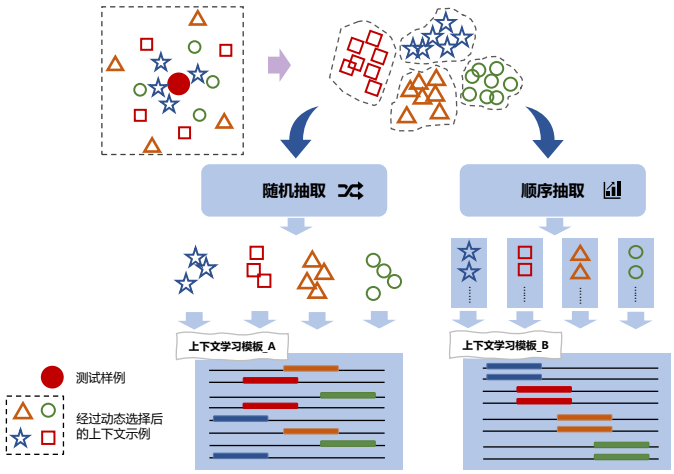

Figure 2: 聚类筛选过程

3.3 排序选择

动态选择和聚类筛选过程从上下文示例的相关性与多样性的角度筛选候选示例，而没有考虑示例的排列顺序对模板质量的影响。然而，由于LLM对于上下文示例的顺序较为敏感，所以对上下文示例进行合理排序是至关重要的。受Min et al.(2022)和高怡et al.(2023)的启发，ICL的性能收益主要来自于独立规范的输入空间和标签空间，以及正确一致的"输入-标签对"的格式，且模型可以通过模板来引入标签信息，因此我们考虑输入空间与标签空间之间的映射关系，为上下文示例选择合理的排序。

首先需要确定待筛选的不同排序方式。对于动态选择和聚类筛选操作后的N个候选示例，具有N!种不同的排列方式。为了节省计算资源，本方法进一步减少示例数目，从N个候选示例中选择n个示例，并进行排列。考虑到时间和计算资源的复杂度，本方法从所有可能的排列空间中随机选取了m种排列进行比较，如公式3所示：

$$R = Random_m\left(\frac{N!}{(N-n)!}\right), n \leq N \tag{3}$$

其中R代表包括m种不同排序的集合，每种排序都包含n个示例，$Random_m$的含义是从n!种排列方式中选取m个。

由于在预测前无法获取测试样例的真实标签，本方法获取其所有可能的标签空间信息，将输入数据的每一个可能的标签记为c。每一种可能的标签c对应多种不同的排序方式r，对于使用排序方式r进行排序的上下文示例文本序列$T_{c,r}$，使用分词器中的编码器对模板进行编码，编码后获得包含正确映射关系的信息分布，记作$emb_{T_{c,r}}$，将$T_{c,r}$与输入样例x_t拼接后，作为模板提示LLM输出预测结果。本文获取预测结果的嵌入空间信息分布，记作$emb_{T_{c,r},x_t}$。由于$T_{c,r}$由带标签的训练数据组成，其编码后具有正确的输入空间和标签空间之间的映射信息分布，因此可以将该映射信息分布作为标准，来判断该排序方式能否为LLM提供正确信息。本方法使用交叉熵损失函数为每种排序方式进行打分，如公式4所示。其中，$\mathcal{C}$代表分类类别标签的数目。将得到的损失值进行归一化后，选取损失最小的排序序列作为最终模板。

$$L(emb_{T_{c,r}}, emb_{T_{c,r},x_t}) = -\sum_{c=1}^{\mathcal{C}} emb_{T_{c,r}} log(emb_{T_{c,r},x_t}) \tag{4}$$

4 实验

4.1 数据集与实验设置

为了验证基于动态聚类与标签空间映射的上下文学习模板构建方法的有效性，本文选取七项自然语言处理领域的分类任务数据集进行实验，包括情感分析数据集SST-2、SST-5，自然语言推理数据集SNLI、MNLI、QNLI，以及主题分类任务数据集Trec、Ag_News。

本实验环境为CPU配置为Intel(R) Core(TM) i9-10980XE CPU@3.00GHz，GPU配置为NVIDIA GeForce RTX 3090的服务器，CUDA版本为12.2，使用Pytorch v2.1.0作为深度学习框架。由于本实验以实例级别进行，不同实例对应的上下文模板长度不同，因此进行推理时将batch size设为1。在实验过程中，将随机数种子分别设为1,2,3，选取最优结果。

4.2 对比实验

本文基于大语言模型GPT2-XL和Bloom，将基于动态聚类与标签空间映射的上下文学习模板构建方法与其他性能优秀的方法进行对比，在不同数据集上进行实验，实验结果如表1所示。

- Random：该方法采用随机选取的策略，为每一个数据集构建一个上下文模板。
- Prompt：该方法属于上下文学习的特殊情况，即不使用任何示例，直接进行预测。
- MI：(Sorensen et al., 2022)该方法选择最大化输入和输出之间的互信息进行模板构建，无需标记示例，也不需要使用模型进行大规模推理预测。
- GlobalE：(Lu et al., 2022)为了避免类别不平衡的预测问题，该方法使用局部熵值作为评价指标，为上下文模板选择合适的示例排列顺序。
- Top-K+LocalE：(Lu et al., 2022)该方法根据模型对输入的预测结果的偏移程度，判断模型的质量以及模型在不同类别之间进行分类的能力。使用Top-k方法进行示例选择后，进一步使用LocalE方法进行示例间排序。
- KATE：(Liu et al., 2022)该方法使用KNN算法辅助示例选择，并使用编码器将样例映射到语义空间中，检索语义上与测试样例相似的示例，构建与其相应的模板。

Model	Method	Dataset						
		SST2	SST5	SNLI	MNLI	QNLI	Trec	Ag_News
GPT2-XL(1.5B)	Random	31.47	26.24	44.75	40.98	52.35	19.40	33.07
	Prompt	47.12	29.46	41.98	39.06	50.45	13.80	29.76
	MI	52.86	35.35	46.02	41.32	50.62	16.00	47.29
	GlobalE	67.27	33.21	46.99	40.46	57.27	28.53	52.01
	Top-K+LocalE	63.21	34.03	51.71	42.46	53.45	35.80	83.42
	KATE	68.75	36.97	58.34	46.33	59.77	40.80	88.87
	Ours	**71.94**	**39.28**	**59.57**	**47.15**	**61.16**	**43.00**	**89.21**
Bloom(1.1B)	Random	34.87	22.31	43.39	41.87	49.65	18.00	32.95
	Prompt	25.04	26.02	42.26	39.85	49.61	5.80	25.13
	Top-K+LocalE	54.31	29.70	51.52	37.95	48.95	28.20	75.41
	KATE	62.49	35.84	**58.74**	43.37	54.99	39.40	85.50
	Ours	**71.39**	**36.97**	57.85	**44.12**	**55.41**	**39.60**	**86.82**

Table 1: 不同上下文模板组织方法在不同任务数据集上的性能对比

从表1的数据中可以看出，本文所提出的基于动态聚类与标签空间映射的上下文学习模板构建方法，相比于其他基于熵值、互信息等方法，在不同数据集、不同LLM上普遍具有优秀性能表现。其中，在SST-2数据集上，GPT2-XL模型准确率提升了3.19%，Bloom模型准确率提升了8.9%，验证了本文提出方法的有效性。本方法选取了比传统Top-K方法更多的相关样本，并过滤掉了无关样本，减少了干扰信息；利用聚类筛选增加了上下文模板中示例的多样性，使模型能够从多角度进行综合预测；合理选择示例排序方式来构建最终的上下文模板，为模型预测提供了正确的输入与标签之间的映射关系，最终提升了模型预测的准确率。

4.3 消融实验

我们使用GPT2-XL模型，分别验证本文提出的动态选择、聚类筛选和排序选择的有效性。

4.3.1 动态选择方法消融实验

图3展示了在七个数据集上针对动态选择方法的消融实验结果，Top-k代表根据欧式距离进行示例选择的传统方法。Topk-D使用本文提出的动态选择方法改进模板。结果显示，在不同数据集上，动态选择方法能够普遍提升模型预测准确率，说明其可以有效评估上下文模板示例与测试样本之间的相关度。而对于Ag_News数据集，由于其包含四种不同标签，映射关系与信息分布较为复杂。此外，不同于其他输入长度较短的数据集，Ag_News数据集的输入长度集中在200-400之间，对于较长文本中的语义信息，以及句子与标签之间的映射关系，相对更难获取，因此还需对示例进行进一步地筛选和排序。

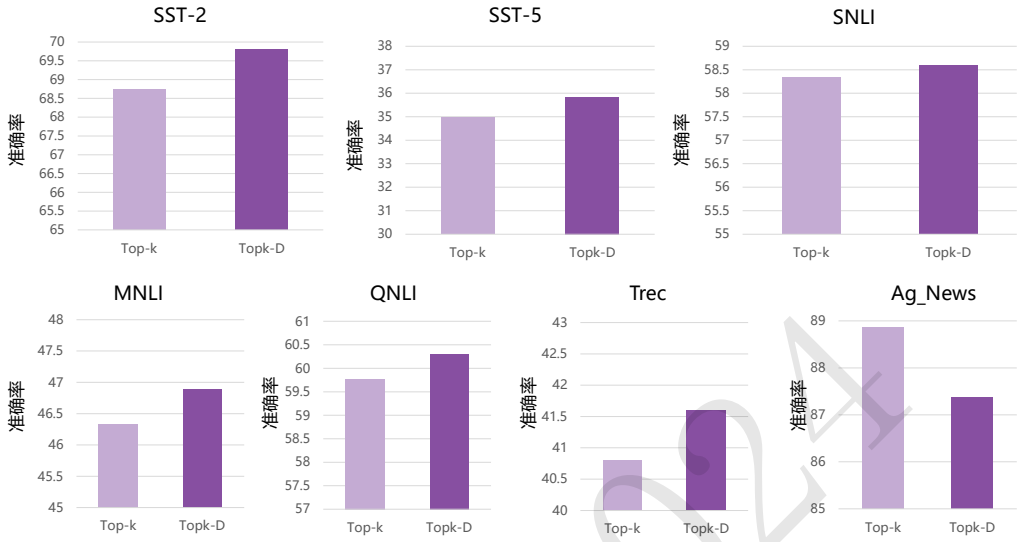

Figure 3: 动态选择方法的对比实验结果

4.3.2 聚类筛选方法消融实验

图4展示了针对聚类筛选方法的消融实验结果。其中“TopK-D+顺序抽取”首先使用TopK-D动态选取示例，随后采用聚类筛选方法中的顺序抽取策略得到示例集；“TopK+随机抽取”指在TopK-D算法动态选取示例后，使用聚类筛选方法中的随机抽取策略得到示例集。

Figure 4: 聚类筛选方法的消融实验结果

第二十三届中国计算语言学大会论文集，第883页-第893页，太原，中国，2024年7月25日至28日。
卷1：主会论文

4.3.3 排序选择方法消融实验

图5是针对排序选择模块进行的消融实验结果。其中“TopK-D+聚类筛选+排序选择”代表先使用TopK-D和聚类筛选算法进行示例选择，随后对示例进行排序选择。从图中的实验结果可以看出，本文提出的排序选择方法能够为LLM提供正确且充分的标签映射关系，在不同数据集上的准确率均得到有效提升，对于标签空间中标签较多、较复杂的数据集，如SST-5、Trec，本文的排序选择方法获得了更为显著的提升效果。

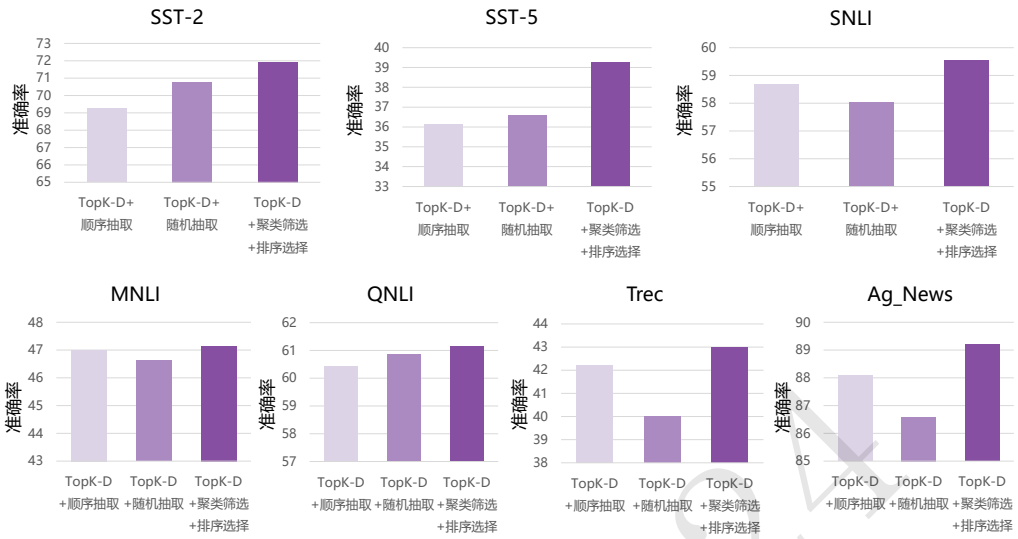

Figure 5: 排序选择的消融实验结果

4.4 超参数实验

为了探索不同超参数的设置对模型性能的影响，我们在Trec数据集上，使用GPT2-XL模型对动态选择模块进行了一系列超参数实验，如图6所示。

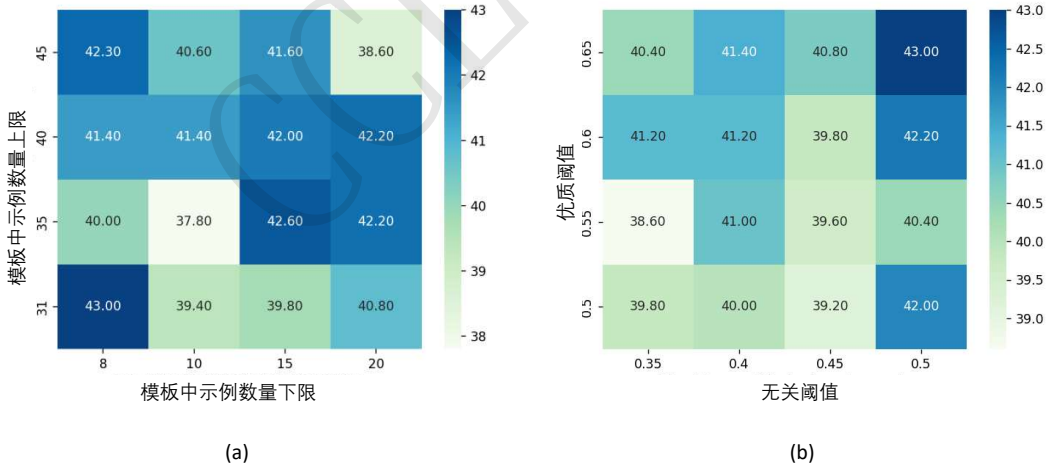

Figure 6: 动态选择模块的超参数实验结果

图6所示的热力图分别展示了动态选择模块中，模板示例数量的上限和下限、优质阈值和无关阈值的取值对性能的影响。图6(a)中结果显示，当模板中示例数量下限为8，上限为31时，LLM达到最优性能43%。当模板示例的数量下限提升时，准确率均有所下降，这说明当嵌入空间的测试样本周围稀疏时，向上下文模板中增加过多示例，存在引入无关示例的风险，会对LLM产生干扰；当模板示例的数量上限提升时，由于此时嵌入空间中的测试样本周

围稠密，相关示例较多，因此会提高LLM预测的准确率，但若此时降低示例的下限，提升上限同样容易引入无关示例，会对标签预测造成干扰。因此综合考虑示例数量的上限和下限，本文选择上限为31，下限为8的超参数组合进行动态选择，使用较少的数据量达到较高的性能。从图6(b)中显示的结果可以看出，当无关阈值取0.5，优质阈值取0.65时，动态选择模块能够使LLM达到43%的最优性能。当使用更小的无关阈值或更大的优质阈值时，准确率普遍有所下降，说明其他阈值超参数的组合对示例与测试样例之间距离的判断无法达到理想的示例动态筛选效果。

4.5 实例对比

为了分析本文提出方法的有效性，从Trec数据集中选取了主题分类任务的实例数据进行对比，如表2所示。

输入 分类标签	What is the full form of .com? ABBR							
Top-K 预测标签	ABBR ENTY	ABBR	ENTY	ABBR	ENTY	ENTY	ENTY	ABBR
Ours 预测标签	ENTY ABBR	ABBR	ABBR	ABBR	LOC	ENTY	ABBR	ABBR

Table 2: 本文方法(Ours)与Top-K方法所构建的模板和输出结果的对比，模板均由八个候选示例排列组成。其中ABBR代表与缩写相关的主题，ENTY代表与实体相关的主题，LOC代表与位置相关的主题

从表2的示例数据中可以看出，本文所提出的方法相比于Top-K方法，包含了更多与标签相同的映射分布，保证了预测的准确率，同时也更具有多样性，表明本文方法能够有效提示LLM做出正确预测。

5 总结

本文面向大语言模型中上下文学习模板的设计与组织这一任务，提出了基于动态聚类与标签空间映射的上下文学习模板构建方法。采用动态选择策略对Top-K算法进行优化，根据不同的测试样例设置优质阈值和无关阈值，以实例级别进行上下文示例的选择，保证模板的相关性；进一步提出了聚类筛选的方法，利用聚类算法将动态选择后的示例进行再次筛选，在保证示例相关性的前提下，增加上下文学习示例的多样性；提出基于映射损失的排序选择方法，选择包含正确映射关系的示例排序方式构建最终模板。实验在情感分析、自然语言推理和主题分类领域中的七个数据集中进行验证，并针对每个模块进行了充分的消融实验，结果表明本文提出的方法可以为模型提供高质量的上下文学习模板，有效提升LLM在不同任务上的准确率。未来的工作将关注如何将外部知识与上下文模板相融合，并提升LLM的鲁棒性，使其在使用不同上下文模板的条件下达到理想的效果。

参考文献

高怡, 纪焘, 吴苑斌, 牟小峰, and 王綻. 2023. 基于标签增强和对比学习的鲁棒小样本事件检测. 中文信息学报, 37(4): 98-108.

穆建媛, 朱毅, 周鑫柯, 李云, 强继朋, and 袁运浩. 2023. 基于提示学习的中文短文本分类方法. 中文信息学报, 37(7): 82–90.

言佳润and 鲜于波. 2023. 面向中文网络对话文本的论辩挖掘——基于微调与提示学习的大模型算法. 中文信息学报, 37(10): 139-148.

张少辉, 亓玉浩, 翟方文, 吕洪波, and 宋亦旭. 2022. 基于冗余距离筛选的UWB定位优化方法. 清华大学学报（自然科学版）, 62(5): 934-942.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems,2020,33:1877-1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al. 2023. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 2023, 24(240): 1-113.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2022. A survey for in-context learning. arXiv preprint arXiv:2301.00234.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying Prompts in Language Models via Perplexity Estimation. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 10136–10148, Singapore. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2022. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. arXiv preprint arXiv:2206.08082.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse Demonstrations Improve In-context Compositional Generalization. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3?. In Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified Demonstration Retriever for In-Context Learning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to Learn In Context. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 3505-3506.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning To Retrieve Prompts for In-Context Learning. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

R.J.Solomonoff. 1964. A formal theory of inductive inference. Information and Control, 7(1).

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 819–862, Dublin, Ireland. Association for Computational Linguistics.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. International Conference on Machine Learning, PMLR.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context Learning for Relation Extraction using Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 3534–3547, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35 (2022): 24824-24837.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc Le. 2023. Symbol tuning improves in-context learning in language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 968–979, Singapore. Association for Computational Linguistics.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1423–1436, Toronto, Canada. Association for Computational Linguistics.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. International Conference on Machine Learning. PMLR.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. arXiv preprint arXiv:2303.18223.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. arXiv preprint arXiv:2304.04675.