2022 Clustering of Short Texts Based on Dynamic Adjustment for Contrastive Learning
202200130267 张宇欣

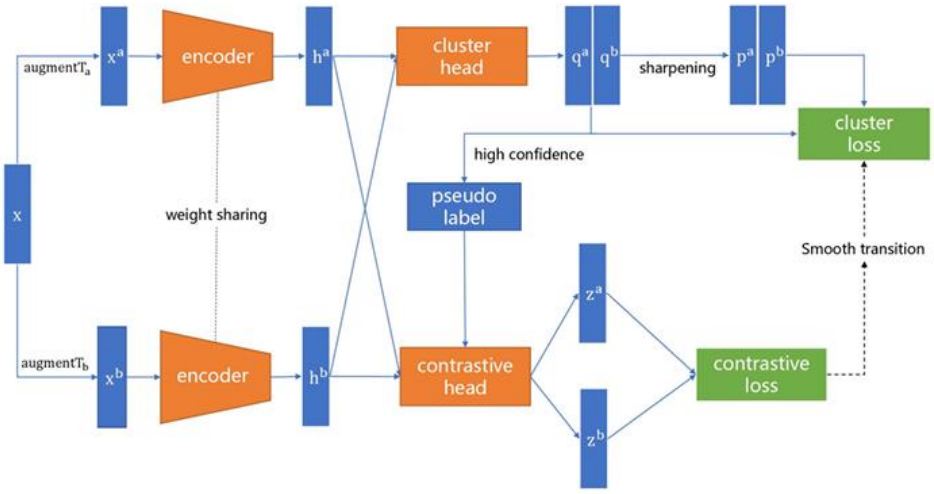
1. 问题陈述

短文本通常具有高维度、高稀疏性和高噪声的特性，传统聚类方法（如 K-Means 和 GMM）由于依赖初始空间的距离度量，无法很好地表征短文本的语义相似性。对比学习专注于样本的实例级特征表示学习，但忽略了数据在语义级别上的相关性，尤其是同一聚类内的样本语义。将对比学习和聚类结合时，现有方法可能会把同一聚类内的样本视为负样本，导致聚类空间稀疏，影响聚类质量。对比学习中的伪标签生成如果不准确，会导致后续训练中误差的累积，从而进一步降低聚类的可靠性。

2. 方法概述

作者提出了一种基于动态调整对比学习权重的聚类方法（DACL）。模型将对比学习和聚类任务结合为一个端到端的联合训练框架。模型的核心思想是通过动态调整对比损失和聚类损失的权重，逐步过渡从特征表示的优化到聚类质量的提升，从而实现更高效的短文本聚类。

模型框架：



该模型由三部分组成，即编码器 f（）、实例级对比头 gI（） 和语义级聚类头 gC（）。编码器负责将样本嵌入到低维特征空间以提取特征，聚类头负责计算分布概率。每个聚类簇分配样本，而对比头负责将样本嵌入到对比空间中。由聚类头计算出的聚类分布概率可用于在对比头中筛选负样本实例。聚类目标损失和对比目标损失用于训练模型。聚类目标损失通过锐化样本的聚类分布概率来提高聚类结果的置信度，而对比目标损失通过最大化正样本对的距离并最小化负样本实例的距离来优化样本的特征表示。在训练过程中，聚类损失和对比损失动态相加形成总损失，总损失权重逐渐从对比损失过渡到聚类损失。

1.编码器：用于将样本嵌入到低维特征空间，提取特征表示。
通过上下文增强生成数据，即对输入文本随机掩码，用预训练语言模型预测并替换被掩盖的

单词。用两个不同的预训练语言模型，对每个样本生成一对增强数据。使用共享参数的深度神经网络作为编码器，从增强数据中提取特征。

2. 语义级聚类头：将样本分成 K 个不相交的聚类，使同一聚类内的语义尽可能相似。

使用 K-Means 初始化每个聚类的质心。使用学生 t-分布计算样本 x_i 与聚类中心 k_k 的相似度，公式为：

$$\boldsymbol{q}_{ik} = \frac{(1 + ||\boldsymbol{h}_i - \boldsymbol{\mu}_k||_2^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{k'=1}^{K}(1 + ||\boldsymbol{h}_i - \boldsymbol{\mu}_{k'}||_2^2/\alpha)^{-\frac{\alpha+1}{2}}} \qquad (1)$$

从而得到样本的软分配概率分布 qi。定义辅助目标函数，通过归一化后的高置信度分配来优化样本表示，

$$p_{ik} = \frac{q_{ik}^2/\sum_{i=1}^{M} q_{ik}}{\sum_{k'=1}^{K} q_{ik}^2/q_{ik'}}$$

通过最小化 KL 散度优化样本的聚类表示：

$$\begin{aligned}\mathcal{L}_C &= \frac{1}{2}(KL[\boldsymbol{p}^b||\boldsymbol{q}^a] + KL[\boldsymbol{p}^a||\boldsymbol{q}^b]) \\ &= \frac{1}{2M}\sum_{i=1}^{M}\sum_{k=1}^{K}(p_{ik}^b log\frac{p_{ik}^b}{q_{ik}^a} + p_{ik}^a log\frac{p_{ik}^a}{q_{ik}^b})\end{aligned}$$

3. 实例级对比头：最大化正样本对的相似性，最小化负样本的相似性。

基于聚类软分配概率分布生成伪标签，样本与伪标签不同的其他样本视为负样本，伪标签相同的样本视为正样本。使用非线性 MLP 投影特征表示到子空间。在子空间中计算正样本对和负样本的对比损失：

$$l_i^a = -log\frac{exp(sim(\boldsymbol{z}_i^a, \boldsymbol{z}_i^b)/\tau)}{\sum_{j\in S_i} exp(sim(\boldsymbol{z}_i^a, \boldsymbol{z}_j^a)/\tau) + exp(sim(\boldsymbol{z}_i^a, \boldsymbol{z}_i^b)/\tau)}$$

其中 sim(·,·) 是余弦相似度。

4. 动态调整的目标损失函数

总损失由聚类损失 Lc 和对比损失 Li 组成：

$$\mathcal{L} = (\theta + (1-\lambda))\mathcal{L}_C + \sigma(\theta + \lambda)\mathcal{L}_I$$

初始阶段主要通过对比损失学习样本特征表示。随着训练的进行，逐渐降低对比损失权重，增加聚类损失权重：

$$\lambda(l) = \frac{1}{2}\cos(\frac{l}{L}\pi) + \frac{1}{2}$$

DACL 模型基于以下方法：

SCCL (Supporting Clustering with Contrastive Learning)：SCCL 是将对比学习与聚类结合的早期框架，强调通过对比学习提高聚类分离性。

基于学生 t 分布的软分配策略：通过 t 分布计算样本与聚类中心的相似性，生成软分配概率。对比学习增强特征表示：借鉴对比学习框架（如 SimCLR），通过正样本对和负样本对优化特征质量。

算法流程图：

Algorithm 1 Clustering With Dynamic Adjustment for Contrastive Learning

Input: dataset X; iteration number L; batchsize M; cluster number K; pre-train encoder $f(\cdot)$; augmentations T_a, T_b; non-linear mlp $g_I(\cdot)$

Output: dataset X corresponding predicted label vector $Y(Y \in \mathbb{R}^{N\times 1})$

//training
initialization cluster center μ by K-Means
for $l = 1$ to L **do**
  randomly select M samples as mini-batch $B = \{\boldsymbol{x}_i\}_{i=1}^{M}$ from dataset X
  compute feature representations $\boldsymbol{h}_i^a, \boldsymbol{h}_i^b$ by $\boldsymbol{h}_i^a = f(T_a(\boldsymbol{x}_i)), \boldsymbol{h}_i^b = f(T_b(\boldsymbol{x}_b))$
  compute cluster soft-assignments $\boldsymbol{q}^a, \boldsymbol{q}^b, \boldsymbol{p}^a, \boldsymbol{p}^b$ by Eq.(1) and Eq.(3)
  comput cluster loss $\mathcal{L}_C$ by Eq.(4)
  comput subspace representations $\boldsymbol{z}_i$ by $\boldsymbol{z}_i = g_I(\boldsymbol{h}_i)$
  acquire peseudo-labels y_i by Eq.(5) and Eq.(6)
  filter negative sets S_i through Eq (7)
  comput contrastive loss $\mathcal{L}_I$ by Eq.(10)
  comput total loss $\mathcal{L}$ by Eq.(11)
  update f, g_I, μ by minimize $\mathcal{L}$
end for
// test
define Y is predicted label vector
for $\boldsymbol{x}$ in X **do**
  compute feature representation $\boldsymbol{h} = f(\boldsymbol{x})$
  compute cluster soft-assignments $\boldsymbol{q}$ by Eq.(1)
  compute cluster assignment by $\boldsymbol{y} = argmax_k\ \boldsymbol{q}$
  $Y = [Y||\boldsymbol{y}]$
end for

3. 实验

实验在八个基准数据集上进行，数据集包括 SearchSnippets、StackOverflow、Biomedical、AgNews、Tweet，以及从 GoogleNews 数据集中分别提取的标题和摘要，数据集中不进行任何预处理。实验数据集：

Dataset	$\|V\|$	Documents		Clusters	
		N^D	Len	N^C	L/S
SearchSnippets	31K	12340	18	8	7
StackOverflow	15K	20000	8	20	1
Biomedical	19K	20000	13	20	1
AgNews	21K	8000	23	4	1
Tweet	5K	2472	8	89	249
GoogleNews-TS	20K	11109	28	152	143
GoogleNews-T	8K	11109	6	152	143
GoogleNews-S	18K	11109	22	152	143

作者未提供实验代码

作者在实验中对比了以下 8 种代表性短文本聚类方法：
BoW：使用词袋模型选取最常出现的 1500 个词。
TF-IDF：使用 TF-IDF 特征构建 1500 维向量，随后应用 K-Means。
K-Means：对模型的初始编码器特征进行聚类。
DEC：使用深度神经网络联合优化特征表示和聚类分配。
STTC：基于 Word2Vec 嵌入的特征空间，结合卷积神经网络和 K-Means。
Self-Train：使用 SIF（简单的句子向量）嵌入，通过自编码器重构短文本向量并进行聚类。
HAC-HD：采用基于层次聚类的方法对稀疏的相似度矩阵进行操作。
SCCL：使用对比学习和锐化分布的联合优化。

实验结果：
模型相较于现有的基线模型表现最佳，尤其是同样采用对比学习与聚类相结合的 SCCL 模型。与 SCCL 相比，DACL 通过动态调整 损失权重和负样本过滤，并通过在语义级聚类损失中将 增强样本的聚类分布推向其对应正样本对的目标分布， 从而取得了更好的性能。对于数据量较少且聚类数量较多的数据集（如 GoogleNews 和 Tweet），传统方法（如 HAC-HD）有时能取得更好的效果。

Model	Searchnippets		StackOverflow		Biomedical		AgNews	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
BoW	24.3	9.3	18.5	14.0	14.3	9.2	27.6	2.6
TF-IDF	31.5	19.2	58.4	58.7	28.3	23.2	34.5	11.9
K-Means	59.0	36.4	60.8	52.3	39.8	32.7	83.9	59.2
DEC	76.9	64.9	74.7	75.3	41.6	37.7	-	-
STCC	77.0	63.2	59.8	54.8	43.6	38.1	-	-
Self-Train	77.1	56.7	64.8	64.8	**54.8**	**47.1**	-	-
HAC-SD	82.7	63.8	64.8	59.5	40.1	33.5	82.8	54.6
SCCL	85.2	71.1	75.5	74.5	46.2	41.5	88.2	68.2
DACL	**86.1**	**73.6**	**77.5**	**76.0**	48.6	40.3	**88.6**	**69.0**
Model	Tweet		GoogleNews-TS		GoogleNews-T		GoogleNews-S	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
BoW	49.7	73.6	57.5	81.9	49.8	73.2	49.0	73.5
TF-IDF	57.0	80.7	68.0	88.9	58.9	79.3	61.9	83.0
K-Means	51.7	79.0	56.0	78.4	62.2	83.3	67.8	87.5
DEC	-	-	-	-	-	-	-	-
STCC	-	-	-	-	-	-	-	-
Self-Train	-	-	-	-	-	-	-	-
HAC-SD	**89.6**	85.2	85.8	88.0	**81.8**	84.2	80.6	83.5
SCCL	78.2	89.2	**89.8**	**94.9**	75.8	88.3	83.1	90.4
DACL	82.0	**90.6**	89.6	94.4	80.5	**89.8**	**84.0**	**91.7**

4. 思考

论文提出了一种动态调整对比学习和聚类损失权重的方法，解决了对比学习与聚类目标之间的不一致性问题。通过筛选负样本，提高了对比学习过程中样本表征的准确性，从而提升聚类性能。提供了针对互联网大规模短文本的高效聚类方法，优化了短文本的分类和表示能力。在 8 个数据集上评估了方法的有效性，表现出比 SOTA 更优或相当的效果。

当前方法在应对动态生成的海量文本数据（如数据流）时存在局限性，无法实时处理数据特征变化。对计算复杂度、运行时间等实际性能指标的分析较少，尤其是针对大规模数据的处理能力，仍需更详细的讨论。

论文提出了一种创新的动态对比学习与无监督聚类相结合的方法，在多个数据集上取得了显著的效果。然而，其在处理大规模数据和实时流数据场景下仍存在挑战。未来研究可以从增量学习、初始化改进和效率优化等方向进一步提升模型的适用性和实际表现。

5. 其它（选填）

需要特别记录的其它笔记