山西大学学报(自然科学版)47(4):727—736,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023065

融入事件知识的新闻事件对比聚类方法

梁晨[1,2],余正涛[1,2],高盛祥[1,2*],朱恩昌[1,2]
(1.昆明理工大学 信息工程与自动化学院,云南 昆明 650500;
2.云南省人工智能重点实验室,云南 昆明 650500)

摘　要:新闻事件聚类旨在从海量新闻文本中挖掘若干个不同主题的事件簇。目前事件聚类大多基于文本语义表征,忽略了事件知识的指导作用,且由于表征学习和目标聚类的迭代进行,不仅容易造成误差累积,还只能处理离线任务,限制了对实时新闻数据的处理。针对上述问题,该文提出一种融入事件知识的新闻事件对比聚类方法,该方法在文本表征的基础上,融入事件关键信息丰富事件表征;将聚类标签作为表示,同时在实例级和簇群级进行对比学习;以端到端的方式联合学习表示和簇群分配,实现对数据流的聚类。实验结果表明,该方法相较于其他基线模型,提高了3%。

关键词:事件聚类;事件表征;对比学习;深度聚类
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:0253-2395(2024)04-0727-10

A Contrastive Clustering Method of News Events Incorporating Event Knowledge

LIANG Chen[1,2], YU Zhengtao[1,2], GAO Shengxiang[1,2*], ZHU Enchang[1,2]
(1.*Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650500*, China;*
2.*Yunnan Key Laboratory of Artificial Intelligence, Kunming* 650500*, China*)

Abstract: News event clustering aims to mine several event clusters of different topics from massive news texts. At present, event clustering is mostly based on text semantic representation, but ignoring the guiding role of event knowledge. Due to the iterative process of representation learning and target clustering, it is easy to cause error accumulation. It can only deal with offline tasks, which limits the processing of real-time news data, and to solve the above problems, this paper proposes a contrastive clustering method of news events incorporating event knowledge. On the basis of text representation, this method incorporates event key information to enrich event representation. The cluster label is used as the representation, and contrastive learning is performed at the instance level and the cluster level. The representation and cluster assignment are jointly learned in an end-to-end manner to realize the clustering of data streams. Experimental results show that the proposed method improves by 3% compared with other baseline models.

Key words: event clustering; event representation; contrastive learning; deep clustering

收稿日期:2022-12-29;**接受日期**:2023-03-23
基金项目:国家自然科学基金(61972186;61732005;U21B2027);云南高新技术产业发展项目(201606);云南省重大科技专项计划(202103AA080015;202002AD080001-5);云南省基础研究计划(202001AS070014);云南省学术和技术带头人后备人才(202105AC160018)
作者简介:梁晨(1997-),女,山西吕梁人,硕士研究生,主要研究方向为自然语言处理。E-mail:liangchenshanxi@126.com
* **通信作者**:高盛祥(GAO Shengxiang),E-mail:gaoshengxiang.yn@foxmail.com
引文格式:梁晨,余正涛,高盛祥,等.融入事件知识的新闻事件对比聚类方法[J].山西大学学报(自然科学版),2024,47(4):727-736. DOI:10.13451/j.sxu.ns.2023065

0 引言

事件是信息的重要载体,涉及特定的时间、地点、人物和组织机构等内容[1]。新闻事件聚类旨在将海量新闻文本聚类成不同主题的事件簇,它是信息检索与文本挖掘领域中的重要组成部分,被广泛应用于舆情分析、信息推荐等任务中。海量新闻数据中存在描述不同事件的新闻文本,而一个新闻事件可以通过所涉及的新闻主题、时间、地点来唯一标识,因此本文融入这三种事件知识作为指导信息,区分不同事件的同时丰富事件表征,增强事件聚类效果。

近年来,事件聚类的研究致力于将深度表示学习和聚类方法结合起来。针对无监督事件聚类效果差、有监督需要大量标注数据的问题,郭恒睿等[2]提出一种半监督事件聚类模型(Semi-supervised Chinese Incremental Event Clustering Model, SemiEC),该模型在小规模标注数据的基础上,通过LSTM(Long Short-term Memory)表征事件,计算相似度后进行增量聚类,利用聚类产生的标注数据对模型再训练,结束后对不确定样本再聚类。实验表明,SemiEC相较于传统模型性能有所提高。但由于在聚类开始前,不同类别数据在表示空间中存在交叉重叠,只改进聚类方法提升效果有限。为此,Zhang等[3]提出基于对比学习的短文本聚类方法(Supporting Clustering with Contrastive Learning, SCCL),利用对比学习拉近同源样本,拉远非同源样本,联合优化聚类损失和对比损失得到更优的表征,从而实现更好的聚类效果。

上述方法在短文本事件聚类任务上取得了较大进展,但在新闻长文本事件聚类上还存在以下问题:(1)新闻文本作为长文本,存在文本内容冗余、事件特征模糊、噪声干扰过多、语义复杂度高等问题,深层语义建模困难;(2)现有基于对比学习的聚类研究,通过迭代进行表征学习和目标聚类以实现更好的聚类结果,但迭代交替过程中会造成误差累积,影响聚类效果;(3)同时,迭代学习只能处理离线任务,无法对实时新闻聚类处理。

基于此,本文提出了一种融入事件知识的新闻事件对比聚类模型NECC(News Event Contrastive Clustering Model by Integrating Event Knowledge)。首先,对原始新闻文本通过Dropout进行数据增强;其次,在获得全局语义表征的基础上,融入新闻主题、时间、地点三种事件知识增强全局重要信息,丰富事件表征;然后,将聚类标签作为表示,同时在实例级和簇群级上进行双重对比聚类;最后,以端到端的方式联合学习事件表征和簇群分配,实现对实时新闻的聚类。实验表明,NECC模型与其他基准模型相比,在各项聚类指标上均得到了提高。

1 相关工作

1.1 文本聚类

文本聚类是一种无监督任务,旨在将相似的文本归档在一起,并不关注文本的具体类别。传统的文本聚类是基于文档中词的相似性,常见方法是通过向量空间模型得到文本表征,再利用余弦相似度计算文本相似性,最后采用K-Means方法[4]或Single-pass[5]进行聚类。但由于词在不同语境下的含义有所区别,所以Hofmann等[6]提出概率潜语义分析算法(Probability Latent Semantic Analysis, PLSA),在文本和词之间加入主题,先让文本和主题产生关联,再从主题中寻找词的概率。Blei等[7]在PLSA的基础上加入Dirichlet先验分布,提出LDA(Latent Dirichlet Allocation)主题模型,解决了模型的过拟合问题。但由于上述方法难以准确提取文本的事件特征,导致表示不充分,所以聚类的准确度较低、稳定性较差。

近年来,研究人员逐渐通过深度表示学习提升聚类效果。深度聚类主要分为两部分,先将输入转换为新的表示,再基于新的表示进行聚类。郭恒睿等[2]针对无监督聚类效果差、有监督需要大量标注数据的问题,提出一种半监督事件聚类模型(SemiEC),该模型在小规模标注数据的基础上,通过LSTM表征事件,计算相似度后进行增量聚类,利用聚类产生的标注数据对模型再训练,结束后对不确定样本再聚类。Zhang等[3]针对聚类开始前不同类别数据在表示空间中存在交叉问题,提出了基于对比学习的短文本聚类方法SCCL,利用对比学习

拉近同源样本,拉远非同源样本,联合优化聚类损失和对比损失得到更优的表征,实现了更好的聚类效果。

1.2 对比学习

对比学习是一种自监督学习方法,通过缩短表示空间中正实例与锚点间距离,同时拉远负实例与锚点间距离,学习到更优的语义表示,以辅助其他下游任务。

对比学习的关键在于使用何种数据增强方式构造相似实例和不相似实例。传统的文本增强有基于同义词的替换、基于掩码语言模型的删除和替换、回译等方法,但需要控制样本数量、少量学习,并且容易改变文本语义。Gao等[8]提出SimCSE(Simple Contrastive Sentence Embedding Framework),使用不同的Dropout实现数据增强;Chuang等[9]提出DiffCSE(Difference-based Contrastive Learning for Sentence Embeddings),同时使用基于Dropout的数据增强和基于掩码建模的单词替换;Fang等[10]提出CERT(Contrastive self-supervised Encoder Representations from Transformers),使用回译实现数据增强;Wu等[11]提出ESimCSE(Enhance SimCSE),执行单词重复操作实现数据增强;Zhang等[12]提出VaSCL(Virtual Augmentation Supported Contrastive Learning of Sentence Representations),通过添加高斯噪声得到正样本,再从k近邻中得到距离最近的k个负样本。

与文本聚类不同,事件聚类旨在根据不同的事件知识实现事件粒度上的文本聚类。传统的文本聚类模型难以准确提取新闻长文本的事件知识,事件聚类准确度较低。我们提出融入新闻主题、时间、地点三种事件知识增强全局重要信息,丰富事件表征。传统基于对比学习的文本聚类方法,通常迭代进行文本表示和目标聚类,容易造成误差累积。我们提出将每个样本的聚类软标签作为其特征表示,同时在实例级和簇群级上进行双重对比聚类,以端到端的方式联合学习事件表征和簇群分配,实现对实时新闻数据的聚类。

2 模型

针对新闻事件聚类任务,本文提出了一种融入事件知识的新闻事件对比聚类模型NECC,模型由数据增强、事件知识抽取、特征提取和对比聚类四部分组成,其模型架构图如图1所示。

2.1 数据增强

文本数据增强可以解决样本少、分布不均衡等问题,有效提升模型的泛化能力。传统文本增强有随机替换、随机删除、随机插入等方法,但这些方法很容易造成文本语义的改变,带来一定的噪声干扰。Gao等[8]发现仅通过Dropout进行数据增强比传统方法效果更好。基于上述思想,本文通过Dropout进行数据增强。Dropout是指在深度学习网络训练过程中,通过随机去除部分神经元的连接从而改变模型输出,防止模型过拟合,提高模型的鲁棒性。

具体来说,对于随机采样的每一个样本集B,大小为M。为B中的每一个实例$\tilde{X}_i$生成一对增强数据X_i^a,X_i^b,最终得到一个增强样本集$B^K=\{X_i\}_{i=1}^{2M}$,公式如下:

$$X_i^a=T^a(\tilde{X}_i), \tag{1}$$

$$X_i^b=T^b(\tilde{X}_i), \tag{2}$$

其中$i\in\{1,\cdots,2M\}$表示增强数据集B^K中的任意实例,增强实例X_i^a,X_i^b是将来自B中的同一个实例$\tilde{X}_i$传递给预先训练好的编码器两次,产生的两个相近表示。T^a,T^b表示不同的Dropout机制。将X_i^a,$X_i^b\in B^a$作为正实例,而B^K中的其他$2M-2$个样本作为$\tilde{X}_i$的负实例。

2.2 事件知识抽取

现有的文本表征大多依赖于文本语义,但新闻文本是长文本,存在文本内容冗余、事件特征模糊、噪声干扰、语义复杂等问题。而一个新闻事件可以通过其涉及的新闻主题、时间和地点来唯一标识。因此,我们在获得全局语义表征的基础上,融入这三种事件知识作为指导信息,区分不同事件的同时,丰富事件表征,增强聚类效果。

命名实体识别可以从新闻文本中识别并提取涉及新闻事件的时间类实体和地点类实体,保留重要信息并过滤掉相对不重要的内容。如图2所示,本文使用BERT(Bidirectional Encoder Representation from Transformers)命名实体识别模型[13]进行命名实体识别任务,对输入的新闻

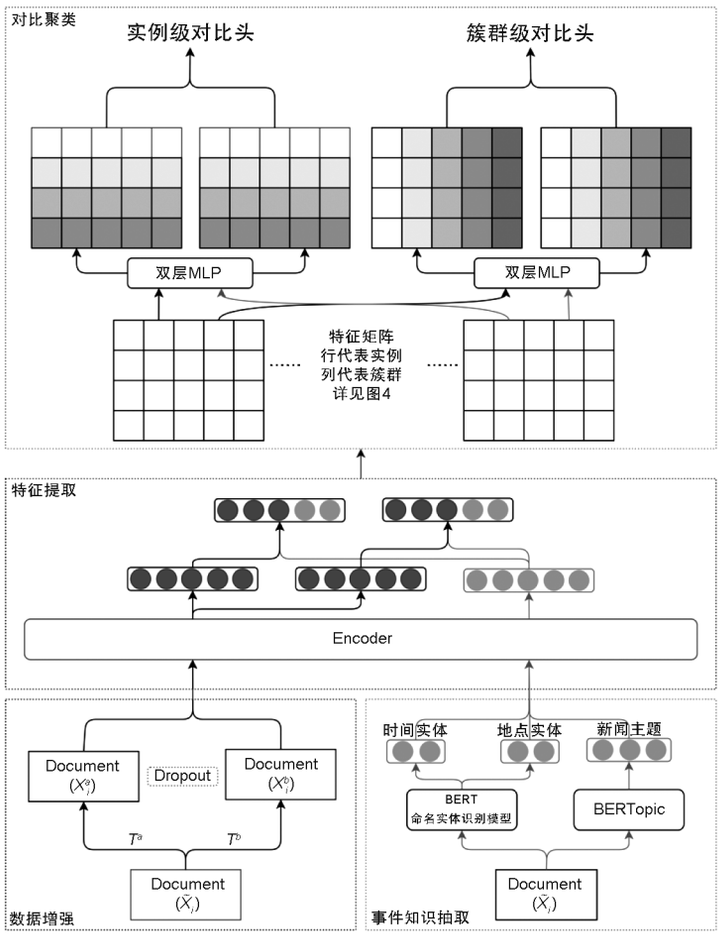

图1 NECC模型架构图

Fig. 1 Diagram of the NECC model architecture

文本预处理,进行分词和词性标注,输出句子中的时间实体和地点实体,拼接形成最后需要的事件知识。

我们采用BERTopic[14]实现新闻文本的主题建模。BERTopic是一种在线主题建模技术,利用BERT嵌入和C-TF-IDF(Class-based Term Frequency-Inverse Document Frequency)创建密集的簇,BERTopic不再使用单个文档,而是将一个簇中的所有文档视为一个文档,比较簇内所有词的重要性得分,这样得到的主题更具有代表性。具体而言,首先使用Sentence-BERT预训练模型[15]计算提取文档的向量表示;其次,基于UMAP(Uniform Manifold Approximation and Projection)方法对高维稀疏的文档向量进行降维;然后,再通过HDBSCAN算法对文本表征进行聚类,语义相近的文本聚成簇;最后,使用基于簇的C-TF-IDF提取每个簇的主题词,实现新闻文本的主题建模,有效缓解了基于密度聚类和基于中心采样的偏差,具体公式如下:

$$W_{t,c}=tf_{t,c}\times\log\left(1+\frac{A}{tf_t}\right), \tag{3}$$

其中t表示某个单词,c表示簇,A表示每个簇的平均词数,$tf_{t,c}$表示词t在簇c中出现的频率,tf_t表示词t在所有簇中出现的频率,$W_{t,c}$表示词t在簇c中的重要性得分,tf_t值越高,越说明t是一个常用词而不是指代某一个簇的词,因此越不能代表这个簇的主题。

2.3 特征提取

将增强后的文本数据X_i和事件知识Y_i作为输入,然后通过特征提取器进行编码,拼接得到最终的事件表征,具体公式如下:

$$H_i=Encoder(X_i)\oplus Encoder(Y_i), \tag{4}$$

$$Encoder(Y_i)=f(T_i^1)\oplus f(T_i^2)\oplus f(L_i), \tag{5}$$

其中H_i为X_i和Y_i经过编码器后的输出表征,Y_i表示$\tilde{X}_i$中的事件知识,T_i^1表示新闻事件主

题,T_i^2表示时间实体,L_i表示地点实体,$\oplus$表示拼接操作,$f(\cdot)$和$Encoder(\cdot)$均表示特征提取操作。

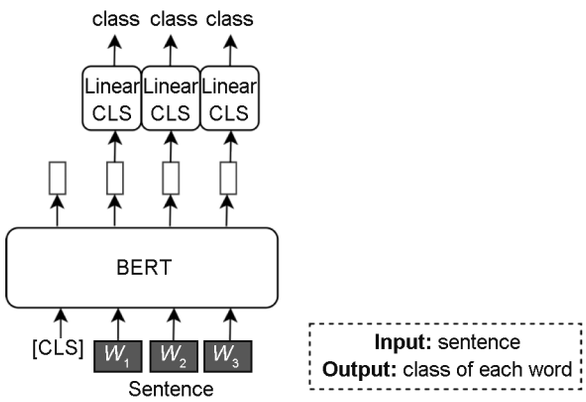

图2 BERT命名实体识别模型图

Fig. 2 Diagram of BERT named entity recognition model

2.4 对比聚类

Zhang等[3]提出了基于对比学习的短文本聚类方法(SCCL),利用对比学习拉近同源样本,拉远非同源样本,联合优化聚类损失和对比损失得到更优的表征,实现更好的聚类效果。其对比学习是在实例层面,横向进行对比。Li等[16]在计算机视觉领域提出了一种单阶段的对比学习聚类方法,提出"每个样本的聚类软标签可作为特征表示"的思想,通过增加一个聚类层面的对比学习,进行纵向对比。受上述思想启发,我们在文本领域提出将聚类任务统一到表示学习框架下,同时进行表示学习和目标聚类。

如图3所示,神经网络在给定的数据集上学习到特征矩阵后,该特征矩阵的每一行可以视为一个实例在不同簇下的表示,而每一列可以视为某个簇下不同样本的表示。所以,特征矩阵的行可以看作是簇分配的概率,列可以看作实例在该簇上的分布,通过同时最小化横向和纵向两个层面的损失函数,最终得到对比聚类下的网络权重,进而得到聚类结果。

2.4.1 实例对比

在实例对比层面,对于每篇新闻文本$\tilde{X}_i$,数据增强后都有一个正样本对和$2M-2$个负样本,融入事件知识Y_i后,拼接得到最终的事件向量$\boldsymbol{H}_i$。

具体而言,给定一个大小为M的Mini-Batch,NECC为每个实例进行两次Dropout数据增强,并拼接所含的三种事件知识,最终产生$2M$个数据样本$\{H_1^a,\cdots,H_i^a,\cdots,H_M^a,H_1^b,\cdots,H_i^b,\cdots,H_M^b\}$。对于一个特定的实例$H_i^a$,我们从$2M-1$对数据中,选择其对应的增强样本$H_i^b$,使其形成正样本对$\{H_i^a,H_i^b\}$,其余$2M-2$对为负样本对。

Wang等[17]提出MLP(Multilayer Perceptron)可以在无监督任务中提升当前数据集的表征能力。为了减轻对比损失带来的信息损失,我们不直接对特征矩阵进行对比学习,而是通过堆叠一个两层的MLP $g_I(\cdot)$,将特征矩阵映射到一个子空间后进行对比学习,通过余弦相似度衡量样本对间的相似性,公式如下:

$$s(z_i^{k_1},z_j^{k_2})=\frac{(z_i^{k_1})(z_j^{k_2})^T}{\|z_i^{k_1}\|\|z_j^{k_2}\|},\tag{6}$$

$$Z_i=g_1(\boldsymbol{H}_i),\tag{7}$$

其中k_1、$k_2\in\{a,b\}$,i、$j\in\{1,M\}$,Z表示H经过实例对比头的输出,单个实例对比损失函数为:

$$l_i^a=-\log\frac{\exp(s(z_i^a,z_j^b)/\tau_{\mathrm{I}})}{\sum_{j=1}^{M}\left[\exp(s(z_i^a,z_j^a)/\tau_{\mathrm{I}})+\exp(s(z_i^a,z_j^b)/\tau_{\mathrm{I}})\right]},\tag{8}$$

其中τ_{I}是温度参数,一般设置为0.5。实例级对比学习在整个数据集上的损失函数为:

$$L_{\mathrm{ins}}=\frac{1}{2M}\sum_{i=1}^{M}(l_i^a+l_i^b)。\tag{9}$$

2.4.2 簇群对比

如图2所示,我们将样本数据映射到一个维度等于簇数的表示空间中,因为样本的聚类软标签可以作为表示,所以当维度等于簇数时,特征的第i个元素也可以理解为它属于第i个簇的概率。

具体而言,令$C\in R^{M\times N}$表示H经过簇群对比头后的输出,$C_{m,n}$表示样本m属于簇n的概率,M表示批量大小,N表示簇的个数。由于每个样本只属于一个簇,所以理想情况下,C的行可以视为一个One-hot编码,同时第i列也可以看作是第i个簇的表示。

我们同样使用一个两层MLP $g_C(\cdot)$将特征矩阵投影到一个N维表示空间中,具体公式如下:

$$C_i=g_C(\boldsymbol{H}_i),\tag{10}$$

其中C_i是C的第i行,也是样本$\tilde{X}_i$的软标签。

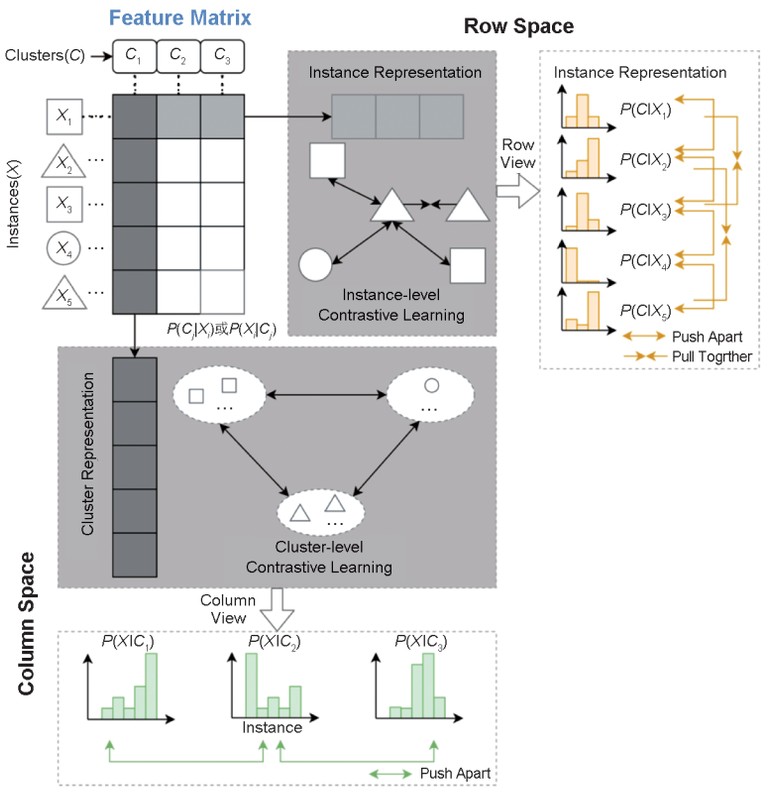

图3 特征矩阵中行、列含义图

Fig. 3 Meaning diagram of rows and columns in feature matrix

令 $\hat{C}_i$ 为 C 的第 i 列，构建簇群级正样本对 $\{\hat{c}_i^a, \hat{c}_i^b\}$，其余 $2N-2$ 个构成负样本对。与实例对比类似，使用余弦相似度衡量簇间的相似性，公式如下：

$$s(\hat{c}_i^{k_1}, \hat{c}_j^{k_2}) = \frac{(\hat{c}_i^{k_1})(\hat{c}_j^{k_2})^T}{\|\hat{c}_i^{k_1}\| \, \|\hat{c}_j^{k_2}\|}, \tag{11}$$

其中，k_1、$k_2 \in \{a, b\}$，i、$j \in \{1, N\}$，单个簇群对比损失函数为：

$$\hat{l}_i^a = -\log \frac{\exp(s(\hat{c}_i^a, \hat{c}_i^b)/\tau_{\mathbb{I}})}{\sum_{j=1}^{N}\left[\exp(s(\hat{c}_i^a, \hat{c}_i^a)/\tau_{\mathbb{I}}) + (s(\hat{c}_i^a, \hat{c}_i^b)/\tau_{\mathbb{I}})\right]}, \tag{12}$$

其中 $\tau_{\mathbb{I}}$ 是温度参数，一般设置为0.5。所有簇的对比损失函数为：

$$L_{\text{clu}} = \frac{1}{2N}\sum_{i=1}^{N}(l_i^a + l_i^b) - E(C), \tag{13}$$

$$E(C) = -\sum_{i=1}^{N}\left[P(\hat{c}_i^a)\log P(\hat{c}_i^a) + P(\hat{c}_i^b)\log P(\hat{c}_i^b)\right], \tag{14}$$

$$P(\hat{c}_i^k) = \frac{1}{M}\sum_{t=1}^{M} Y_{ti}^k, \tag{15}$$

其中，$k \in \{a, b\}$，$P(\hat{c}_i^k)$ 表示簇分配概率，$E(C)$ 表示簇分配概率 $P(\hat{c}_i^k)$ 的熵，可以避免大量实例分配至同一个簇。

整体损失函数由实例损失和簇群损失组成，即：

$$L = L_{\text{ins}} + L_{\text{clu}}。 \tag{16}$$

3 实验结果分析

3.1 数据集

本文在构建的中文新闻事件数据集上对NECC模型进行实验验证。我们从新浪新闻、中国日报网、腾讯新闻等国内主流新闻媒体上根据事件关键词爬取涉及当下热点事件的新闻文本，经过人工清洗与数据过滤后，最终获得涉及8种关键词、8 835篇新闻文档。构建的新闻事件数据集详细信息如表1所示。

3.2 评价指标

本文的评价指标使用的是 *Purity*、*ARI*（RandIndex）和 *NMI*（Normalized Mutual Information）。其中，*Purity* 是指计算准确的文本数量占总文本数量的比例。*ARI* 对获取到的标签与真实标签之间的差异进行对比，改善了 *RI* 惩罚不足的

问题。*NMI*是一个基于熵的评价指标。

表1 数据集统计结果

Table 1 Results of dataset statistics

文本关键词	文档数/篇	平均长度/字
东航飞机失事	1 115	896
俄乌冲突	733	591
安倍遇刺	1 031	614
冬奥	462	552
新冠疫情	2 527	477
洪灾	971	802
乡村振兴	1 692	1 511
孟晚舟回国	304	683

*Purity*在0~1之间取值,越接近1表示聚类得到的结果越准确,具体计算公式如下:

$$Purity(Y,C)=\frac{1}{N}\sum_{k}\max_{j}|O_k\cap C_j|, \quad (17)$$

其中$Y=\{O_1,O_2,\cdots,O_k\}$表示的是经过计算划分出来的簇,$C=\{C_1,C_2,\cdots,C_j\}$表示真实簇的划分,N表示样本总数,k是计算得到事件簇的总个数,j是真实事件簇的总个数。

*ARI*和*RI*的计算公式如下:

$$RI=\frac{a+b}{C_N^2}, \quad (18)$$

$$ARI=\frac{RI-\mathrm{mean}(RI)}{\max(RI)-\mathrm{mean}(RI)}, \quad (19)$$

其中a表示真实的簇与经过计算得到的簇中相同的元素的对数,b表示真实的簇与经过计算得到的簇中不同的元素的对数。mean(•)是用来计算平均值。ARI在0~1之间取值,越接近1表示聚类得到的结果越准确。

*NMI*计算公式如下所示:

$$NMI(Y,C)=\frac{I(Y,C)}{[H(Y)+H(C)]/2}, \quad (20)$$

其中I表示互信息,计算公式如下所示:

$$I(Y,C)=\sum_{k}\sum_{j}\frac{|O_k\cap C_j|}{N}\log\frac{N|O_k\cap C_j|}{|O_k||C_j|}。 \quad (21)$$

H表示熵,计算公式如下所示:

$$H(Y)=-\sum_{i}\frac{O_i}{N}\log\frac{O_i}{N}, \quad (22)$$

*NMI*在0~1之间取值,越接近1表示聚类得到的结果越准确。

3.3 实验设置

本文模型在Pytorch框架上开发,使用Sentence Transformers库[18]中的distilbert-basenli-stsb-mean-token作为预训练编码器,最大输入长度为512,使用Adam优化器[19],激活函数为ReLU,编码时的学习率设置为2,实例头和簇群头的学习率为1×10^{-4}。本文在(0,1]范围内尝试不同的τ值,发现τ=0.5性能最优。

3.4 基线模型

为证明所提方法的有效性,选取以下6个基线模型进行对比分析:

Singlepass[5]:使用VSM(Vector Space Model)得到文本表征,使用余弦相似度计算相似度,最后通过增量聚类实现文本的聚类;

K-Means[4]:使用VSM得到文本表征,使用余弦相似度计算相似度,最后通过*K*-Means实现文本的聚类;

LSH(locality-sensetive hashing)[20]:使用VSM得到文本表征,使用余弦相似度计算相似度,最后通过哈希算法实现文本的聚类;

Hadifar[21]:采用SIF Embedding得到文本表征,通过深度聚类算法实现文本的聚类;

Wang[22]:使用LSTM得到文本表征,通过线性网络模型计算相似度,通过Singlepass算法进行聚类;

BERT:将Wang的模型换成BERT词向量,利用线性神经网络模型计算文本相似度,通过增量聚类算法进行聚类;

SemiEC[2]:使用LSTM得到文本表征,使用余弦相似度计算相似度,通过增量聚类实现文本的聚类;然后对模型再训练,不确定样本再聚类;

SCCL[3]:使用对比学习进行表示学习,并使用锐化的置信分布进行聚类学习。通过联合优化对比度损失和聚类损失来执行聚类。

3.5 实验结果分析

本文所提模型与基线模型在新闻事件数据集上的对比分析结果如表2所示。

如表2所示,通过与基线模型相比,NECC在新闻事件数据集上的性能优于所有的基线模型。与传统的Singlepass、K-Means、LSH、Hadifar相比,NECC有很大的提升,这是因为传统模型无法获取优质的表征,聚类效果有限。Wang采用LSTM表征文本后,聚类效果出现大

幅上升,之后的BERT更优于Wang的LSTM。SemiEC是对Wang模型的改进,其利用聚类产生的标注数据对模型再训练,结束后对不确定样本再聚类,提升了聚类准确性和聚类的纯度。SCCL提出引入对比学习获得更优表征,再次提高了聚类效果。而NECC的出现,缓解了SCCL中因迭代学习带来的误差累积,实现了更优的聚类。

表2 与基线模型的实验对比

Table 2 Experimental comparison with baseline models

模型	Purity	ARI	NMI
Singlepass	0.59	0.49	0.61
K-Means	0.76	0.61	0.73
LSH	0.58	0.47	0.43
Hadifar	0.59	0.54	0.52
Wang	0.73	0.72	0.66
BERT	0.75	0.75	0.69
SemiEC	0.78	0.71	0.75
SCCL	0.81	0.75	0.76
NECC	**0.84**	**0.76**	**0.79**

3.6 消融实验

3.6.1 不同数据增强方法效果对比

为了探究数据增强对聚类结果的影响,我们将NECC中基于Dropout的数据增强分别和不使用数据增强、使用同义词增强(使用WordNet同义词替换)、使用上下文增强(利用预训练语言模型找到输入文本的前n个合适的词进行插入或替换)、使用反译(将输入文本翻译成英语,然后再翻译回中文来生成输入文本的释义)等方法进行比较,对比结果如图4所示。

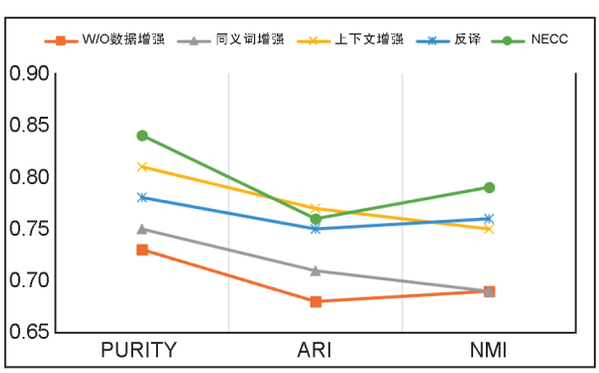

图4 不同数据增强方法效果对比图

Fig. 4 Comparison of the effects of different data augmentation methods

由图4可知,数据增强方法可以有效提升聚类效果,基于上下文增强的方法比传统同义词替换和反译效果更佳,这可能是由于上下文增强的方法可以得到文本的深层信息,最大限度地避免文本语义在增强过程中被改变。

3.6.2 融入不同事件知识效果对比

为了探究事件知识对聚类结果的影响,故将三种事件知识替换为不融入事件知识、只融入事件时间、只融入事件地点以及融入事件主题知识,其对比结果如图5所示。

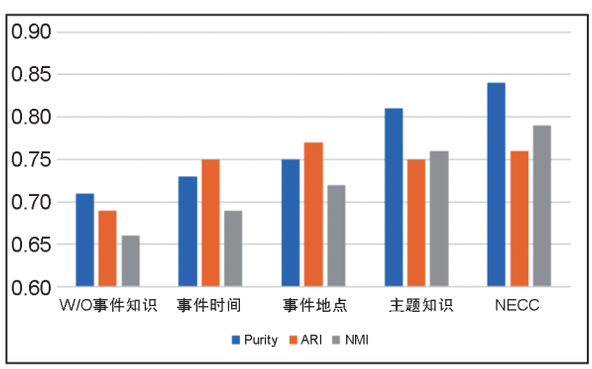

图5 融入不同事件知识效果对比图

Fig. 5 Comparison diagram of the effect of incorporating different event knowledge

由图5可知,NECC所提出的三种事件知识(新闻主题、时间、地点)具有约束作用。观察发现,只融入事件时间或事件地点效果不佳,原因是单一的时间或地点并不能代表某一事件。而事件主题知识比事件的时间或地点更能提升聚类效果,事件主题很大程度上已经代表某一事件,可以作为指导信息,区分主题不同的事件。但NECC使用主题信息+时间+地点的方法,更好地丰富事件表征,增强聚类效果。

3.6.3 不同对比学习和聚类的组合方式效果对比

为了探究不同对比学习和目标聚类组合的效果,我们对不使用对比学习直接聚类、顺序进行对比学习和目标聚类、迭代进行对比学习和目标聚类、联合进行对比学习和目标聚类(NECC)这几种组合方式进行实验,实验结果如图6所示。

由图6可知,对比学习可以有效提升聚类效果,原因是聚类开始前,不同类别的数据在表示空间中存在交叉重叠现象,对比学习拉近同源样本,拉远非同源样本,可以实现隐式聚

类。迭代进行对比学习和目标聚类效果优于顺序进行,原因可能是迭代学习有助于获得更深层的表征信息,辅助提升聚类效果。而NECC联合进行对比学习和目标聚类效果明显优于其他组合方式,原因可能是迭代学习不可避免地会导致误差累积,只能得到次优结果。

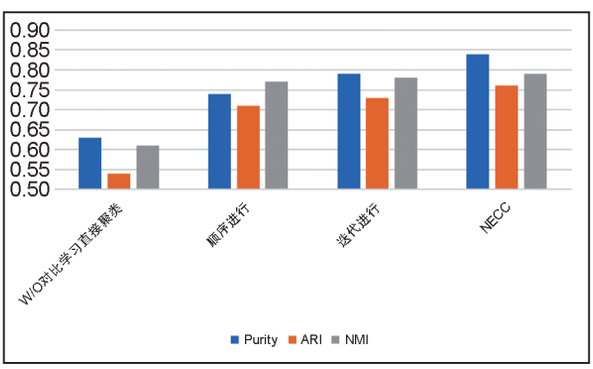

图6 不同对比学习和聚类组合方式效果对比图

Fig. 6 Effect comparison of different contrastive learning and clustering combination methods

4 结论

本文提出了一种融入事件知识的新闻事件对比聚类模型NECC用于实现实时新闻事件聚类,使用Dropout对原始文本进行数据增强,通过新闻主题、时间、地点三种事件知识丰富事件表征,基于“每个样本的聚类软标签可作为特征表示”的思想,将聚类任务统一到表示学习框架下,同时进行表示学习和目标聚类,减少误差累积。通过端到端模型实现对实时数据的处理,解决了新闻长文本存在的文本内容冗余、事件特征模糊、噪声干扰、语义复杂等问题,以及传统迭代表征聚类的方法易造成误差累积,并通过实验验证了所提方法的有效性。

参考文献:

[1] CORDEIRO M, GAMA J. Online Social Networks Event Detection: A Survey[M]//Solving Large Scale Learning Tasks. Challenges and Algorithms. Cham: Springer International Publishing. 2016: 1-41. DOI: 10.1007/978-3-319-41706-6_1.

[2] 郭恒睿, 王中卿, 朱巧明, 等. 基于半监督学习的中文社交文本事件聚类方法[J]. 中文信息学报, 2022, **36**(2): 152-159. DOI: 10.3969/j.issn.1003-0077.2022.02.017.
GUO H R, WANG Z Q, ZHU Q M, *et al*. Event Clustering Method for Chinese Social Text Based on Semi-supervised Learning[J]. *J Chin Inf Process*, 2022, **36**(2): 152-159. DOI: 10.3969/j.issn.1003-0077.2022.02.017.

[3] ZHANG D, NAN F, WEI X, *et al*. Supporting Clustering with Contrastive Learning[EB/OL]. arXiv Preprint: 2103.12953, 2021. https: //arXiv.org/abs/2103.12953.

[4] LIKAS A, VLASSIS N, VERBEEK J J. The Global K-means Clustering Algorithm[J]. *Pattern Recognit*, 2003, **36**(2): 451-461. DOI: 10.1016/s0031-3203(02)00060-2.

[5] PAPKA R, ALLAN J. On-line New Event Detection Using Single Pass Clustering[M]. Amherst: University of Massachusetts, 1998.

[6] HOFMANN T. Probabilistic Latent Semantic Analysis [EB/OL]. arXiv Preprint: 1301.6705, 2013. https: //arXiv.org/abs/1310.6705.

[7] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. *J mach Learn res*, 2003, **3**(Jan): 993-1022. DOI: 10.1162/jmlr.2003.3.4-5.993

[8] GAO T, YAO X, CHEN D. SimCSE: Simple Contrastive Learning of Sentence Embeddings[EB/OL]. arXiv Preprint: 2104.08821, 2021. https: //arXiv.org/abs/2104.08821.

[9] CHUANG Y S, DANGOVSKI R, LUO H, *et al*. DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings[EB/OL]. arXiv Preprint: 2204.10298, 2022.https: //arXiv.org/abs/2204.10298.

[10] FANG H C, XIE P T. An End-to-end Contrastive Self-supervised Learning Framework for Language Understanding[J]. *Trans Assoc Comput Linguist*, 2022, **10**: 1324-1340. DOI: 10.1162/tacl_a_00521.

[11] WU X, GAO C, ZANG L, *et al*. ESimCSE: Enhanced Sample Building Method for Contrastive Learning of Unsupervised Sentence Embedding[EB/OL]. arXiv Preprint: 2109.04380, 2021.https: //arXiv.org/abs/2109.04380.

[12] ZHANG D, XIAO W, ZHU H, *et al*. Virtual Augmentation Supported Contrastive Learning of Sentence Representations[EB/OL]. arXiv Preprint: 2110.08552, 2021.https: //arXiv.org/abs/2110.08552.

[13] JIA C, SHI Y F, YANG Q R, *et al*. Entity Enhanced BERT Pre-training for Chinese NER[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 6384-6396. DOI: 10.18653/v1/2020.emnlp-main.518.

[14] GROOTENDORST M. BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure[EB/OL]. arXiv Preprint: 2203.05794, 2022.https: //arXiv.org/abs/2203.05794.

[15] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-networks[EB/OL]. arXiv Preprint: 1908.10084, 2019. https: //arXiv.

org/abs/1908.10084.

[16] LI Y F, HU P, LIU Z T, *et al*. Contrastive Clustering[J]. *Proc AAAI Conf Artif Intell*, 2021, **35**(10): 8547−8555. DOI: 10.1609/aaai.v35i10.17037.

[17] WANG Y Z, TANG S X, ZHU F, *et al*. Revisiting the Transferability of Supervised Pretraining: An MLP Perspective[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 9173−9183. DOI: 10.1109/CVPR52688.2022.00897.

[18] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-networks[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP−IJCNLP). 2019: 3982−3992. DOI: 10.18653/v1/d19-1410.

[19] ZHANG Z J. Improved Adam Optimizer for Deep Neural Networks[C]//2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). 2019: 1−2. DOI: 10.1109/IWQoS.2018.8624183.

[20] PETROVIĆ S, OSBORNE M, LAVRENKO V. Streaming First Story Detection with Application to Twitter [C]//HLT '10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. New York: ACM, 2010: 181−189. DOI: 10.5555/1857999.1858020.

[21] HADIFAR A, STERCKX L, DEMEESTER T, *et al*. A Self-training Approach for Short Text Clustering[C]// Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP−2019). 2019: 194−199.

[22] WANG Z Q, ZHANG Y. A Neural Model for Joint Event Detection and Summarization[J]. *IJCAI Int Jt Conf Artif Intell*, 2017, **0**: 4158−4164.