基于多粒度语义间隔损失的半监督文本分类方法

廖国安[1],徐计[1,2],陈艳平[1,2]

(1.贵州大学公共大数据国家重点实验室,贵州 贵阳 550025;2.贵州大学计算机科学与技术学院,贵州 贵阳 550025)

摘 要: 半监督文本分类使用少量标注数据和大量未标注数据训练文本分类模型。然而,现有基于伪标签的方法存在由于决策边界欠拟合和伪标签偏差导致的误差累积问题。为此,本文提出一个基于多粒度语义间隔损失的交叉标注模型。首先通过基于 Transformer 的双向编码器表示(BERT)预训练模型提取包含上下文信息的向量序列,然后利用注意力网络和文本卷积神经网络(TextCNN)分别捕获全局语义特征和局部语义特征,并将这两部分特征融合形成新的语义特征,同时考虑不同类别之间的语义相似性,以在特征嵌入空间中将相似类别的样本分开,从而减轻误差累积的影响。最后,增加一个与类别相关的间隔,通过自适应局部阈值学习得到的伪标签样本与间隔进行正负样本损失计算,使每个类别样本在嵌入空间形成高密度分布,从而缓解决策边界的欠拟合。实验结果表明,该模型能够有效融合多粒度语义信息,生成高内聚伪标签样本,提升半监督文本分类的性能。

关键词: 文本分类;伪标签;间隔损失;特征提取;决策边界;半监督学习;自然语言处理

源代码链接: https://github.com/PandaA0520/SMLCL

中图分类号: TP391.1 **文献标志码:** A **DOI:** 10.19678/j.issn.1000-3428.0070533

Semi-Supervised Text Classification Method Based on Multi-Granularity Semantic Margin Loss

LIAO Guoan[1], XU Ji[1,2], CHEN Yanping[1,2]

(1.State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, Guizhou, China;
2.College of Computer Science and Technology, Guizhou University, Guiyang 550025, Guizhou, China)

【Abstract】 Semi-supervised text classification uses a limited set of labeled data in conjunction with a large corpus of unlabeled data to develop text classification models. Current pseudo-labeling techniques often face challenges such as pseudo-label bias stemming from decision boundary underfitting and error accumulation. To address these issues, a cross-annotation model based on the multi-granularity semantic margin loss is introduced. Initially, the Bidirectional Encoder Representations from Transformers (BERT)-pretrained model is employed to extract vector sequences enriched with contextual information. Subsequently, an attention mechanism and Text Convolutional Neural Network (TextCNN) are utilized to capture global and local semantic features, respectively. These features are subsequently fused to create enriched semantic representations, factoring in semantic similarities across different categories to separate similar classes within the feature-embedding space, thereby reducing the impact of error accumulation. Additionally, a category-specific margin to calculate the positive and negative sample loss between pseudo-labels derived through adaptive local threshold learning is introduced. The margin enables a high-density distribution of samples within each class in the embedding space, thereby mitigating decision boundary underfitting. Experimental evaluations demonstrate that the proposed model effectively integrates multi-granular semantic information, achieves high cohesion among pseudo-labeled samples, and significantly improves the performance of semi-supervised text classification.

【Key words】 text classification; pseudo-label; margin loss; feature extraction; decision boundary; semi-supervised learning; natural language processing

0 引言

传统文本分类[1-2]通过利用大量标注数据取得了很多成果。然而,在许多现实世界的场景中,获取大规模标注数据的成本很高,这使得使用这种方法训练有效的分类器变得很困难。半监督文本分类[3-5]是一种机器学习方法,它利用少量标注数据和大量未标注数据来训练模型。这种方法可以有效地

基金项目: 国家自然科学基金(62366008, 61966005)。

作者简介: 廖国安(CCF 学生会员),男,硕士研究生,主研方向为自然语言处理、粒计算、机器学习;徐计(通信作者),特聘教授、博士;陈艳平,特聘教授、博士、博士生导师。

收稿日期: 2024-10-24 **修回日期:** 2025-01-07 **E-mail:** jixu@gzu.edu.cn

降低手动标注的成本和时间，同时增强模型的泛化能力。在半监督学习中，常用的方法是通过模型对未标注数据进行初步预测，生成伪标签。这些伪标签用于进一步训练模型，以增强模型的表现。然而，伪标签可能包含错误，这些错误会被模型进一步放大，从而影响模型的准确性。NOVOTNEY 等[6]利用自动转录作为语言模型估计的先验已被证明可以显著提高预测模型的准确性，从而有效缓解因伪标签错误而带来的负面影响。

伪标签学习的挑战在于如何生成高质量的伪标签，避免引入错误信息或产生过拟合。同时，标注数据长尾分布也是一个问题，少数标签下包含大多数训练样本，而剩余标签下只包含少数样本。这种不平衡的分布会导致模型结果偏向那些频繁出现的标签，从而降低对稀有类别的识别能力，进而影响模型的整体泛化性能。近年来，伪标签方法[7-9]已经引起了广泛的关注，这些方法将伪标签分配给未标注的数据实例作为额外的监督信息，同时遵循低密度分离假设，在高密度特征空间中为未标注数据生成伪标签，帮助识别类别之间的边界。通过使用标注数据和伪标签数据的混合数据进行训练，模型获得了更丰富的监督信息，并减少了经验误差。然而，这些方法往往忽视了对数据进行进一步语义分析，并且常常面临模型偏向于较容易类别以及训练过程中误差累积的问题。

在实际模型的训练中，决策边界[10-11]的学习至关重要。特别是在伪标签学习中，如何有效地利用决策边界附近的伪标签数据成为一个关键问题。为了更好地构建模型，通常这些方法会学习一个良好的嵌入表示，使得同一类的样本聚集在一起，而不同类的样本彼此远离。此外，为了使学习到的特征具有足够的区分性，通常会在 Softmax 损失函数中加入一个间隔。但是在利用决策边界附近的伪标签数据时，需要更加谨慎地挑选伪标签样本，以避免引入额外的误差并影响模型的泛化能力。

鉴于上述情况，当前的半监督学习任务中生成伪标签需要解决两个主要问题：1）如何生成高质量的伪标签，以减少标签样本长尾分布的发生；2）如何降低误差累积的概率。产生这些问题的主要原因还是模型决策边界欠拟合，特征选择不足，输入的特征信息不够丰富，模型无法捕捉到数据中潜在的复杂模式，且数据本身可能存在噪声或不足以反映所有重要特征，导致决策边界过于简单。在机器学习任务中，原始数据通常是高维的，往往存在着隐含的复杂关系，因此本文对数据进行深层次的语义特征提取，旨在从数据中挖掘出更加丰富且具备语义信息的特征，帮助模型理解数据的上下文，进而提升其对数据的理解和分类能力。伪标签的质量依赖于模型对数据的理解深度，当模型能有效地从数据中提取深层次、多粒度的语义特征时，生成的伪标签会更准确，同时如果模型处于欠拟合状态，那么伪标签可能会被错误生成，进而产生误差累积，影响后续训练。

因此，本文提出了一个语义间隔损失交叉标注模型（SMLCL）以解决决策边界的欠拟合和伪标签误差累积两个问题。SMLCL 通过一个语义特征提取模块和一个间隔损失模块来提升半监督文本分类任务的性能。为了增强模型的泛化能力，采用了两种数据增强方法来预处理文本数据。第一种方法是强数据增强，该方法主要是回译，即将原始英文文本翻译成俄语，然后再翻译回英语，改变句子结构同时保持总体语义不变。第二种方法是弱数据增强，包括同义词替换、随机删除和随机插入。这两种方法不会显著改变句子结构，从而提高模型的泛化能力。在语义特征提取模块中，本文采用预训练的基于 Transformer 的双向编码器表示（BERT）[12]模型作为基础模型，然后添加一个注意力层和一个文本卷积神经网络（TextCNN）[13]层，以捕捉长文本间的依赖关系和局部语义特征，实现全局特征与局部特征的融合，提升后续伪标签质量，缓解误差累积影响。此外，本文添加了一个与类别相关的间隔，通过语义特征提取模块对经过自适应局部阈值后得到的伪标签与间隔进行正负样本损失计算，以更好地区分不同类别的样本，使同类样本高密度分布同时低耦合，从而缓解决策边界的欠拟合问题。除此之外，本文通过初始化两个具有不同参数的模型来生成伪标签，并让这两个模型以交叉标注的方式相互学习。通过这种交叉监督的方式，每个模型不仅可以利用自身生成的伪标签，还可以从另一个模型中获得额外的反馈，从而有效地减少单一模型可能带来的偏差和错误累积。

本文的主要贡献包括 4 个方面：

1）提出了一个复合语义特征提取模型，旨在从数据增强后的文本中提取长文本依赖关系和局部语义特征，从而实现全局特征与局部特征多粒度信息融合，增强模型的泛化能力。

2）提出了一个与类别相关的间隔损失函数，该函数根据样本类别之间的语义相似度，使用间隔值

与预测值之间的差值来区分不同类别的样本，并在指定区间内最大化正确类别与错误类别的预测分数之间的间隔。

3）采用交叉标注策略的语义间隔损失模型选择位于间隔附近的伪标签，以减轻伪标签误差累积的影响。

4）通过实验研究表明，本文提出的方法在 6 个常用英文基准数据集上结果优于目前的先进方法，并且通过综合分析证实了 SMLCL 的有效性。

1 相关工作

1）半监督文本分类。

FixMatch[14]使用从弱增强的未标注数据生成的伪标签来监督同一数据的强增强版本。SAT[15]通过训练一个元学习器，根据它们与原始数据的相似度重新排序增强方法，以改进 FixMatch。MixText[16]利用 Mixup[17]在潜在空间中对未标注和标注数据进行插值，避免了对有限标注数据的过拟合。CrisisMatch[18]研究了几种流行的半监督组件，并将 Mixup 与伪标签集成用于灾难推文分类。PGPL[19]提出了一种原型引导的半监督模型，该模型集成了原型锚定对比策略和原型引导的伪标签策略，以解决伪标签偏差问题。

FlexMatch[20]和 FreeMatch[21]都估计每个类别的学习状态，以自适应的方式调整局部阈值，缩小类别之间产生的伪标签数据数量差异。为了解决误差累积的问题，BLUM 等[22]提出了 Co-training 方法，该方法同时训练两个网络，并使用来自对方网络生成的伪标签来交叉监督自身网络。JointMatch[23]使用双网络架构来优化模型，通过交叉标注的方式进行训练，并根据每个类别在不同时间的学习情况动态设置阈值来适应训练，这有助于生成高质量伪标签。

2）语义分析模型。

注意力机制最初是由 MNIH 等[24]提出的，其模仿人类如何逐步关注视觉场景中的不同部分，使用类似于人眼中央凹的机制，其中视觉场中心的焦点比边缘更清晰。随后，VASWANI 等[25]引入了能够捕捉输入和输出之间全局依赖关系的注意力机制，使得模型能够并行处理数据并更有效地捕获长文本依赖。这种架构已成为自然语言处理领域广泛发展的基础，包括 BERT 和 GPT[26]等模型。HDP[27]结合多标签的语义背景，以降维的方式将文档映射成主题集的随机概率混合完成文本分类。LSET[28]基于上下文表示类别词去噪并采用距离监督自训练实现弱监督文本分类。

TextCNN[13]使用卷积层从文本数据中提取更高级别的特征，并捕捉局部语义特征。PAJO[29]探索了临床实践指南的自动文献筛选，它结合了文章标题和摘要编码的文本特征以及期刊特征。本文结合了这两种模型的优势，捕捉并结合长文本全局语义信息和局部特征来减少类别偏见。

3）间隔损失函数。

基于间隔的 Softmax 损失函数被广泛用于训练模型。ArcFace[30]采用角度间隔以稳定训练过程，进一步提高了人脸识别模型的区分能力。CosFace[31]采用附加间隔以最大化类间方差并最小化类内方差。AdaFace[32]则采用自适应间隔，根据样本的图像质量为不同难度的样本分配不同的重要性。这些方法的共同点是通过添加间隔损失计算来进一步增强性能。尽管上述间隔损失在视觉识别任务上取得了较好的结果，但它们并非为半监督文本分类任务设计。本文将间隔建模为提升伪标签质量的函数，建模后的函数会决定哪些样本在训练期间贡献更多的决策比重。为了最小化损失，本文方法中间隔驱动特征向量分别向正负方向发展，决策边界也根据正负样本的分布进行调整。

2 方法

2.1 总体模型构建

本文提出了一个半监督文本分类模型 SMLCL，结合语义特征提取模块和间隔损失模块，用以解决伪标签误差累积问题和决策边界的欠拟合情况。本节首先展示 SMLCL 模型的整体结构，其主要由 4 个模块组成：语义特征提取模块，自适应局部阈值模块，交叉标注模块，边界间隔损失模块。

SMLCL 的主要框架如图 1 所示（彩色效果见《计算机工程》官网 HTML 版，下同）。本文将原始数据分成两个部分：标注数据和无标注数据，针对每一批无标注数据和标注数据应用两种数据增强技术：弱数据增强和强数据增强。随后，将这些数据结合注意力层和 TextCNN 层的 BERT 模型来捕获语义特征。通过自适应局部阈值模块处理后，置信度超过自适应局部阈值的预测值被选为相对应的类别伪标签，再通过预设边界与真实值之间的间隔，使得预测值向真实标签方向靠近，远离错误标签。然后，两个模型通过交换它们生成的伪标签预测，交叉迭代地相互训练。

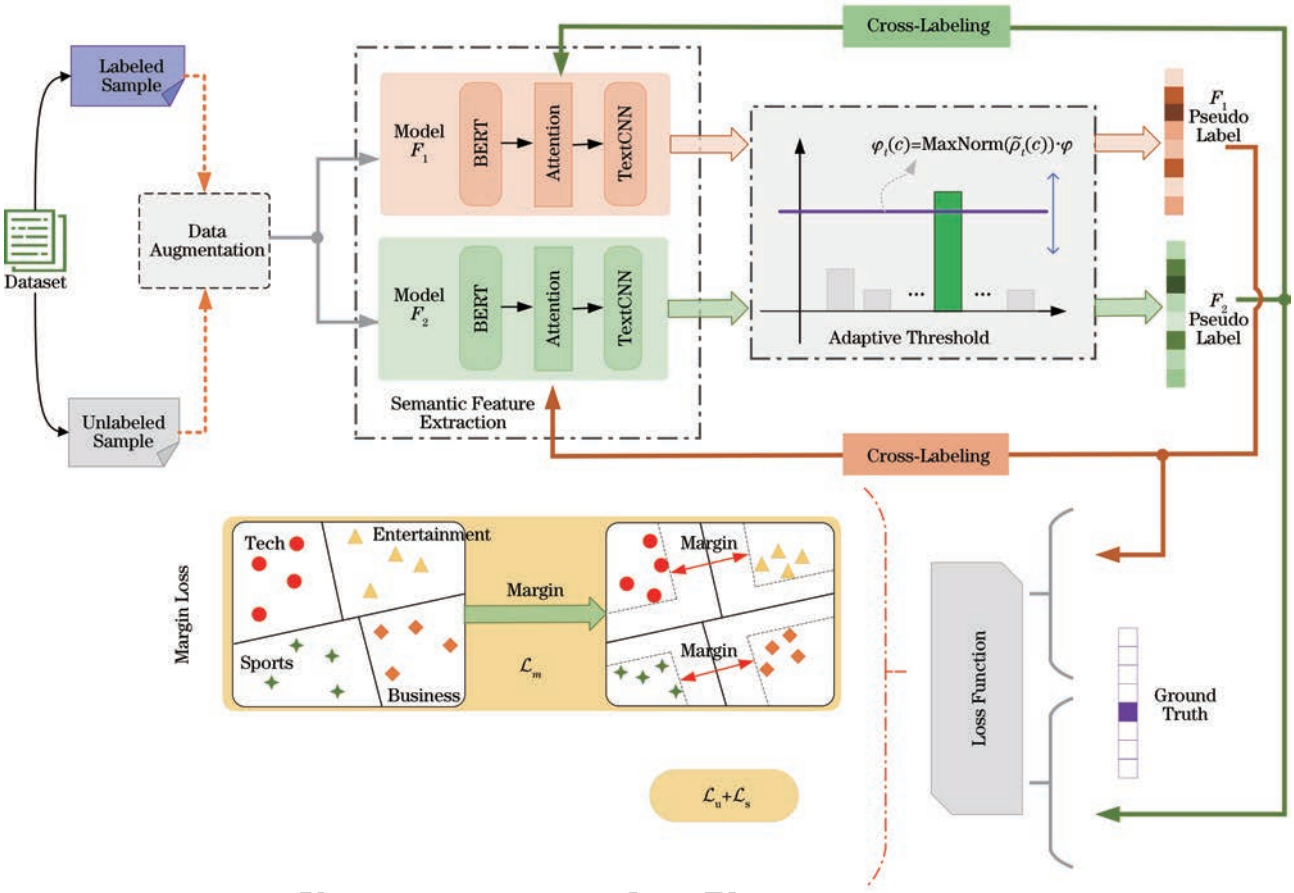


图 1 整体模型示意图

Fig. 1 Illustration of the overall model

2.2 语义特征提取模块

因为对于每个句子而言，每个词的重要程度不同，同时还要考虑句子之间上下文的语义关系，所以本文引入注意力机制来提取句子中重要单词的语义信息，捕捉粗粒度长文本语义特征，并再添加一层 TextCNN 以捕捉细粒度局部语义依赖，使得模型能够同时处理全局特征和局部特征，从多粒度的方向增强模型的表现力和泛化能力。

文本中，不同单词或短语对语义的贡献不同，注意力机制能够动态分配权重，更关注语义关键的部分，而忽略噪声或无关部分，并优先提取与目标相关的单词，例如后续实验部分的 GoEmotions 情感数据集中，注意力机制会更多关注表达情感的词（如“hate”“oh my god”等），而忽略中性词。

注意力层的主要公式如下：

$$\boldsymbol{u}_i=\tanh(\boldsymbol{W}_i\boldsymbol{h}_i+\boldsymbol{b}_i) \tag{1}$$

$$\boldsymbol{\alpha}_i=\frac{\exp(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_{\mathrm{w}})}{\sum_{i=0}^{L}\exp(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_{\mathrm{w}})} \tag{2}$$

$$\boldsymbol{h}_i=\sum_{i=0}^{L}\boldsymbol{a}_i\boldsymbol{h}_i \tag{3}$$

式中：$\boldsymbol{h}_i$ 为随机初始化的上下文向量，在训练的过程中更新；$\boldsymbol{u}_i$ 为隐藏层向量 $\boldsymbol{h}_i$ 经过一个全连接运算得到的结果；$\boldsymbol{u}_{\mathrm{w}}$ 表示一个可训练的上下文向量或注意力权重向量；$\boldsymbol{W}_i$ 和 $\boldsymbol{b}_i$ 分别为注意力计算的权值矩阵和偏置项；$\boldsymbol{\alpha}_i$ 为句子中第 i 个单词的注意力分数。

在本文中，注意力层为了适应 BERT 的输出，隐藏层数量设置为 768，用注意力权重矩阵乘以 BERT 模型输出的隐层状态矩阵，然后用 Softmax 计算得分并将其标准化为概率分布，最后这个分布用于计算 BERT 模型的最后一层隐藏层状态，得到 $\boldsymbol{O}_{\mathrm{att}}$，这是加权后的输出，包含了文本中所有单词的注意力加权表示：

$$\boldsymbol{O}_{\mathrm{att}}=\mathrm{Softmax}(\boldsymbol{\sigma W})\boldsymbol{o}(\boldsymbol{x}_b) \tag{4}$$

式中：$\boldsymbol{X}=\{(\boldsymbol{x}_b,\boldsymbol{y}_b):b\in(1,2,\cdots,B)\}$ 是标注样本数据批；$\boldsymbol{o}(\boldsymbol{x}_b)$ 为弱数据增强数据经过 BERT 的隐藏状态；$\boldsymbol{W}$ 为可学习的注意力权重矩阵；$\boldsymbol{O}_{\mathrm{att}}$ 为注意力层输出。

在 TextCNN 模型中，包含了多个卷积层，卷积核大小分别为 3、4、5，旨在捕捉不同跨度的 n-gram 特征，即分别捕捉 3、4、5 个字符之间的关系，滤波器的大小设置为 128。卷积核大小决定了所提取的局部 n-gram 特征的范围（即感受野），例如，卷积核大

小为 3 时对应捕捉 3-gram 特征(如连续的 3 个单词)。由于本文所使用的英文数据集为短文本数据集,句子长度通常为 10～30 个单词,关键信息往往包含在 n-gram 中,如三元组(主谓宾)或短语(如“not very good”)。假设数据集中的句子长度服从正态分布,均值为 15 个单词,标准差为 5。对于句子长度为 L 的文本,卷积核大小为 k 的卷积层产生的有效感受野为 $L-k+1$,而句子的覆盖率则为 k/L。因此,对于卷积核大小分别为 3、4、5 的情况,针对均值为 15 个单词的句子,其有效感受野和覆盖率分别为 13、12、11 以及 20%、27%、33%。这种设置能够同时覆盖局部信息,并捕捉不同范围的上下文特征,从多个粒度角度有效提取文本的关键信息,避免单一卷积核带来的信息丢失。

在前向传播函数中,首先将输入 $\boldsymbol{O}_{\text{att}}$ 进行转置以匹配卷积层的输入格式,然后对每个卷积层应用 ReLU 激活函数和最大池化操作,并将所有卷积层的输出拼接在一起:

$$\boldsymbol{O}^{i}_{\text{conv}}=\text{ReLU}(\boldsymbol{f}_i(\boldsymbol{O}^{\text{T}}_{\text{att}})) \tag{5}$$

$$\boldsymbol{O}^{i}_{\text{pool}}=\text{MaxPool}(\boldsymbol{O}^{i}_{\text{conv}}) \tag{6}$$

$$\boldsymbol{O}_{\text{cnn}}=\text{Concat}(\boldsymbol{O}^{i}_{\text{pool}}) \tag{7}$$

式中:$\boldsymbol{O}^{i}_{\text{pool}}$ 表示池化层输出;$\boldsymbol{O}^{i}_{\text{conv}}$ 表示第 i 个卷积操作的输出结果;$\boldsymbol{f}_i$ 为卷积核,i 是卷积核索引;$\boldsymbol{O}^{\text{T}}_{\text{att}}$ 为 $\boldsymbol{O}_{\text{att}}$ 的转置;$\boldsymbol{O}_{\text{cnn}}$ 表示所有池化层拼接后得到的 TextCNN 层输出。将式(5)～式(7)整合为式(8):

$$\boldsymbol{O}_{\text{cnn}}=\text{Concat}(\text{MaxPool}(\text{ReLU}(\boldsymbol{f}_i*\boldsymbol{O}_{\text{att}}))) \tag{8}$$

式中:$*$ 表示卷积操作。在此模块中,本文利用注意力机制动态分配权重,以捕捉粗粒度的全局语义信息和长距离依赖,从而增强模型的解释性并过滤无关噪声,然后通过 TextCNN 提取 n-gram 特征,在短文本场景中有效捕捉细粒度局部上下文关系,并与注意力机制互补,实现了多粒度的特征融合,提升了语义特征提取的准确性。

2.3 自适应阈值模块

很多半监督文本分类的伪标签算法都通过固定阈值选取高置信度的无标注数据作为伪标签样本进行训练[33-34]。但这忽略了在不同时间步下学习不同类的难度,可能导致容易类比困难类产生更多的伪标签用于学习,并且这种情况会导致模型后续训练不断增加偏差而使得性能降低。本文不使用固定阈值,而是估计不同类的学习状态,并自适应地调整不同类的局部阈值,动态调整各个类别生成伪标签的能力,减小伪标签长尾分布的影响。

在每个时间步 t,每个类学习状态为 $\tilde{\boldsymbol{p}}_t=[\tilde{p}_t(1),\tilde{p}_t(2),\cdots,\tilde{p}_t(C)]$,通过模型在未标注数据上的预测概率的指数移动平均(EMA)来估计分类学习状态:

$$\tilde{p}_t(c)=\begin{cases}\dfrac{1}{C}, & t=0\\ \lambda\tilde{p}_{t-1}(c)+(1-\lambda)\dfrac{1}{\mu B}\displaystyle\sum_{b=1}^{\mu B}q_b, & \text{otherwise}\end{cases} \tag{9}$$

式中:$\boldsymbol{U}=\{\boldsymbol{u}_b:b\in(1,2,\cdots,\mu B)\}$ 是无标注样本数据批;C 是类别数;B 是标注数据的批大小;μ 是未标注数据与样本数据的比值;$\lambda\in(0,1)$ 是 EMA 衰减系数。

接着对估计的学习状态进行归一化处理,并从初始阈值 φ 中调整每个类 C 的局部阈值:

$$\varphi_t(c)=\text{MaxNorm}(\tilde{p}_t(c))\cdot\varphi=\frac{\tilde{p}_t(c)}{\max\{\tilde{p}_t(c):c\in[C]\}}\cdot\varphi \tag{10}$$

式中:φ 是初始阈值。MaxNorm 是归一化方法,具体定义如下:在时间步 t 下,每个类别的学习状态权重通过除以当前所有类别中的最大值来进行归一化。通过采用这一技术,在当前学习状态中难以生成高置信度伪标签的类别将具有较低的局部阈值,从而允许生成更多的伪标签,将这些伪标签用于后续训练过程。

2.4 交叉标注模块

伪标签的另一个关键问题是误差累积:如果生成的预测是不正确的,模型会产生越来越多的噪声伪标签,并继续在错误样本上训练模型,模型会变得越来越差。因此,本文给定两个不同的初始化网络 F_1 和 F_2,每个网络使用另一个网络的伪标签作为自己下一轮迭代使用的数据。通过两个模型交叉验证和标注,可以提高最终模型对新数据的适应能力,而且有助于减少单一模型偏见的影响。然后,将这些伪标签用于计算未标注数据损失 L_{u},以更新其对等网络的参数:

$$P_b^{F_2}=\left(\max\left(q_b^{F_2}\right)\geqslant\varphi_t^{F_2}\left(\underset{b}{\text{argmax}}\left(q_b^{F_2}\right)\right)\right) \tag{11}$$

$$P_b^{F_1}=\left(\max\left(q_b^{F_1}\right)\geqslant\varphi_t^{F_1}\left(\underset{b}{\text{argmax}}\left(q_b^{F_1}\right)\right)\right) \tag{12}$$

$$L_{\text{u}}^{F_1}=\frac{1}{\mu B}\sum_{b=1}^{\mu B}\mathbb{I}\left(P_b^{F_2}\right)\mathbb{Z}\left(\hat{q}_b^{F_2},Q_b^{F_1}\right) \tag{13}$$

$$L_{\text{u}}^{F_2}=\frac{1}{\mu B}\sum_{b=1}^{\mu B}\mathbb{I}\left(P_b^{F_1}\right)\mathbb{Z}\left(\hat{q}_b^{F_1},Q_b^{F_2}\right) \tag{14}$$

式中:P_b 代表模型在置信度超过阈值时才生成伪标签;q_b 是弱增强未标注数据预测类别分布;$\hat{q}_b$ 是由

q_b 转换而来的独热(one-hot)标签；Q_b 是强增强未标注数据预测类别分布；$\mathbb{Z}$ 是交叉熵损失；$\mathbb{I}(\cdot)$是指示函数。该方法有助于避免单个网络内的直接误差积累。

2.5 间隔损失模块

间隔损失模块是模型的核心部分。在半监督学习中，尤其在使用伪标签方法时，决策边界的准确性和鲁棒性对于模型的性能至关重要。如果决策边界没有被正确定位，模型可能会在不同类别之间混淆，从而导致伪标签质量下降并引发误分类的连锁反应。因此，调整合适的决策边界以确保其能够有效地区分不同类别的样本显得尤为重要。通过在不同类别的预测伪标签之间加入一个边界与预测值之间的间隔，以学习一个判别性嵌入空间，有助于增加之前模块中正负伪标签样本之间的外部距离，从而更易区分不同类别的样本，如图 2 所示。

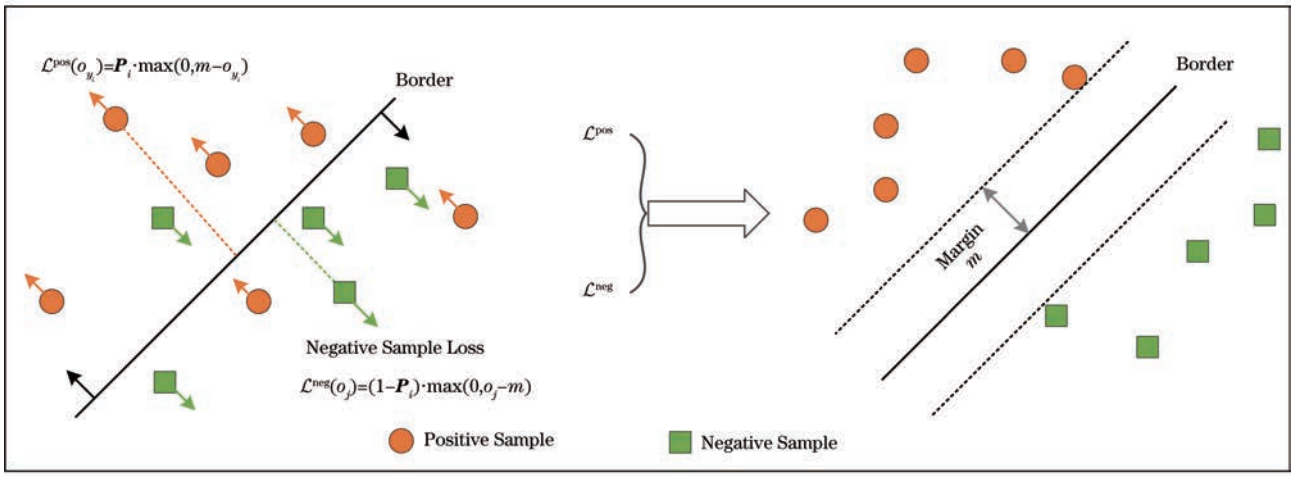


图 2 间隔损失示意图

Fig. 2 Illustration of the margin loss

图 2 展示了损失函数如何通过调整真实类别和非真实类别样本相对于预定义边界之间的间隔的输出来改善模型的决策过程。为了最小化损失，边界驱使正负样本沿着边界的不同方向移动，决策边界也会根据样本的分布进行调整。分类正确的样本将在角度上接近于真实类别权重向量 $\boldsymbol{W}_{y_i}$，分类错误的样本则会接近于 $\boldsymbol{W}_j$。为了实现这一点，本文提出了一种间隔损失函数，其公式为：

$$\boldsymbol{L}_m=\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j\neq y_i}\left[\mathrm{pos}(o_{y_i})+\mathrm{neg}(o_j)\right]\right) \tag{15}$$

式(15)主要由两个组成部分构成：正样本损失和负样本损失。对于间隔值 m，本文设置为 0.5，这是因为对于每个样本 i，前面模块的伪标签预测输出值 o_{y_i} 应至少大于间隔值。如果 o_{y_i} 小于间隔值，那么 $\max(0,m-o_{y_i})$将产生一个正损失值，从而鼓励模型增加 o_{y_i} 的值。相反地，对于所有非真实类别预测值 o_j，其值应小于间隔值。如果 o_j 大于间隔值，那么 $\max(0,o_j-m)$也将产生一个正损失值，从而促使模型减少 o_j 的值。具体正样本损失和负样本损失的计算如下：

$$\mathrm{pos}(o_{y_i})=\boldsymbol{P}_i\cdot\max(0,m-o_{y_i}) \tag{16}$$

$$\mathrm{neg}(o_j)=(1-\boldsymbol{P}_i)\cdot\max(0,o_j-m) \tag{17}$$

式中：$\boldsymbol{P}_i$ 表示样本 i 真实标签的 one-hot 编码向量；N 是样本的总数；m 是用于定义类别之间最小距离的间隔值。

除了计算间隔损失以外，在更新无监督损失时还需要考虑两个网络在初始训练阶段具备不同的学习能力，然而随着训练的进行，这些网络会逐渐收敛并退化为自训练，再次受到误差累积的影响。因此，本文为每个未标注数据 $\boldsymbol{u}_b$ 计算数据一致性损失权重 w_b 为更多的不一致数据进行加权，以保持两个网络的发散：

$$w_b=\eta\,\mathbb{I}\left(\hat{q}_b^{F_1}=\hat{q}_b^{F_2}\right)+(1-\eta)\mathbb{I}\left(\hat{q}_b^{F_1}\neq\hat{q}_b^{F_2}\right) \tag{18}$$

式中：数据一致性权重 $\eta\in(0,0.5)$。因此，最终无标注数据损失函数定义如下所示：

$$L_{\mathrm{u}}^{F_1}=\frac{1}{\mu B}\sum_{b=1}^{\mu B}w_b\,\mathbb{I}\left(P_b^{F_2}\right)\mathbb{Z}\left(\hat{q}_b^{F_2},Q_b^{F_1}\right) \tag{19}$$

$$L_{\mathrm{u}}^{F_2}=\frac{1}{\mu B}\sum_{b=1}^{\mu B}w_b\,\mathbb{I}\left(P_b^{F_1}\right)\mathbb{Z}\left(\hat{q}_b^{F_1},Q_b^{F_2}\right) \tag{20}$$

式中：L_{u} 是无监督损失，在计算了两个网络的无标注数据损失之后，将其与标注数据的损失相加，以获得最终的损失函数。对于有监督数据的损失函数，定义如下：

$$L_{\mathrm{s}}=\frac{1}{B}\sum_{b=1}^{B}\mathbb{Z}(y_b,o(x_b)) \tag{21}$$

到目前为止，已经计算了间隔损失 L_m、无监督

损失 L_u 和有监督损失 L_s，将这三个部分结合在一起得到最终的全局损失函数：

$$L=L_s+w_uL_u+L_m \tag{22}$$

2.6 算法描述

SMLCL 整体算法如算法 1 所述。首先，将原始数据进行数据增强，输入复合语义模型中进行特征提取，通过注意力层获取粗粒度全局特征，接着通过 TextCNN 层获取细粒度局部特征，然后将这两者得到的特征融合以获取初步的嵌入表示。接下来，对每个类不同时间步下的学习状态进行估计，根据这些估计动态调整置信度阈值进行伪标签的筛选，而后根据伪标签预测值与真实值的差异计算更新无监督损失时的权重，同时计算正负伪标签样本相较于间隔的损失，最后返回总体的损失函数，两个模型交叉标注学习迭代更新。这种语义特征提取计算间隔损失交叉标注的机制使模型能够保留更多的细节信息，并且减小伪标签误差累积的影响和决策边界的欠拟合。

算法 1 SMLCL 算法

输入 模型 F_1，F_2，类别数 C，无标注样本 $\boldsymbol{U}=\{\boldsymbol{u}_b:b\in(1,2,\cdots,\mu B)\}$，标注样本 $\boldsymbol{X}=\{(x_b,y_b):b\in(1,2,\cdots,B)\}$

输出 模型总损失

1. For model F_1，F_2 do：
2. 计算文本单词注意力加权表示 O_{att}；//式(4)
3. 计算对 O_{att} 进行卷积池化后得到的 O_{cnn}；//式(8)
4. 计算标注数据的损失；//式(21)
5. 估计不同类别的学习状态；//式(9)
6. For c=1 to C do：
7. 根据每个类的学习状态调整局部阈值；//式(10)
8. End For
9. End For
10. For b=1 to μB do：
11. 为预测标签不一致的数据计算损失权重；//式(18)
12. End For
13. 计算间隔损失；//式(15)
14. 计算两个网络无标注数据的损失；//式(19)、式(20)

3 实验与结果分析

本文实验使用的显卡为 A100，显存大小为 80 GB，Python 版本为 3.8，PyTorch 版本为 1.9。评价指标选择准确率和 F1 值(Macro-F1)两项。准确率是分类问题中最直观的性能指标，它定义为正确预测的数量占总预测数量的比例；Macro-F1 考虑了类别不平衡的情况，计算精确率和召回率的平均值。

3.1 数据集

为了全面准确地评估实验结果，选用了 6 个标准的英文半监督文本分类基准数据集评估 SMLCL，分别是 AGNews[35]、Yahoo! Answer[36](后文简称为 Yahoo)、IMDB[37]、Empathetic Dialogues[38](后文简称为 Empath)、GoEmotions[39](后文简称为 GoEmo)、Hurricane[40](后文简称为 Hu)。AGNews 由新闻文章构成；Yahoo 包含多种类型的用户生成内容，样本包括问题和相应最佳回答；IMDB 由电影评论构成；Empath 是一个专门用于训练具备同理心对话能力模型的数据集；GoEmo 主要由 Reddit 上收集的英文对话片段组成，是一个大型情感分类数据集，专用于识别文本中的细粒度情感；Hu 是一个专用于研究和分析灾害期间社交媒体数据的情感反应的数据集。表 1 显示了 6 个数据集的基本信息。

表 1 数据集基本信息

Table 1 Information statistics of datasets

数据集	数据类别	类别数/个	数据总数/条	数据集划分
AGNews	新闻主题	4	35 600	5 000/2 000/1 900
Yahoo	问答主题	10	130 000	5 000/2 000/6 000
IMDB	电影评论	2	37 000	5 000/1 000/12 500
Empath	对话主题	32	24 850	610/2 770/2 547
GoEmo	Reddits	27	29 425	870/2 956/2 984
Hu	推特帖子	8	12 800	1 280/1 280/1 280

为了公平对比其他方法，在原始测试集和训练集中使用随机采样来构建本文的无标注样本训练集和验证集，同时在所有基准数据集中均采用相同的数据增强技术。对于回译，先将文本翻译成俄语，然后再翻译回英语；对于同义词替换，用 WordNet 做同义词随机替换 30%的词。3 个数据集的划分以标注类别为基本单位，每个数据集的无标注样本数均为 5 000。以 AGNews 为例，AGNews 的验证集和测试集分别包含 8 000、7 600 条样本，由于 AGNews 有 world、sports、business、sci/tech 4 个标签类别，因此对于每个标签类别，验证集和测试集的大小为 2 000、1 900，见

表 1 中“数据集划分”一列，其他数据集同理。

3.2 超参数设置

在模型训练过程中，采用 Adam 优化器进行优化。为了进一步验证模型的稳定性和可靠性，本实验对模型进行了 5 次独立评估，并计算了平均准确率和标准差。这种多次评估的方法能够有效减小因数据随机性或模型初始化差异带来的评估偏差。为了保持与对比的基线方法所设置的超参数一致，本文实验对于不同的数据集进行不同的参数设置，具体超参数值如表 2 所示。

表 2 模型参数

Table 2 Model parameters

数据集	学习率/10^{-5}	批大小	无监督权重	EMA	初始阈值	间隔	标签数量/条
AGNews	2	8	1.00	0.9	0.98	0.5	10
Yahoo	2	8	1.00	0.9	0.98	0.5	20
IMDB	1	4	0.05	0.9	0.98	0.5	10
Empath	1	4	1.00	0.9	0.98	0.5	10
GoEmo	1	4	1.00	0.9	0.98	0.5	10
Hu	1	4	1.00	0.9	0.98	0.5	10

3.3 基线方法

本文选取了近几年的方法：BERT[12]，UDA[4]，MixText[16]，FixMatch[14]，SAT[15]，FreeMatch[21]，SoftMatch[41]，以及目前最先进的方法 JointMatch[23]作为实验对比方法，以进一步说明 SMLCL 的有效性。BERT 使用基于 Transformer 的双向编码器来理解上下文进行文本分类；UDA 利用未标注数据，通过数据增强来提升模型的泛化能力；MixText 在 Mixup 基础上进一步考虑了文本的结构和语义，以生成更具挑战性的训练样本，从而帮助模型学习更复杂的决策边界；FixMatch 将一致性正则化和伪标签结合实现文本分类；SAT 通过动态调整每个样本的权重来减少噪声标签的影响，从而使模型能够在复杂或受干扰的数据集上更加有效地学习；FreeMatch 提出一种新的半监督学习方法，核心思想是利用自适应阈值机制和一致性正则化，解决传统半监督学习中固定阈值导致的不灵活问题，从而提高模型在未标注数据上的利用效率；SoftMatch 通过软标签和标签平滑技术来改善半监督学习的性能，允许模型探索更多可能的标签分布，从而提高其泛化能力和准确性；JointMatch 结合多种技术处理半监督文本分类，包括数据增强、伪标签更新、优化标注数据的监督损失、交叉标注等，从而达到高准确率。

3.4 实验结果

表 3 总结了 SMLCL 在不同文本分类数据集上与各基准模型的比较结果，其中粗体表示在数据集上的最佳性能，而性能排名第二的数据则使用下划线进行标注，下同。对于标签数量为 10 的 AG News 来说，SMLCL 与最佳基准 JointMatch 相比准确率和 F1 值分别提高了 4.15%和 4.20%；在标注数量为 20 的 Yahoo 上分别提升 6.14%和 5.93%；在标注数量为 10 的 IMDB 上则分别提升 15.73%和 15.90%。JointMatch 虽然使用了自适应阈值和交叉标注，但忽略了文本信息的语义相似度和伪标签之间的类别差异，本文通过复合语义特征提取来评估每个词和句子上下文的重要性，并利用这些特征与真实值之间的间隔区分不同类别样本，这种策略使得模型在不同粒度层次上捕捉和保留了文本关键信息，在情感分析数据集 IMDB 上的巨大性能提升表明了本文提出的模型在语义处理上具有优越性。

表 3 在 AGNews、Yahoo、IMDB 数据集上的准确率对比

Table 3 Accuracy comparison on AGNews, Yahoo, and IMDB datasets %

模型	AGNews(C=4)		Yahoo(C=10)		IMDB(C=2)	
	准确率	F1 值	准确率	F1 值	准确率	F1 值
BERT	69.18±3.70	68.27±3.50	58.11±1.60	57.38±1.90	63.16±1.40	62.93±1.60
UDA	76.69±3.20	76.51±3.00	59.32±2.00	58.47±2.30	64.88±1.70	64.57±1.50
MixText	78.07±2.80	77.23±3.50	59.93±1.90	59.24±1.80	65.22±1.10	65.78±1.20
FixMatch	80.22±2.40	78.98±2.10	60.17±1.70	59.86±1.50	64.52±1.60	64.31±1.40
SAT	85.43±1.20	85.30±1.50	61.33±1.50	60.96±1.40	68.96±1.70	68.92±1.60
JointMatch	<u>87.68±0.50</u>	<u>87.64±0.50</u>	<u>66.58±0.70</u>	<u>66.09±0.80</u>	<u>76.91±4.50</u>	<u>76.79±4.60</u>
SMLCL(Ours)	**91.32±0.30**	**91.32±0.31**	**70.67±1.01**	**70.01±1.18**	**89.01±0.98**	**89.00±0.98**
相对 JointMatch 提升	4.15	4.20	6.14	5.93	15.73	15.90

为了对不同数量标注数据的实验结果进行详细比较，在 AG News 和 Yahoo 数据集上进行了实验，标注数据的数量各不相同。这些实验表明 SMLCL 相比于 FixMatch 和 JointMatch 在准确性上的提升。如表 4 和表 5 所示，SMLCL 在这两个数据集上的所有标注数据数量条件下都稳定地优于 JointMatch 和 FixMatch。在标签数量较少的情况下，如在标签数量为 5 的情况下，SMLCL 的准确率能够比标签数量为 100 的 JointMatch 达到更高的水平。在 Yahoo 数据集上，在标签数量为 5、10、15、25、100 的设置下，SMLCL 相对于 JointMatch 准确率分别提升了 8.58%、7.71%、5.69%、6.90%、3.83%，在 AGnews 数据集上则分别提升 4.79%、3.90%、3.91%、3.46%、3.56%。实验结果表明，SMLCL 不仅相比 JointMatch 达到了更高的准确率，而且在有限的标注数据情况下模型准确率也更加稳定。

表 4 Yahoo 数据集不同标签数量下的准确率

Table 4 Accuracy under different numbers of labeled data on Yahoo dataset %

模型	标签数量为 5	标签数量为 10	标签数量为 15	标签数量为 25	标签数量为 100
FixMatch	56.13	59.81	65.07	63.85	68.07
JointMatch	64.08	65.52	67.51	67.24	69.74
SMLCL(Ours)	**69.58**	**70.57**	**71.35**	**71.88**	**72.41**
相对 JointMatch 提升	8.58	7.71	5.69	6.90	3.83

表 5 AGNews 数据集不同标注数量下的准确率

Table 5 Accuracy under different numbers of labeled data on AGNews dataset %

标签数量	标签数量为 5	标签数量为 10	标签数量为 15	标签数量为 25	标签数量为 100
FixMatch	70.29	80.25	82.32	85.96	87.79
JointMatch	86.00	88.18	88.33	88.50	89.07
SMLCL(Ours)	**90.12**	**91.62**	**91.78**	**91.56**	**92.24**
相对 JointMatch 提升	4.79	3.90	3.91	3.46	3.56

图 3 为模型在 AGNews 和 Yahoo 两个数据集不同标注数量下的准确率可视化表现，如图 3 所示，SMLCL 曲线更加平稳，这是因为 SMLCL 通过调整边界与伪标签预测值之间的间隔解决了决策边界欠拟合的问题，将伪标签有效区分开，减小了伪标签误差带来的误差累积。同时实验结果表明，SMLCL 可以更高效地利用少量标注数据，凸显了其在大量无标注数据情况下具有更大的潜力。

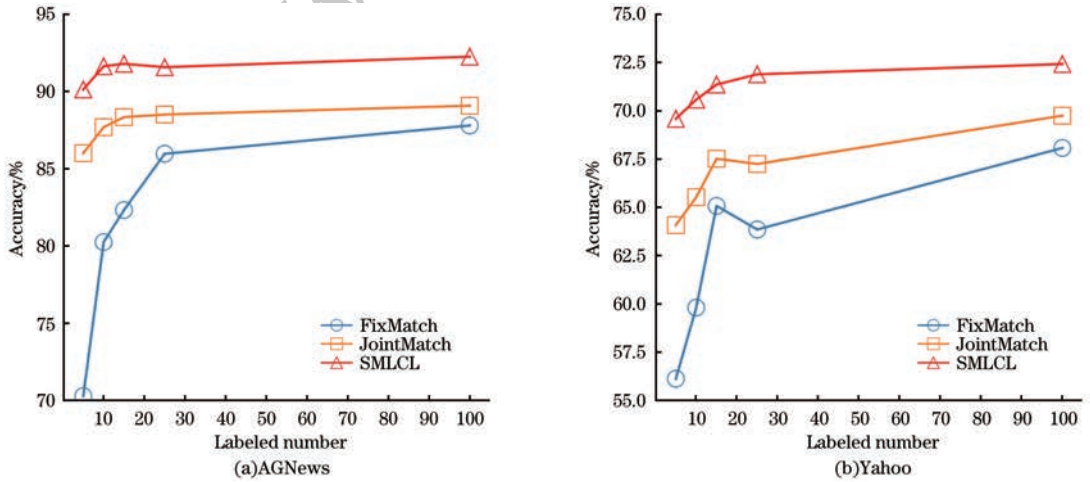


图 3 不同标注数量下的准确率对比

Fig. 3 Accuracy comparison under different numbers of labels

本文也在另外 3 个具有更多标签类别的数据集 Empath、GoEmo、Hu 上进行实验分析，以展示 SMLCL 的泛化能力。如表 6 所示，标注样本数均设为 10，可见 SMLCL 相比所有基准方法性能均有提升。这 3 个数据集样本大多带有情感色彩描述，之前的方法如 FreeMatch 核心特点是自适应阈值机制，动态调整不同类别伪标签置信度阈值，解决传统固定阈值对不同类别不适配的问题；SoftMatch

核心特点是利用软标签(概率分布)代替硬标签,使模型探索更多可能的标签分布,这两者都只是针对伪标签的某一细节问题优化,缺乏对整体架构的探索,因此最终结果平均差异不是很大。而JointMatch将两者优点结合在一起,构建了自适应阈值模块与软标签模块,性能提升相对较大。本文在这些方法的基础上加入多粒度语义嵌入,捕捉复杂情感特征,再利用所提出的针对不同类别之间的伪标签正负样本间隔有效地分割开样本,使得样本分布高内聚、低耦合,相对JointMatch准确率分别提升了24.00%、39.19%、6.33%。

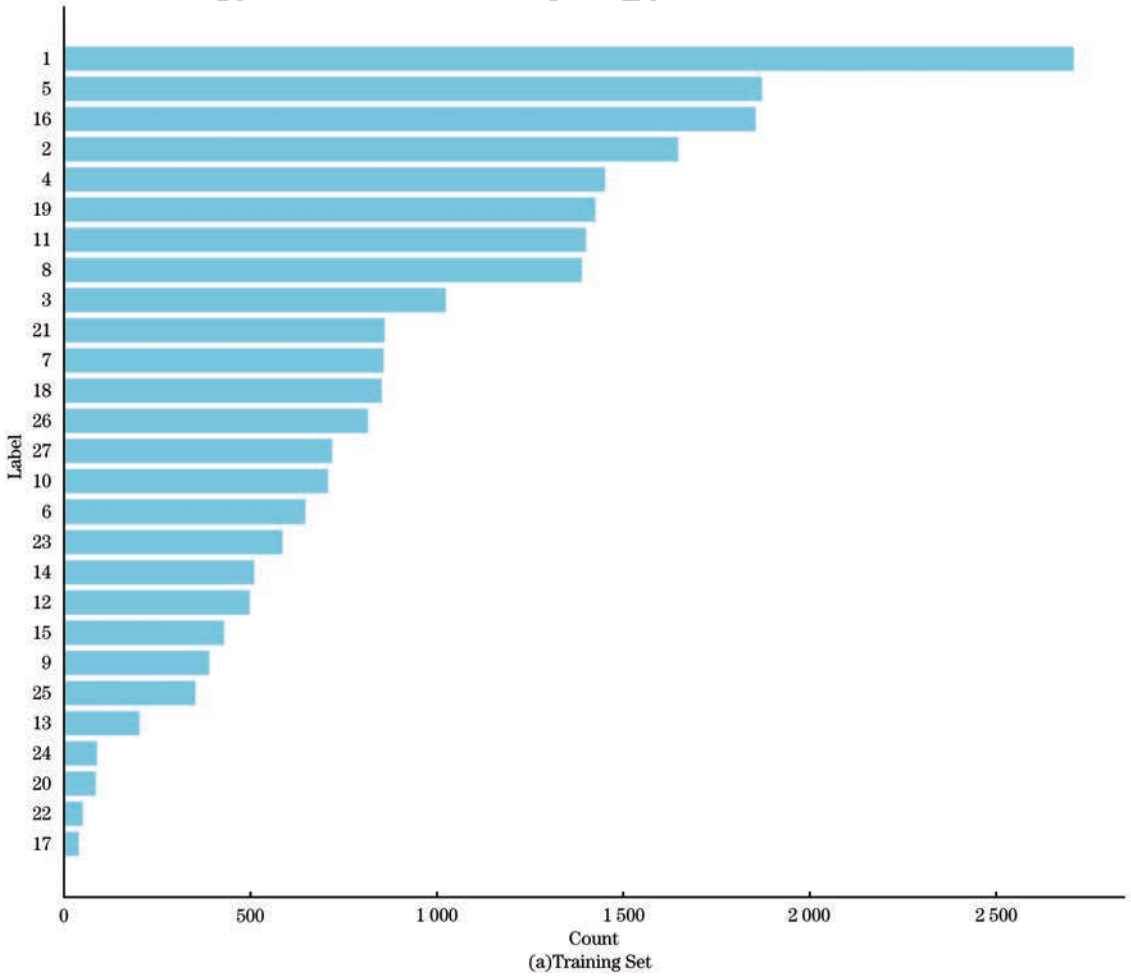
表 6 在多类别数据集上的准确率对比

Table 6 Accuracy comparison on multi-class datasets %

数据集	Empath	GoEmo	Hu
BERT	21.63	28.86	71.72
FixMatch	24.40	30.06	73.44
SAT	29.25	30.36	74.06
SoftMatch	29.96	31.84	74.38
FreeMatch	30.15	32.74	77.03
JointMatch	34.67	38.40	78.98
SMLCL(Ours)	**42.99**	**53.45**	**83.98**
相对 JointMatch 提升	24.00	39.19	6.33

为了验证本文模型在长尾数据集上有效,进一步分析在GoEmo数据集上的结果。如图4(a)所示,GoEmo情感分类数据集共有27个类别,前8个头部类占据大多数样本,而最后尾部类占据很少样本,如24、20、22、17几类样本数量均少于100条。图4(b)为模型对于各类别预测准确率的热力图,可以看出整体所有分类颜色高亮点都保持在对角线上,存在一些亮点在对角线外区域,如17、24等,但核心仍位于对角线,证明本文模型能够有效捕捉长尾数据集稀有类别的关键语义特征信息,并通过添加边界间隔提升伪标签的识别准确率。

通过图5可视化比较可以清晰地观察到,相对于之前的方法,本文方法在每个数据集上性能都有显著提升,3个数据集主要用于情感分析,说明SMLCL对于理解和分类文本中的复杂情感具有优势,其通过语义特征提取模块后的数据确保了文本的核心信息在池化过程中得以完整保留,随后通过自适应局部阈值和类别相关间隔两次筛选后得到的伪标签更具有代表性和鲁棒性,为后续训练更新提供了坚实的基础。



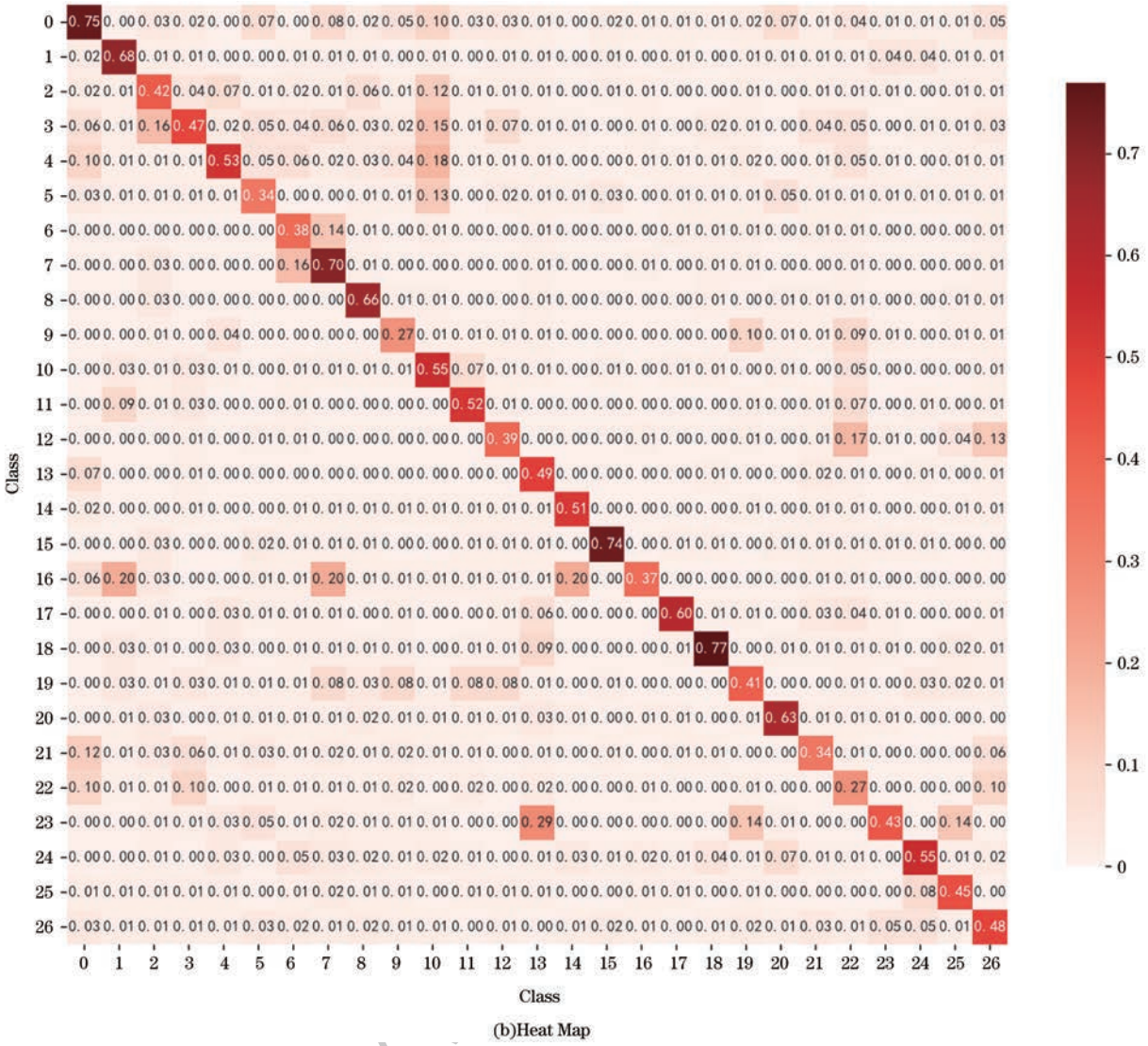


图 4 GoEmo 训练集及准确率热力图

Fig. 4 GoEmo training set and accuracy heat map

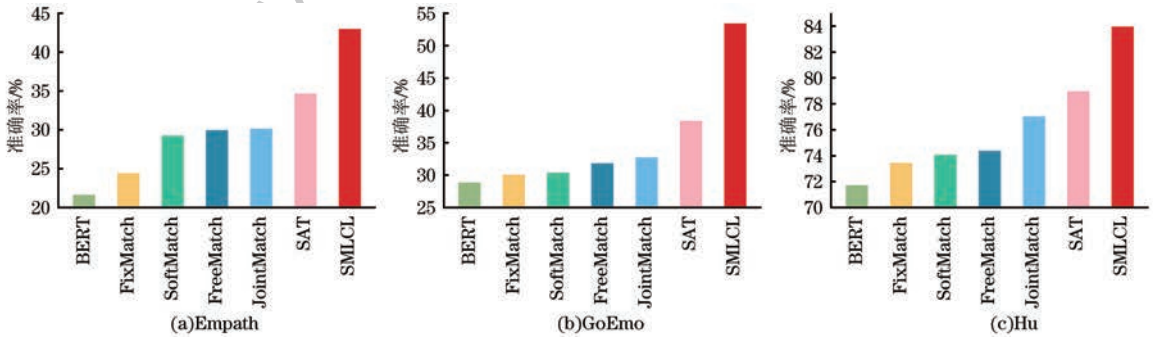


图 5 多类别数据集准确率示意图

Fig. 5 Illustration of accuracy among multi-class datasets

3.5 消融实验

为了说明 SMLCL 中各组成部分的有效性，本文还研究了去掉不同模块后对 SMLCL 的准确率会产生什么影响。分别在标注样本数设置为 10、25、20 的 AGNews、Yahoo 和 IMDB 数据集上进行以下消融实验：移除语义特征提取模块，移除间隔损失模块、移除语义特征提取模块、间隔损失模块，各子模块去除后的准确率下降情况如表 7 所示。

总体来看，依次剥离各个组件及全部组件后，准确率均有所下降，这表明 SMLCL 中的所有组件对最终性能都有贡献。在 3 个数据集上去除间隔损失模块后，SMLCL的性能下降最显著。间隔损失模

表 7 消融实验结果

Table 7 Ablation experimental results %

方法	AGNews	Yahoo	IMDB
SMLCL	**91.62**	**71.68**	**89.99**
−语义特征提取	88.40	68.80	88.11
−间隔损失	87.89	67.43	87.87
−语义特征提取、间隔损失	88.39	67.24	82.86

块用于分离伪标签中各类别之间的距离，使得正负样本差异明显，从而缓解决策边界附近欠拟合问题，去掉该模块会丢失类别差异信息，导致模型性能下降。语义特征提取模块用于提取长文本的全局语义特征和文本中的局部特征，能够有效地将这两部分特征结合在一起，从而增强针对不同跨度文本的语义理解。去掉语义特征提取模块后，模型就不能利用不同粒度下的语义信息，影响了模块的判断和选择。这些语义信息可以帮助模型排除一些不符合条件的负样本，例如一些不带感情色彩的中性词，同时在 3 个数据集中，IMDB 的准确率变化最大，笔者推测这是因为该数据集主要由评论文本组成。语义特征的捕捉使模型具有更强的适应性，因此在 IMDB 数据集上体现出较大的改进。

3.6 不同间隔值对性能的影响分析

SMLCL 使用语义间隔来区分不同类别之间的伪标签样本，通过学习一个判别性嵌入空间，有助于增大之前模块中获得的正负伪标签样本之间的外部距离，从而使正负样本高密度分布，在选择间隔损失中的参数 m 时，本文综合考虑了多个因素，以确保最佳性能和模型鲁棒性。首先，理论上，间隔 m 的选择旨在有效反映分类决策边界，以保证正样本的伪标签预测值至少大于 m，而负样本则应低于该值。此外，本文还考虑了数据的分布特性，尤其是在类别不均衡的情况下。在这一过程中，本文对模型的鲁棒性进行了测试，通过加入噪声或扰动输入，确认所选间隔能够有效区分不同类别样本。因此，本文探究这两个因素共同作用下不同间隔值对模型性能的影响。针对输入噪声，选择随机改变部分样本的标签，同时随机删除某些单词，实验选择数据集为 IMDB，标注样本数设置为 10，除间隔值外其余超参数均和表 2 一致，最终实验结果如表 8 所示。

实验结果表明，选择间隔值为 0.5 时，虽然召回率较间隔值为 0.7 的情况略低一些，但模型在各项指标上综合表现最佳，当间隔值较低，如 0.1、0.3 时，模型表现不是很好，显示出较小的间隔值可能不足以建立良好的分类边界。当间隔值增加到 0.9 时，模型表现最差，因为更大的间隔使得决策边界过于稀疏，不能准确保证样本高密度聚类。因此，合适的间隔值选择不仅能提升模型在测试集上的准确性，也能增强模型在更广阔范围内的泛化能力。

表 8 不同间隔值的性能对比

Table 8 Performance comparison of different margins %

间隔值	准确率	召回率	F1 值
0.1	87.33	86.89	87.34
0.3	88.23	87.46	88.34
0.5	89.01	88.76	89.00
0.7	88.94	88.91	88.96
0.9	86.65	86.13	86.74

3.7 不同初始化对性能的影响分析

本文模型初始化都是采用高斯噪声初始化，因为它能够更好地平衡梯度传播，尤其是当网络使用非线性激活函数(如 ReLU)时。为了探索其他噪声优化方法及加权一致性损失对模型的性能影响，本文采用高斯噪声初始化作为对比方法，如 Kaiming Uniform 初始化方法，通过设置不同学习率和加权一致性损失的消融实验，验证本文模型损失函数的有效性。实验选择 IMDB 数据集，标注样本数设置为 10，除上述参数设置外，其余超参数均和表 2 一致，最终实验结果如表 9 所示。

从实验结果可以看到，加权一致性损失会带来一定程度的性能提升。例如，在使用高斯噪声初始化并设置学习率为 1×10^{-3} 时，加入加权一致性损失后，模型的准确率从 78.98%提升到 80.37%，同时还可看出学习率是影响模型准确率最大的因素，其余相同条件下，将学习率从 1×10^{-5} 调整到 1×10^{-3}，准确率从 89.01%降到 80.37%。相比之下，当使用均匀噪声初始化时，虽然效果较差，但依然可以通过加权一致性损失改善性能，表明该损失函数的潜力。

表 9 不同初始化设置的性能对比

Table 9 Performance comparison of different initialization settings %

初始化方法	学习率	加权一致性损失	准确率	F1 值
高斯噪声初始化	1×10^{-5}	无	86.43	86.22
高斯噪声初始化	1×10^{-5}	有	89.01	89.00
高斯噪声初始化	1×10^{-4}	无	85.66	85.54
高斯噪声初始化	1×10^{-4}	有	87.28	87.21
均匀噪声初始化	1×10^{-5}	无	83.38	83.39
均匀噪声初始化	1×10^{-5}	有	84.72	84.68
均匀噪声初始化	1×10^{-4}	无	81.03	81.24
均匀噪声初始化	1×10^{-4}	有	82.31	82.33
高斯噪声初始化	1×10^{-3}	无	78.98	79.13
高斯噪声初始化	1×10^{-3}	有	80.37	80.36

4 结束语

本文针对现有半监督文本分类方法中未充分考虑语义特征信息及决策边界的欠拟合，导致伪标签质量良莠不齐和误差累积问题，提出了一种基于语义间隔损失交叉标注的半监督学习模型 SMLCL，用于半监督文本分类任务。该模型应用了语义特征提取分析和与类别相关的边界间隔损失策略。具体来说，语义特征提取从多粒度角度捕捉长文本上下文依赖和局部语义特征并将两者结合，缓解了伪标签误差累积的问题。此外，与类别相关的间隔将附近可靠伪标签数据推向更大的正负样本间距离，并利用交叉标注的方式训练模型，使得同类别样本达到高密度分布，从而缓解决策边界附近的欠拟合问题。实验结果表明，本文模型相较先进方法性能显著提升。

本文模型虽然在性能上取得了一定的提升，但仍有完善空间，下一步研究方向主要是探索模型在中文数据集上的表现，针对这一跨语言场景，主要挑战有以下三点：一是分词处理，英文文本以空格作为单词分隔符，因此分词只需基于空格和标点符号进行切割，但中文文本缺乏显式词边界，单个字符既可为词，也可为词的一部分，如“南京市/长江大桥”可能会被误分为“南京/市长/江大桥”；二是数据处理，数据增强及后续训练缺乏适配的中文词向量模型和预训练模型；三是计算资源，较之英文处理，中文处理更加复杂，计算嵌入时会耗费更多资源，增加耗时。未来的工作会继续跟进这些问题，旨在进一步验证本文模型的鲁棒性和泛化性。

参考文献

[1] 张铭泉，周辉，曹锦纲. 基于注意力机制的双 BERT 有向情感文本分类研究[J]. 智能系统学报，2022，17(6)：1220-1227.
ZHANG M Q，ZHOU H，CAO J G. Dual BERT directed sentiment text classification based on attention mechanism [J]. CAAI Transactions on Intelligent Systems，2022，17(6)：1220-1227.（in Chinese）

[2] 古丽娜孜·艾力木江，乎西旦·居马洪，孙铁利，等. 基于支持向量的最近邻文本分类方法[J]. 智能系统学报，2018，13(5)：799-807.
Gulnaz Alimjan，Hurxida Jumahun，SUN T L，et al. The nearest neighbor text classification method based on support vector[J]. CAAI Transactions on Intelligent Systems，2018，13(5)：799-807.（in Chinese）

[3] SOSEA T，CARAGEA C. Leveraging training dynamics and self-training for text classification[C]//Proceedings of the Findings of the Association for Computational Linguistics：EMNLP 2022. Abu Dhabi，United Arab Emirates：Association for Computational Linguistics，2022：4750-4762.

[4] XIE Q Z，DAI Z H，HOVY E，et al. Unsupervised data augmentation for consistency training[J]. Advances in Neural Information Processing Systems，2020，33：6256-6268.

[5] 武红鑫，韩萌，陈志强，等. 监督和半监督学习下的多标签分类综述[J]. 计算机科学，2022，49(8)：12-25.
WU H X，HAN M，CHEN Z Q，et al. Survey of multi-label classification based on supervised and semi-supervised learning[J]. Computer Science，2022，49(8)：12-25.（in Chinese）

[6] NOVOTNEY S，SCHWARTZ R，KHUDANPUR S. Getting more from automatic transcripts for semi-supervised language modeling[J]. Computer Speech & Language，2016，36：93-109.

[7] LEE J H，KO S K，HAN Y S. SALNet：semi-supervised few-shot text classification with attention-based lexicon construction[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S. l.]：AAAI，2021：13189-13197.

[8] SAJJADI M，JAVANMARDI M，TASDIZEN T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning[EB/OL]. [2024-09-24]. https://arxiv.org/abs/1606.04586.

[9] YIN J，LIU X Y，YANG Z W. A deep multiple-instance text binary classification for topic relevant content extraction on social media[J]. Journal of King Saud University-Computer and Information Sciences，2024，36(1)：101883.

[10] GIDARIS S，KOMODAKIS N. Dynamic few-shot visual learning without forgetting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA：IEEE Press，2018：4367-4375.

[11] HUANG Y G，WANG Y H，TAI Y，et al. CurricularFace：adaptive curriculum learning loss for deep face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle，USA：IEEE Press，2020：5900-5909.

[12] DEVLIN J，CHANG M W，LEE K，et al. BERT：pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2024-09-24]. https://arxiv.org/abs/1810.04805.

[13] LAI S W，XU L H，LIU K，et al. Recurrent convolutional neural networks for text classification[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. [S. l.]：AAAI，2015：2267-2273.

[14] SOHN K，BERTHELOT D，LI C L，et al. FixMatch：simplifying semi-supervised learning with consistency and confidence[EB/OL]. [2024-09-24]. https://arxiv.org/abs/2001.07685.

[15] CHEN H，HAN W，PORIA S. SAT：improving semi-supervised text classification with simple instance-adaptive self-training[C]//Proceedings of the Findings of the Association for Computational Linguistics：EMNLP 2022. Abu Dhabi，United Arab Emirates：Association for Computational Linguistics，2022：6141-6146.

[16] CHEN J A，YANG Z C，YANG D Y. MixText：linguistically-informed interpolation of hidden space for semi-supervised text classification[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online：Association for Computational Linguistics，2020：2147-2157.

[17] ZHANG H Y，CISSE M，DAUPHIN Y N，et al. Mixup：beyond empirical risk minimization[EB/OL]. [2024-09-24]. https://arxiv.org/abs/1710.09412.

[18] ZOU H P，ZHOU Y，CARAGEA C，et al. CrisisMatch：semi-supervised few-shot learning for fine-grained disaster tweet classification[EB/OL]. [2024-09-24]. https://arxiv.org/abs/2310.14627.

[19] YANG W Y，ZHANG R C，CHEN J F，et al. Prototype-guided pseudo labeling for semi-supervised text classification[C]//

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, 2023: 16369-16382.
[20] ZHANG B W, WANG Y D, HOU W X, et al. FlexMatch: boosting semi-supervised learning with curriculum pseudo labeling[EB/OL]. [2024-09-24]. https://arxiv.org/abs/2110.08263.
[21] WANG Y D, CHEN H, HENG Q, et al. FreeMatch: self-adaptive thresholding for semi-supervised learning[EB/OL]. [2024-09-24]. https://arxiv.org/abs/2205.07246.
[22] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the 11th Annual Conference on Computational Learning Theory. New York, USA: ACM Press, 1998: 92-100.
[23] ZOU H P, CARAGEA C. JointMatch: a unified approach for diverse and collaborative pseudo-labeling to semi-supervised text classification[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023: 7290-7301.
[24] MNIH V, HEESS N M O, GRAVES A, et al. Recurrent models of visual attention[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2014: 2204-2212.
[25] VASWANI A. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2017: 6000-6010.
[26] FLORIDI L, CHIRIATTI M. GPT-3: its nature, scope, limits, and consequences[J]. Minds and Machines, 2020, 30(4): 681-694.
[27] 李永忠, 郑滔. 基于标签的半监督 HDP 文本分类主题模型[J]. 模式识别与人工智能, 2017, 30(12): 1138-1148.
LI Y Z, ZHENG T. Semi-supervised labeled hierarchical dirichlet process topic model for document categorization[J]. Pattern Recognition and Artificial Intelligence, 2017, 30(12): 1138-1148. (in Chinese)
[28] 林呈宇, 王雷, 薛聪. 标签语义增强的弱监督文本分类模型[J]. 计算机应用, 2023, 43(2): 335-342.
LIN C Y, WANG L, XUE C. Weakly-supervised text classification with label semantic enhancement[J]. Journal of Computer Applications, 2023, 43(2): 335-342. (in Chinese)
[29] LIN Y C, LI J, XIAO H, et al. Automatic literature screening using the PAJO deep-learning model for clinical practice guidelines [J]. BMC Medical Informatics and Decision Making, 2023, 23(1): 247.
[30] DENG J K, GUO J, XUE N N, et al. ArcFace: additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE Press, 2020: 4685-4694.
[31] WANG H, WANG Y T, ZHOU Z, et al. CosFace: large margin cosine loss for deep face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE Press, 2018: 5265-5274.
[32] KIM M, JAIN A K, LIU X M. AdaFace: quality adaptive margin for face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE Press, 2022: 18729-18738.
[33] HOSSEINI M, CARAGEA C. Semi-supervised domain adaptation for emotion-related tasks[C]//Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics, 2023: 5402-5410.
[34] XU H M, LIU L Q, ABBASNEJAD E. Progressive class semantic matching for semi-supervised text classification[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA: Association for Computational Linguistics, 2022: 3003-3013.
[35] ZHANG X, ZHAO J B, LECUN Y. Character-level convolutional networks for text classification [EB/OL]. [2024-09-24]. https://arxiv.org/abs/1509.01626.
[36] CHANG M W, RATINOV L A, ROTH D, et al. Importance of semantic representation: dataless classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S. l.]: AAAI, 2008: 830-835.
[37] MAAS A L, DALY R E, PHAM P T, et al. Learning word vectors for sentiment analysis[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, USA: Association for Computational Linguistics, 2011: 142-150.
[38] RASHKIN H, SMITH E M, LI M, et al. Towards empathetic open-domain conversation models: a new benchmark and dataset[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 5370-5381.
[39] DEMSZKY D, MOVSHOVITZ-ATTIAS D, KO J, et al. GoEmotions: a dataset of fine-grained emotions [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 4040-4054.
[40] ALAM F, QAZI U, IMRAN M, et al. HumAID: human-annotated disaster incidents data from Twitter with deep learning benchmarks [EB/OL]. [2024-09-24]. https://arxiv.org/abs/2104.03090.
[41] CHEN H, TAO R, FAN Y, et al. SoftMatch: addressing the quantity-quality trade-off in semi-supervised learning[EB/OL]. [2024-09-24]. https://arxiv.org/abs/2301.10921.

文字编辑 金胡考
栏目编辑 赖玉玲