基于 u-wordMixup 的半监督深度学习模型

唐焕玲[1,3,4†]，宋双梅[2]，刘孝炎[1]，窦全胜[1,3,4]，鲁明羽[5]

(1. 山东工商学院 计算机科学与技术学院，山东 烟台 264005；
2. 山东工商学院 信息与电子工程学院，山东 烟台 264005；
3. 山东省高等学校协同创新中心: 未来智能计算，山东 烟台 264005；
4. 山东省高校智能信息处理重点实验室 (山东工商学 院)，山东 烟台 264005；
5. 大连海事大学 信息科学技术学院，辽宁 大连 116026)

摘 要: 当标注样本匮乏时，半监督学习利用大量未标注样本来解决标注瓶颈的问题，但由于未标注样本和标注样本来自不同领域，可能造成未标注样本存在质量问题，使得模型的泛化能力变差，导致分类精度下降。为此，本文基于 wordMixup 方法，提出了针对未标注样本进行数据增强的 u-wordMixup 方法，并结合一致性训练框架和 Mean Teacher 模型，提出了一种基于 u-wordMixup 的半监督深度学习模型 (Semi-supervised Deep learning model based on u-wordMixup,SD-uwM)。该模型利用 u-wordMixup 方法对未标注样本进行数据增强，在有监督交叉熵和无监督一致性损失的约束下，能够提高未标注样本质量，减少过度拟合。在 AGNews、THUCNews 和 20 Newsgroups 数据集上的对比实验结果表明，所提方法能够提高模型的泛化能力，同时也有效提高了时间性能。

关键词: 半监督学习；数据增强；深度学习；文本分类

中图分类号: TP181　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.1789　　开放科学 (资源服务) 标识码 (**OSID**):

Semi-supervised Deep learning model based on u-wordMixup

TANG Huanling[1,3,4†]，*SONG Shuangmei*[2]，*LIU Xiaoyan*[1]，*Dou Quansheng*[1,3,4]，*LU Mingyu*[5]

(1.School of Computer Science and Technology, Shandong Technology and Business University, Yantai, Shandong 264005, China；2. School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai, Shandong 264005, China；3.Co-innovation Center of Shandong Colleges and Universities: Future Intelligent Computing, Yantai, Shandong 264005, China；4.Key Laboratory of Intelligent Information Processing in Universities of Shandong (Shandong Technology and Business University), Yantai, Shandong 264005, China；5.Information Science and Technology College, Dalian Maritime University, Dalian, Liaoning 116026, China)

Abstract: When labeled data are deficient, semi-supervised learning uses a large number of unlabeled data to solve the bottleneck problem of labeled data. However, the unlabeled data and labeled data come from different fields, it may cause quality problem of unlabeled data, which makes the generalization ability of the model poor and leads to the degradation of classification accuracy. Therefore, based on the wordMixup method, this paper proposes the u-wordMixup method for data augmentation of unlabeled data, and a semi-supervised deep learning model based on u-wordMixup (SD-uwM) by combining the consistent training framework and Mean Teacher model. The model utilizes the u-wordMixup method to augment data of unlabeled data, which can improve the quality of unlabeled data and reduce overfitting under the constraints of supervised cross-entropy and unsupervised consistency loss. The comparative experimental results on the datasets of AGNews, THUCNews and 20 Newsgroups show that the proposed method can improve the generalization ability of the model and also effectively improve the time performance.

Keywords: semi-supervised learning;data augmentation;deep learning;text categorization

0 引 言

传统的监督学习[1-4] 需要大量的标注样本，代价昂贵，存在标注瓶颈问题。半监督学习[5-6] 是有效解决该问题的方法之一，典型方法有 Co-training[7] 和 Tri-training[8] 等。但由于未标注样本和标注样本来自不同领域，可能导致半监督模型精度下降。近年来，深度学习得到广泛关注，虽然已有很多半监督深度学习模型，但大多用于图像分类，用于文本分

收稿日期: 2021-10-18；**录用日期:** 2022-03-15.
基金项目: 国家自然科学基金项目 (61976124, 61976125, 62176140, 61873177, 61972235, 82001775).
†通讯作者. E-mail: th101@163.com.

类的研究不多，特别是关于未标注文本的增强研究。

数据增强有采用噪声输入的简单方法[9-10]，如对数据添加高斯噪声、对抗性噪声；也有根据样本的特征进行增强变换的高级增强方法，如同义词替换[11-13] 和词向量替换[14] 等。简单数据增强对提高模型的泛化能力不理想，最近几年高级数据增强方法的研究逐渐增多。但增强的样本需提供额外的归纳偏差才更有效。因此，如何选择数据增强至关重要。Mixup 方法[15] 在图像分类中表现优异，结合文本特征的 wordMixup(词混合) 方法[16] 应运而生，尽管取得了优异的结果，但是在有标注样本的情况下，模型泛化能力提高空间较小。

本文通过一致性训练框架[17]，基于 wordMixup 数据增强方法，提出了一种针对未标注样本的 u-wordMixup(未标注样本词混合) 数据增强方法，在有监督交叉熵损失和无监督一致性损失的约束下，增强生成高质量的附加训练样本，减少过度拟合。

基于 u-wordMixup 方法，利用 Mean Teacher 模型[18]，进行一致性训练，提出了一种新的半监督深度学习模型 (SD-uwM)。(1) u-wordMixup 数据增强方法对未标注样本进行增强，以降低无监督一致性损失为目标，约束增强未标注训练样本质量，减少模型的过度拟合；(2) 目标损失函数结合有监督交叉熵损失和无监督一致性损失，利用 MeanTeacher 方法进行一致性训练，提高模型的泛化能力。

1 wordMixup 数据增强

wordMixup 是一种针对有标注样本进行数据增强的方法，其思想是对两个样本的词嵌入向量进行插值，生成新的样本词嵌入矩阵，作为增强的样本。

给定一对有标注样本 (x_i, y_i) 和 (x_j, y_j)，进行词嵌入 (word embedding) 得到 $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ 和 $\left(\boldsymbol{x}_j, \boldsymbol{y}_j\right)$，其中，$\boldsymbol{x}_i \in R^{N\times d}$，$\boldsymbol{x}_j \in R^{N\times d}$ 是文本 x_i 和 x_j 的词嵌入矩阵，N 表示单词数，d 表示词向量维度，$\boldsymbol{y}_i$ 和 $\boldsymbol{y}_j$ 是相应的类标签。然后按照式 (1) 和 (2) 进行插值 (interpolation)，得到新样本 $\left(\tilde{\boldsymbol{x}}_{ij}, \tilde{\boldsymbol{y}}_{ij}\right)$，$\tilde{\boldsymbol{x}}_{ij}$ 为增强样本的词嵌入矩阵，$\tilde{\boldsymbol{y}}_{ij}$ 为其类标签。

$$\tilde{\boldsymbol{x}}_{ij}^{k} = \lambda \boldsymbol{x}_i^k + (1-\lambda)\, \boldsymbol{x}_j^k, k = 1..N. \tag{1}$$

$$\tilde{\boldsymbol{y}}_{ij} = \lambda \boldsymbol{y}_i + (1-\lambda)\, \boldsymbol{y}_j. \tag{2}$$

其中，$\boldsymbol{x}_i^k$ 和 $\boldsymbol{x}_j^k$ 分别表示文本 x_i 和 x_j 中第 k 个单词的词向量，$\lambda \in [0,1]$ 为插值权重因子，$\tilde{\boldsymbol{x}}_{ij}^k$ 为插值生成新样本的第 k 个词的词向量。对 $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ 和 $\left(\boldsymbol{x}_j, \boldsymbol{y}_j\right)$ 中的每个词一一对应进行词向量插值，得到新样本的嵌入矩阵 $\tilde{\boldsymbol{x}}_{ij}$，$\tilde{\boldsymbol{y}}_{ij}$ 为 $\tilde{\boldsymbol{x}}_{ij}$ 的类标签，$(\tilde{x}_{ij}, \tilde{y}_{ij})$ 即增强后的附加训练样本。

wordMixup 方法在有监督文本分类中取得了较好的效果。但半监督学习中的未标注没有标签，如何为其插值生成伪标签？为此，本文在 wordMixup 的基础上，针对未标注样本提出一种改进的数据增强方法 u-wordMixup，将在 2.2 节重点介绍。

2 SD-uwM 半监督深度学习模型

2.1 SD-uwM 模型

半监督深度学习模型 (SD-uwM) 的框架如图 2 所示，其利用 Mean Teacher 模型思想构造教师模型 T 与学生模型 S，采用有标注样本和未标注样本，基于有监督交叉熵损失和无监督的一致性损失的目标函数，进行半监督深度学习。

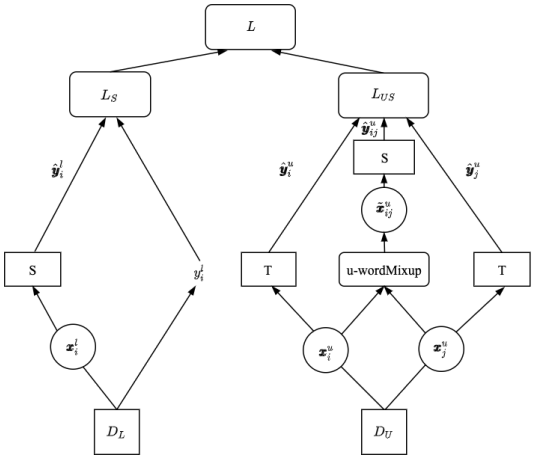


图 1 ：SD-uwM 模型

如图 1 所示，其中 $D_L = \left\{\left(\boldsymbol{x}_i^l, y_i^l\right)\right\}_{i=1}^{N^l}$ 表示有标注训练样本集，$D_U = \{(\boldsymbol{x}_i^u)\}_{i=1}^{N^u}$ 表示未标注训练样本集，N^l 表示有标注样本数，N^u 表示未标注样本数。T 为教师模型，S 为学生模型，T 和 S 具有相同网络结构。L_S 表示有监督损失，L_{US} 表示无监督损失，L 表示一致性目标损失函数。

SD-uwM 基于一致性目标损失函数 L 对有标注样本和未标注样本同时训练学习。如图 1 中左半部分所示，学生模型 S 在 D_L 上计算有监督交叉熵损失 L_S。同时，如图 1 右半部分所示，学生模型 S 利用 u-wordMixup 方法对未标注样本增强，根据学生模型 S 对增强样本的预测，以及教师模型 T 对未标注样本的预测，计算无监督一致性损失 L_{US} ,L_S 和 L_{US} 共同构成 SD-uwM 模型的一致性目标损失函数 L。经过多次迭代，SD-uwM 模型训练输出学生模型 S 的参数，作为最终分类模型的参数。

2.2 u-wordMixup 数据增强

利用 Mean Teacher 模型思想和一致性训练，SD-uwM) 的目标损失函数 L 兼顾了有监督交叉熵损失

和无监督一致性损失，定义如 (3) 所示。

$$L = L_S + \beta L_{US}. \tag{3}$$

其中，L_S 是有标注样本 D_L 上的有监督交叉熵损失，L_{US} 是未标注样本 D_U 上的无监督一致性损失，β 是比例系数。L_S 为学生模型 S 在有标注样本集 D_L 上的有监督损失，计算如式 (4)。

$$L_S = E_{\boldsymbol{x}_i^l, y_i^l \in D_L}[-y_i^l log p_{\theta'}(\boldsymbol{x}_i^l)]. \tag{4}$$

其中，y_i^l 是有标注样本 $\boldsymbol{x}_i^l$ 的真实标签，θ' 表示学生模型 S 的参数，$p_{\theta'}\left(\boldsymbol{x}_i^l\right)$ 表示学生模型 S 对样本 $\boldsymbol{x}_i^l$ 的预测伪标签，即 $\hat{\boldsymbol{y}}_i^l$。

在 wordMixup 的基础上，提出了针对未标注样本的 u-wordMixup 数据增强方法，作为 SD-uwM 模型的一部分。与 wordMixup 方法的不同，u-wordMixup 方法的插值操作对象没有真实类标签，u-wordMixup 方法如图 2 所示。

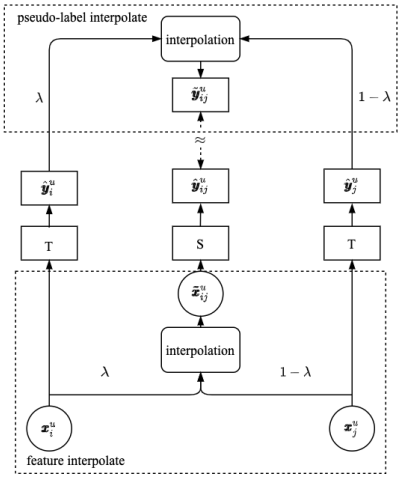


图 2 ：u-wordMixup 方法

如图 2 所示，$\boldsymbol{x}_i^u$ 和 $\boldsymbol{x}_j^u$ 表示两个未标注样本，其 $\boldsymbol{x}_i^u \in R^{N\times d}$，$\boldsymbol{x}_j^u \in R^{N\times d}$，$N$ 表示单词数，d 表示词向量维度。对 $\boldsymbol{x}_i^u$ 和 $\boldsymbol{x}_j^u$ 中的每个词一一对应进行词向量的特征插值 (feature interpolation)，得到新的未标注样本的嵌入矩阵 $\tilde{\boldsymbol{x}}_{ij}^u$，作为附加训练样本。教师模型 T 对未标注样本 $\boldsymbol{x}_i^u$ 和 $\boldsymbol{x}_j^u$ 预测生成伪标签 $\hat{\boldsymbol{y}}_i^u$ 和 $\hat{\boldsymbol{y}}_j^u$，对 $\hat{\boldsymbol{y}}_i^u$ 和 $\hat{\boldsymbol{y}}_j^u$ 做伪标签插值 (pseudo-label interpolation) 得到 $\tilde{\boldsymbol{y}}_{ij}^u$，作为增强样本 $\tilde{\boldsymbol{x}}_{ij}^u$ 的伪标签。然后学生模型 S 对增强样本 $\tilde{\boldsymbol{x}}_{ij}^u$ 预测，得到预测标签 $\hat{\boldsymbol{y}}_{ij}^u$。其中，特征插值和伪标签插值的计算分别如 (5) 和 (6)。

$$\tilde{\boldsymbol{x}}_{ij}^{uk} = q_\lambda(\boldsymbol{x}_i^u, \boldsymbol{x}_j^u) = \lambda \boldsymbol{x}_i^{uk} + (1-\lambda)\boldsymbol{x}_j^{uk}, k = 1..N. \tag{5}$$

$$\tilde{\boldsymbol{y}}_{ij}^u = q_\lambda\left(\hat{\boldsymbol{y}}_i^u, \hat{\boldsymbol{y}}_j^u\right) = \lambda \hat{\boldsymbol{y}}_i^u + (1-\lambda)\hat{\boldsymbol{y}}_j^u. \tag{6}$$

其中，$q_\lambda\left(\boldsymbol{x}_i^u, \boldsymbol{x}_j^u\right)$ 表示 u-wordMixup 数据增强变换，$\boldsymbol{x}_i^{uk}$ 和 $\boldsymbol{x}_j^{uk}$ 分别表示样本 $\boldsymbol{x}_i^u$ 和 $\boldsymbol{x}_j^u$ 的第 k 个单词的词向量，$\lambda \in [0,1]$ 为插值权重因子，$\tilde{\boldsymbol{x}}_{ij}^{uk}$ 为插值生成的增强样本的第 k 个词的词向量。$\hat{\boldsymbol{y}}_i^u$ 为教师模型 T 对 $\boldsymbol{x}_i^u$ 的预测伪标签，$\hat{\boldsymbol{y}}_j^u$ 为教师模型 T 对 $\boldsymbol{x}_j^u$ 的预测伪标签，$\tilde{\boldsymbol{y}}_{ij}^u$ 为插值生成的伪标签，即 $\tilde{\boldsymbol{x}}_{ij}^u$ 的伪标签。

基于一致性训练，学生模型 S 对增强的未标注样本 $\tilde{\boldsymbol{x}}_{ij}^u$ 预测伪标签为 $\hat{\boldsymbol{y}}_{ij}^u$，应该尽可能地与插值生成的伪标签 $\tilde{\boldsymbol{y}}_{ij}^u$ 一致，即 $\tilde{\boldsymbol{y}}_{ij}^u \approx \hat{\boldsymbol{y}}_{ij}^u$，理想情况二者相等。因此，无监督一致性损失 L_{US} 的计算如 (7)。

$$L_{US} = E_{\boldsymbol{x}_i^u, \boldsymbol{x}_j^u \in D_U} E_{\tilde{\boldsymbol{x}}_{ij}^u \sim q_\lambda(\boldsymbol{x}_i^u, \boldsymbol{x}_j^u), \lambda \in [0,1]}[- (\lambda p_\theta(\boldsymbol{x}_i^u) + (1-\lambda) p_\theta(\boldsymbol{x}_j^u)) log p_{\theta'}(\tilde{\boldsymbol{x}}_{ij}^u)]. \tag{7}$$

其中 θ 表示教师模型 T 的参数，θ' 表示学生模型 S 的参数，θ 是 θ' 的移动均值。$p_\theta(\boldsymbol{x}_i^u)$ 表示教师模型 T 对样本 $\boldsymbol{x}_i^u$ 的预测伪标签，$p_\theta(\boldsymbol{x}_j^u)$ 表示教师模型 T 对样本 $\boldsymbol{x}_j^u$ 的预测伪标签，$p_{\theta'}(\tilde{\boldsymbol{x}}_{ij}^u)$ 表示学生模型 S 对新生成样本 $\tilde{\boldsymbol{x}}_{ij}^u$ 的预测伪标签。$\lambda \in [0,1]$ 为插值权重因子，$q_\lambda(\boldsymbol{x}_i^u, \boldsymbol{x}_j^u)$ 表示 u-wordMixup 数据增强变换。

深度半监督学习 SD-uwM 模型，采用 u-wordMixup 方法对未标注样本进行增强，无监督一致性损失 L_{US} 是以减少一致性损失为目标，约束增强后的未标注样本的质量。并结合 Mean Teacher 模型构造教师模型 T 和学生模型 S，对有标注样本和未标注样本进行训练，L_S 和 L_{US} 加权求和作为模型 SD-uwM 的目标函数 L。

2.3 SD-uwM 算法描述

SD-uwM 的算法描述如算法 1 所示。

算法的目标函数 L 兼顾了有监督交叉熵损失 L_S 和无监督一致性损失 L_{US}，约束增强生成未标注训练样本。每次迭代，根据目标函数 L，优化学生模型 S 的参数 θ'。经过多次迭代，最终得到最优的学生模型 S 的参数 θ'。

3 实验分析

3.1 数据集和预处理

本文选择 AGNews、20 Newsgroups 和 THUCNews 三个数据集，AGNews 选择“世界”、“体育”、“业务”和“科技”4 个类，20 Newsgroups 选择“alt.atheism”、“soc.religion.christian”、“comp.graphics”和“sci.med”4 个类，THUCNews 选择“财经”、“彩票”、“教育”和“科技”4 个类。算法 1 中比例系数 β 设置为 1。

3.2 实验结果及分析

实验中的对比方法如下：

1) SD-uwM：本文提出的基于 u-wordMixup 数据

增强的半监督深度学习模型；

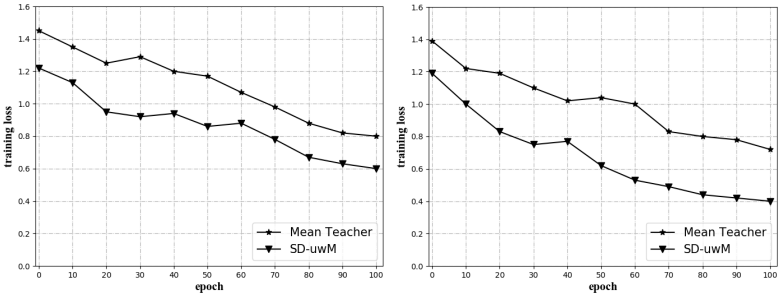
2) wM-SL：文献 [16] 中基于 wordMixup 数据增强方法的监督文本分类方法；

3) SL：无数据增强的监督文本分类方法;

4) Mean Teacher：文献 [18] 中应用于图像分类的半监督方法，修改后用于文本分类任务。

3.2.1 SD-uwM 与 Mean Teacher 的目标损失比较

为验证 u-wordMixup 方法的有效性，对 SD-uwM 和 Mean Teacher 模型进行了对比实验，训练损失变化如图 3。



(a) N^l=200, N^u=3000(THUCNews 数据集)　(b) N^l=300, N^u=5000(AGNews 数据集)

图 3 ：选用 TextCNN 时 SD-uwM 与 Mean Teacher 的训练损失比较

Algorithm 1 SD-uwM 算法

Input: 有标注样本集 D_L，未标注样本集 D_U，比例系数 β，插值权重因子 λ，迭代数 R，移动平均率 α

Initialization: θ，$\theta^{'}$

for $r=1$ to R **do**

　Sample$\left\{\left(\boldsymbol{x}_i^l, y_i^l\right)\right\}_{i=1}^{N^l} \sim D_L$

　$\hat{\boldsymbol{y}}_i^l = p_{\theta'}\left(\boldsymbol{x}_i^l\right)$

　$L_S = E_{\boldsymbol{x}_i^l, y_i^l \in D_L}[-y_i^l log p_{\theta'}(\boldsymbol{x}_i^l)]$

　Sample$\{\boldsymbol{x}_i^u\}_{i=1}^{N^u}$，$\left\{\boldsymbol{x}_j^u\right\}_{j=1}^{N^u} \sim D_U$

　$\hat{\boldsymbol{y}}_i^u = p_\theta\left(\boldsymbol{x}_i^u\right)$，$\hat{\boldsymbol{y}}_j^u = p_\theta\left(\boldsymbol{x}_j^u\right)$

　$\tilde{\boldsymbol{x}}_{ij}^u = q_\lambda\left(\boldsymbol{x}_i^u, \boldsymbol{x}_j^u\right)$

　$\hat{\boldsymbol{y}}_{ij}^u = p_{\theta'}\left(\tilde{\boldsymbol{x}}_{ij}^u\right)$

　$\tilde{\boldsymbol{y}}_{ij}^u = q_\lambda\left(\hat{\boldsymbol{y}}_i^u, \hat{\boldsymbol{y}}_j^u\right)$

　$L_{US} = E_{\boldsymbol{x}_i^u, \boldsymbol{x}_j^u \in D_U} E_{\tilde{\boldsymbol{x}}_{ij}^u \sim q_\lambda(\boldsymbol{x}_i^u, \boldsymbol{x}_j^u), \lambda \in [0,1]}[-$

　　$(\lambda p_\theta(\boldsymbol{x}_i^u) + (1-\lambda) p_\theta(\boldsymbol{x}_j^u)) log p_{\theta'}(\tilde{\boldsymbol{x}}_{ij}^u)]$

　$L\left(\theta^{'}\right) = L_S + \beta L_{US}$

　$g_{\theta'} \leftarrow \nabla L\left(\theta^{'}\right)$

　$\theta = \alpha\theta + (1-\alpha)\,\theta^{'}$

　$\theta^{'} \leftarrow \text{Adam}\left(\theta^{'}, g_{\theta'}\right)$

end for

Output: 学生模型的参数 $\theta^{'}$

从图 3 可以看出，相比于 Mean Teacher 模型，利用 u-wordMixup 数据增强方法的 SD-uwM 模型的训练损失更低。这是由于 SD-uwM 模型目标损失函数更贴合实际，其中 L_{US} 是结合 u-wordMixup 方法，以降低无监督一致性损失为目标，能够提高未标注样本质量，从而提高了模型的性能。

3.2.2 SD-uwM 模型与其它方法的分类性能比较

(1)SD-uwM 模型与其它方法的正确率比较

在 AGNews(N^l=300, N^u=5000)、THUCNews(N^l=300, N^u=5000) 和 20 Newsgroups(N^l=200, N^u=2000) 上，SD-uwM 模型与 Mean Teacher、wM-SL、SL 模型的对比实验结果如表 1。

表 1 四种模型在三种数据集上的分类结果比较

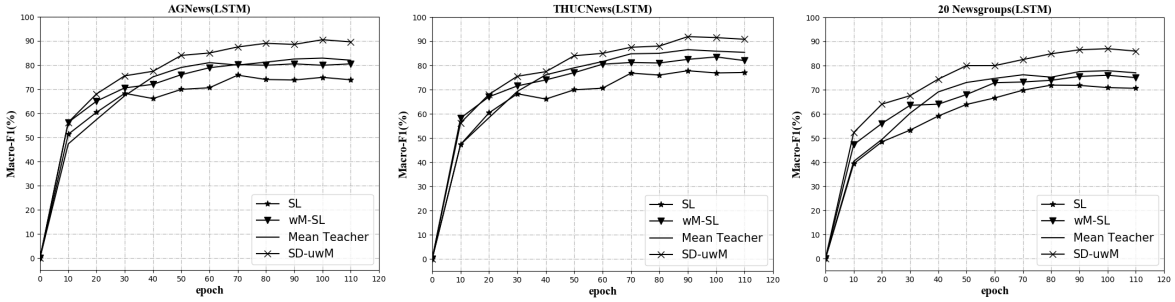
模型	网络结构	Accuracy（%）		
		AGNews	THUCNews	20 Newsgroups
SL	LSTM	75.4±1.1	77.5±1.3	71.5±1.3
wM-SL	LSTM	80.4±1.3	83.2±1.2	75.4±1.3
Mean Teacher	LSTM	82.1±1.3	86.1±1.3	77.5±1.1
SD-uwM	LSTM	**90.4±1.2**	**91.4±1.3**	**85.4±1.2**
SL	TextCNN	76.4±1.2	78.4±1.4	71.2±1.2
wM-SL	TextCNN	80.5±1.2	84.5±1.3	75.3±1.2
Mean Teacher	TextCNN	83.6±1.1	86.1±1.5	78.1±1.3
SD-uwM	TextCNN	**91.2±1.3**	**92.2±1.3**	**86.2±1.1**

从表 1 中可以看出，在三个数据集上，无论网络结构选用 LSTM 还是 TextCNN，SD-uwM 模型的分类精度均优于 SL、wM-SL 和 Mean Teacher 模型。

SL 模型是监督学习方法，需要大量的有标注样本才能取得较好的性能，wM-SL 模型只对有标注样本进行增强，Mean Teacher 模型未使用 u-wordMixup 方法对样本进行增强。而 SD-uwM 模型利用 u-wordMixup 方法对未标注样本进行数据增强，利用无监督一致性损失约束，提高模型的泛化能力。

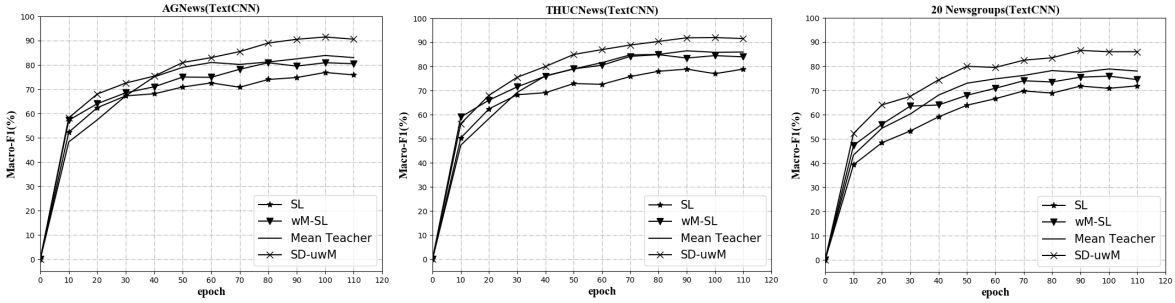
(2) 随迭代增加 SD-uwM 模型的分类性能比较

固定有标注样本数和未标注样本数，随着迭代次数增加，对比分析 SD-uwM 模型与 SL、wM-SL、Mean Teacher 模型的 Macro-F1 值变化。在 AGNews(N^l=300, N^u=5000)、THUCNews(N^l=300, N^u=5000) 和 20 Newsgroups(N^l=200, N^u=2000) 上，实验结果如图 4 和图 5 所示。



(a)AGNews 数据集 (b)THUCNews 数据集 (c)20 Newsgroups 数据集

图 4 ：选用 LSTM 在三种数据集上各模型 Macro-F1 值随迭代次数变化比较



(a)AGNews 数据集 (b)THUCNews 数据集 (c)20 Newsgroups 数据集

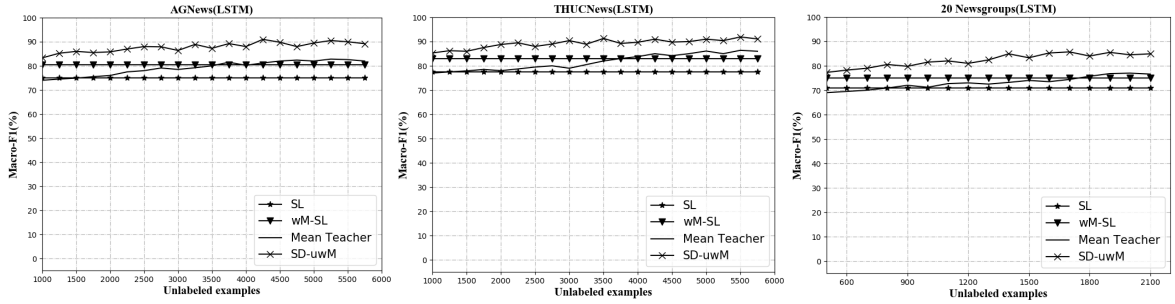
图 5 ：选用 TextCNN 在三种数据集上各模型 Macro-F1 值随迭代次数变化比较

从图 4 和图 5 可以看出，随着迭代次数增加，虽然 SL、wM-SL、Mean Teacher 和 SD-uwM 模型的指标大体呈上升趋势，并收敛至某个上限，但 SD-uwM 模型分类性能明显优于 SL、wM-SL 和 Mean Teacher 模型。如图 5(a) 中，选用 LSTM 在 AGNews 上，对比 Mean Teacher 、wM-SL 和 SL，SD-uwM 的 Macro-F1 达到了 90.3%，分别提高了 8%、9.9% 和 14.5%。

这是由于 SD-uwM 利用 u-wordMixup 方法以无监督一致性损失为目标，进行未标注样本增强，能够减少过度拟合，从而提高分类性能。

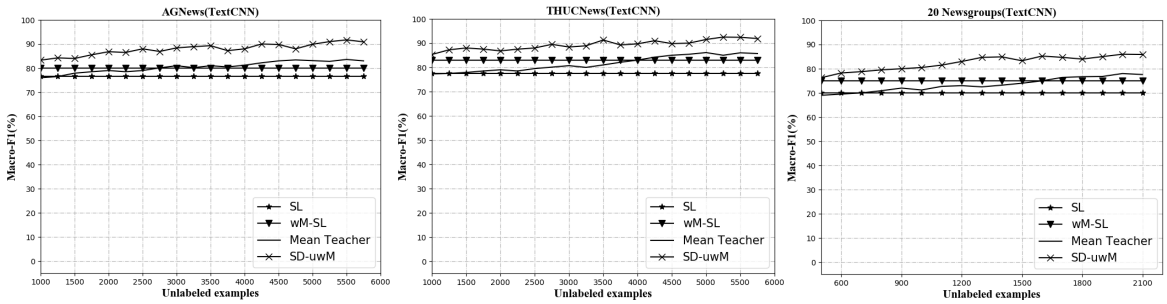
3.2.3 未标注样本对 SD-uwM 模型的影响

为验证未标注样本对 SD-uwM 模型的影响，在 AGNews 和 THUCNews 上固定有标注样本数 N^l=300，20 Newsgroups 上固定有标注样本数 N^l=200，不断增加未标注样本，对比 SD-uwM 模型与 SL、wM-SL、Mean Teacher 模型的分类结果如图 6 和图 7 所示。



(a)AGNews 数据集 (b) THUCNews 数据集 (c)20 Newsgroups 数据集

图 6 ：选用 LSTM 在三种数据集上各模型 Macro-F1 值随未标注样本数变化对比



(a)AGNews 数据集 (b)THUCNews 数据集 (c)20 Newsgroups 数据集

图 7 ：选用 TextCNN 在三种数据集上各模型 Macro-F1 值随未标注样本数变化对比

从图 6 和图 7 可以看出，随着未标注样本的增加，SD-uwM 和 Mean Teacher 的指标呈上升趋势，但 SD-uwM 的分类结果明显优于 Mean Teacher、wM-SL 和 SL。如图 7(b) 中，选用 LSTM 在 THUCNews 上，对比 Mean Teacher 、wM-SL 和 SL，SD-uwM 的 Macro-F1 达到了 91.4%，分别提高了 5.3%、8.2% 和 13.9%。

可以看出，由于 SD-uwM 模型利用 u-wordMixup 方法对未标注样本进行增强，并结合 Mean Teacher 模型思想，因此能够提高文本分类性能。

3.2.4 SD-uwM 模型的时间性能分析

通常的半监督学习算法在选择未标注样本时，需要计算未标注样本与有标注样本相似度矩阵，这将增加时间复杂度。而本文的 SD-uwM 模型是随机采样，不需计算两种样本的相似度。SD-uwM 模型与典型半监督学习方法 Co-training 的时间性能比较如表 2 所示。

表 2 SD-uwM 和 Co-training 的时间性能比较

dataset	model	Accuracy(%)	time(s)	ratio of train time (SD-uwM/Co-training)
20 Newsgroups	SD-uwM	**86.4±1.3**	**0.01**	1/3000
$N^l = 200, N^u = 2000$	Co-training	83.3±1.2	30	
THUCNews	SD-uwM	**90.5±1.3**	**0.02**	1/2200
$N^l = 300, N^u = 4000$	Co-training	88.4±1.2	44	

从表 2 中可以看出，SD-uwM 模型的分类准确率高于 Co-training 模型，同时，SD-uwM 模型的时间性能明显优于 Co-training。原因在于，在选择未标注训练样本时，SD-uwM 模型是随机采样，时间复杂度为 O(1)，而 Co-training 方法需要计算样本相似度矩阵，时间复杂度为 O($N^l * N^u$)。

4 结论

本文提出了一种对未标注样本进行数据增强的 u-wordMixup 方法，并结合一致性训练框架和 Mean Teacher 方法，提出了一种半监督深度学习模型 SD-uwM。该模型利用 u-wordMixup 方法对未标注样本进行数据增强，兼顾有监督交叉熵损失和无监督一致性损失构建新的目标函数。实验表明，SD-uwM 模型能够提高模型的泛化能力和时间性能。下一步，我们将针对类别不均衡的数据集，开展数据增强和半监督深度学习方法研究。

参考文献 (References)

[1] Miyato, Takeru and Maeda, Shin Ichi and Koyama, Masanori and Ishii, Shin. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 44(8): 1979-1993.

[2] YANG H F, LIN K, CHEN C S. Supervised learning of semantics-preserving hash via deep convolutional neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(2): 437–451.

[3] 唐焕玲, 窦全胜, 于立萍, 宋英杰, 鲁明羽. 有监督主题模型的 SLDA-TC 文本分类新方法, 电子学报, 2019,47 (6): 1300-1308.
(Tang Huanling, Dou Quansheng, Yu Liping, Song Yingjie, Lu Mingyu. A new approach to SLDA-TC text classification with supervised topic models,Journal of Electronics, 2019,47 (6): 1300-1308.)

[4] 唐焕玲, 邬俊, 鲁明羽. 基于投票信息熵的 AdaBoost 改进算法. 控制与决策. 2010,25(4),481-492.
(Tang Huanling, Wu Jun, Lu Mingyu. Improved AdaBoost Algorithm Based on Voting Information Entropy. Control and Decision. 2010,25(4),481-492.)

[5] 朱建勇, 周振辰, 杨辉, 等. 基于 Hessian 正则的自适应损失半监督特征选择 [J]. 控制与决策, 2021, 36(8): 1862-1870.
(Zhu Jianyong, Zhou Zhenchen, Yang Hui, et al. Adaptive loss semi-supervised feature selection based on Hessian regularization[J]. Control and Decision, 2021, 36(8): 1862-1870.)

[6] Berthelot, David, et al. MixMatch: A Holistic Approach to Semi-Supervised Learning[J]. Advances in Neural Information Processing Systems, 2019, 32: 5049–5059.

[7] BLUM A, MITCHELL T. Combining labeled and

unlabeled data with co-training [C]. Proceedings of the 11th Annual Conference on Computational Learning Theory. Madison: ACM, 1998: 92–100.

[8] Zhou Z H, Li M. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Transactions on knowledge and Data Engineering, 2005, 17(11): 1529-1541.

[9] Sajjadi M, Javanmardi M, Tasdizen T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning[J]. Advances in neural information processing systems, 2016, 29: 1163-1171.

[10] Sajjadi M, Javanmardi M, Tasdizen T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning[J]. Advances in neural information processing systems, 2016, 29: 1163-1171.

[11] Clark, K. and Luong, M. T. and Manning, C. D. and Le, Quoc. Semi-Supervised Sequence Modeling with Cross-View Training[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 1914–1925.

[12] Wei J, Zou K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks[J]. 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: 6381–6387.

[13] Thyagarajan A. Siamese Recurrent Architectures for Learning Sentence Similarity[C]. Thirtieth Aaai Conference on Artificial Intelligence. AAAI Press, 2016, 30(1).

[14] Xiaoqi Jiao, Yichun Yin, Lifeng Shang and Xin Jiang. TinyBERT: Distilling BERT for Natural Language Understanding[J]. Findings of the Association for Computational Linguistics: EMNLP 2020, 2020: 4163–4174.

[15] Xiaoqi Jiao, Yichun Yin, Lifeng Shang and Xin Jiang. mixup: Beyond empirical risk minimization[J]. International Conference on Learning Representations, 2017.

[16] Guo H, Mao Y, Zhang R. Augmenting Data with Mixup for Sentence Classification: An Empirical Study[J]. arXiv preprint arXiv:1905.08941, 2019.

[17] Tokozume Y, Ushiku Y, Harada T. Between-class learning for image classification[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 5486–5494.

[18] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[J]. arXiv preprint arXiv:1703.01780, 2017.

作者简介

唐焕玲 (1970–), 女, 教授, 博士, 从事机器学习、人工智能、数据挖掘等研究, E-mail: th101@163.com;

宋双梅 (1997–), 女, 研究生, 从事机器学习、人工智能、数据挖掘的研究, E-mail: 2425439857@qq.com;

刘孝炎 (1997–), 男, 研究生, 从事机器学习、人工智能、数据挖掘的研究, E-mail: lxy15058247683@aliyun.com;

窦全胜 (1971–), 男, 教授, 博士, 从事人工智能、机器学习、演化计算等研究, E-mail: li_dou@163.com;

鲁明羽 (1963–), 男, 教授, 博士生导师, 从事机器学习、人工智能、数据挖掘等研究, E-mail: lumingyu@dlmu.edu.cn.