控制与决策
Control and Decision

基于u-wordMixup的半监督深度学习模型

唐焕玲, 宋双梅, 刘孝炎, 窦全胜, 鲁明羽

引用本文:
唐焕玲,宋双梅,刘孝炎,窦全胜,鲁明羽. 基于u-wordMixup的半监督深度学习模型[J]. 控制与决策, 2023, 38(6): 1646-1652.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.1789

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. 山东工商学院 山东省高校智能信息处理重点实验室，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, 基于wordMixup方法, 提出针对未标注样本进行数据增强的u-wordMixup方法, 结合一致性训练框架和Mean Teacher模型, 提出一种基于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, 所提出方法能够提高模型的泛化能力, 同时有效提高时间性能.

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):**

引用格式: 唐焕玲, 宋双梅, 刘孝炎, 等. 基于 u-wordMixup 的半监督深度学习模型 [J]. 控制与决策, 2023, 38(6): 1646-1652.

Semi-supervised deep learning model based on u-wordMixup

TANG Huan-ling[1,3,4†], *SONG Shuang-mei*[2], *LIU Xiao-yan*[1], *DOU Quan-sheng*[1,3,4], *LU Ming-yu*[5]

(1. School of Computer Science and Technology，Shandong Technology and Business University，Yantai 264005，China；2. School of Information and Electronic Engineering，Shandong Technology and Business University，Yantai 264005，China；3. Co-innovation Center of Shandong Colleges and Universities：Future Intelligent Computing，Yantai 264005，China；4. Key Laboratory of Intelligent Information Processing in Universities of Shandong，Shandong Technology and Business University，Yantai 264005，China；5. Information Science and Technology College，Dalian Maritime University，Dalian 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, as the unlabeled data and labeled data come from different fields, quality problems of unlabeled data would be callsed, 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 the u-wordMixup (SD-uwM) by combining the consistent training framework and the Mean Teacher model. The model utilizes the u-wordMixup method to augment the 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

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).

责任编委: 胡清华.

†通讯作者. E-mail: th101@163.com.

*本文附带电子附录文件, 可登录本刊官网该文“资源附件”区自行下载阅览.

0 引　言

传统的监督学习[1-4]需要大量标注样本,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,而用于文本分类的研究不多,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],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[10-12]和词向量替换[13]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[14]在图像分类中表现优异,结合文本特征的wordMixup(词混合)方法[15]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[16],基于wordMixup数据增强方法提出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[17]进行一致性训练,提出一种新的半监督深度学习模型(SD-uwM). 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;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),对其进行词嵌入(word embedding)得到$(\boldsymbol{x}_i, \boldsymbol{y}_i)$和$(\boldsymbol{x}_j, \boldsymbol{y}_j)$,其中:$\boldsymbol{x}_i \in R^{N\times d}$,$\boldsymbol{x}_j \in R^{N\times d}$为文本$x_i$和$x_j$的词嵌入矩阵,$N$为单词数,$d$为词向量维度,$\boldsymbol{y}_i$和$\boldsymbol{y}_j$为相应的类标签. 按照式(1)和(2)进行插值(interpolation)得到新样本$(\tilde{\boldsymbol{x}}_{ij}, \tilde{\boldsymbol{y}}_{ij})$,$\tilde{\boldsymbol{x}}_{ij}$为增强样本的词嵌入矩阵,$\tilde{\boldsymbol{y}}_{ij}$为其类标签,有

$$\tilde{\boldsymbol{x}}_{ij}^k = \lambda \boldsymbol{x}_i^k + (1-\lambda)\boldsymbol{x}_j^k,\ k = 1, 2, \ldots, N; \quad (1)$$

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

其中:$\boldsymbol{x}_i^k$和$\boldsymbol{x}_j^k$分别为文本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$(\boldsymbol{x}_j, \boldsymbol{y}_j)$中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,$(\tilde{x}_{ij}, \tilde{y}_{ij})$即为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? 对此,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.

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1所示,其利用Mean Teacher模型思想构造教师模型T和学生模型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.

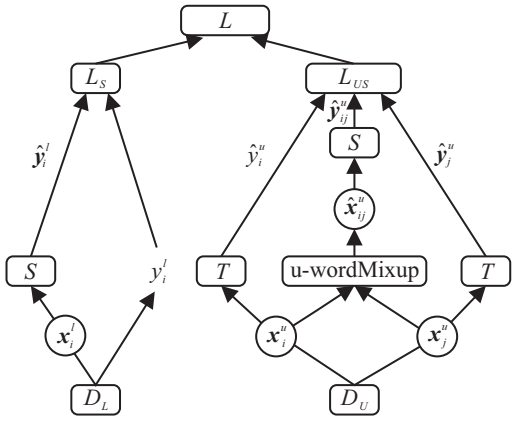


图 1　SD-uwM模型

如图1所示,$D_L = \{(\boldsymbol{x}_i^l, y_i^l)\}_{i=1}^{N^l}$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,定义如下:

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

其中:L_S为有标注样本D_L上的有监督交叉熵损失,L_{US}为未标注样本D_U上的无监督一致性损失,β为比例系数. L_S为学生模型S在有标注样本集D_L上的有监督损失,计算如下:

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

其中:y_i^l为有标注样本$\boldsymbol{x}_i^l$的真实标签;θ'为学生模型S的参数;$p_{\theta'}(\boldsymbol{x}_i^l)$为学生模型$S$对样本$\boldsymbol{x}_i^l$的预测伪标签,即$\hat{\boldsymbol{y}}_i^l$.

在wordMixup基础上,提出针对未标注样本的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,如图2所示.

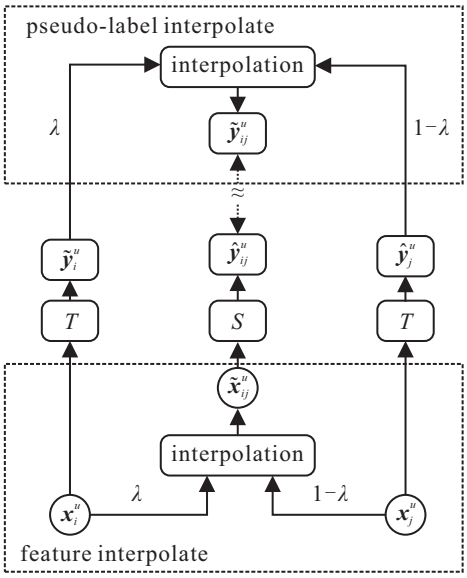


图 2 u-wordMixup方法

图2中,$\boldsymbol{x}_i^u$和$\boldsymbol{x}_j^u$表示两个未标注样本,$\boldsymbol{x}_i^u \in R^{N\times d}$,$\boldsymbol{x}_j^u \in R^{N\times d}$,$N$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. 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为

$$\tilde{\boldsymbol{x}}_{ij}^{uk} = q_\lambda(\boldsymbol{x}_i^u, \boldsymbol{x}_j^u) = \lambda \boldsymbol{x}_i^{uk} + (1-\lambda)\boldsymbol{x}_j^{uk},$$
$$k = 1, 2, \ldots, N; \tag{5}$$

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

其中:$q_\lambda(\boldsymbol{x}_i^u, \boldsymbol{x}_j^u)$为u-wordMixup数据增强变换;$\boldsymbol{x}_i^{uk}$和$\boldsymbol{x}_j^{uk}$分别为样本$\boldsymbol{x}_i^u$和$\boldsymbol{x}_j^u$第$k$个单词的词向量;$\lambda \in [0,1]$为插值权重因子;$\tilde{\boldsymbol{x}}_{ij}^{uk}$为插值生成增强样本第$k$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$L_{US}$计算为

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

其中:θ为教师模型T的参数,θ'为学生模型S的参数,θ为θ'的移动均值,$p_\theta(\boldsymbol{x}_i^u)$为教师模型$T$对样本$\boldsymbol{x}_i^u$的预测伪标签,$p_\theta(\boldsymbol{x}_j^u)$为教师模型$T$对样本$\boldsymbol{x}_j^u$的预测伪标签,$p_{\theta'}(\tilde{\boldsymbol{x}}_{ij}^u)$为学生模型$S$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,$q_\lambda(\boldsymbol{x}_i^u, \boldsymbol{x}_j^u)$为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}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. 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算法描述如下.

算法1 SD-uwM算法.

1: 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.

2: initialization: θ, θ'

3: for $r = 1$ to R do

4: sample $\{(\boldsymbol{x}_i^l, y_i^l)\}_{i=1}^{N^l} \sim D_L$

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

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

7: sample $\{\boldsymbol{x}_i^u\}_{i=1}^{N^u}, \{\boldsymbol{x}_j^u\}_{j=1}^{N^u} \sim D_U$

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

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

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

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

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

13: $L(\theta') = L_S + \beta L_{US}$

14: $g_{\theta'} \leftarrow \nabla L(\theta')$

15: $\theta = \alpha\theta + (1-\alpha)\theta'$

16: $\theta' \leftarrow \mathrm{Adam}(\theta', g_{\theta'})$

17: end for

18: output: 学生模型的参数θ'.

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. 每次迭代中,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

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;

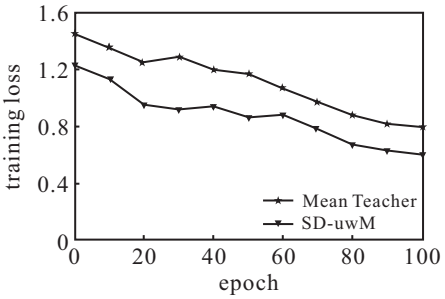
2) wM-SL: 文献[15]中基于wordMixup数据增强方法的监督文本分类方法;

3) SL: 无数据增强的监督文本分类方法;

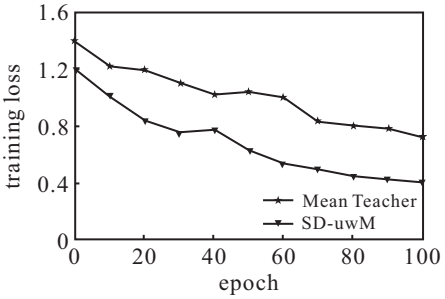
4) Mean Teacher: 文献[17]中应用于图像分类的半监督方法,修改后用于文本分类任务.

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模型进行对比实验,训练损失变化如图3所示. 由图3可见,相比于Mean Teacher模型,利用u-wordMixup数据增强方法的SD-uwM模型训练损失更低. 这是由于SD-uwM模型目标损失函数更贴合实际,其中L_{US}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,从而提高模型性能.



(a)　N^l=200, N^u=3 000(THUCNews数据集)



(b)　N^l=300, N^u=5 000(AGNews数据集)

图3　选用TextCNN时SD-uwM与Mean Teacher的训练损失对比

3.2.2　SD-uwM模型与其他方法的分类性能比较

1) SD-uwM模型与其他方法的正确率比较.

在AGNews($N^l = 300,\ N^u = 5\,000$)、THUCNews($N^l = 300,\ N^u = 5\,000$)和20 Newsgroups($N^l = 200,\ N^u = 2\,000$)上,SD-uwM模型与Mean Teacher、wM-SL、SL模型的对比结果如表1所示.

表1　4种模型在3种数据集上的分类结果比较

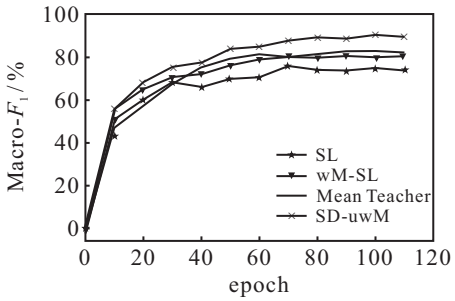
模型	网络结构	accuracy / %		
		AGNews	THUCNews	20 Newsgroups
SL	LSTM	75.4±1.1	77.5±1.3	71.5±1.3
wM-SL	LSTM	80.4±1.3	83.2±1.2	75.4±1.3
Mean Teacher	LSTM	82.1±1.3	86.1±1.3	77.5±1.1
SD-uwM	LSTM	**90.4±1.2**	**91.4±1.3**	**85.4±1.2**
SL	TextCNN	76.4±1.2	78.4±1.4	71.2±1.2
wM-SL	TextCNN	80.5±1.2	84.5±1.3	75.3±1.2
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**

由表1可见,在3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.

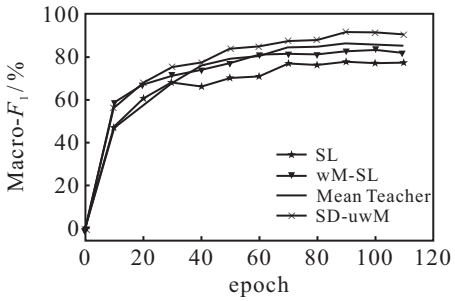
SL模型是监督学习方法,需要大量有标注样本才能取得较好的性能,wM-SL模型只对有标注样本进行增强,Mean Teacher模型未使用u-wordMixup方法对样本进行增强. 而SD-uwM模型利用u-wordMixup方法对未标注样本进行数据增强,利用无监督一致性损失约束提高模型的泛化能力.

2) 随迭代增加SD-uwM模型的分类性能比较.

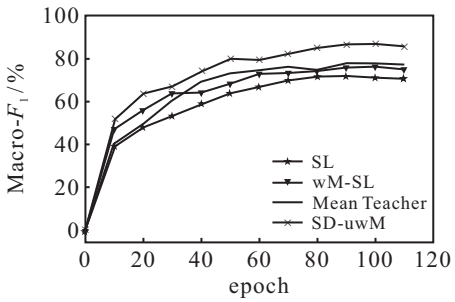
固定有标注样本数和未标注样本数,随着迭代次数增加,对比分析SD-uwM模型与SL、 wM-SL、Mean Teacher模型的Macro-F_1值变化情况. 在AGNews ($N^l = 300,\ N^u = 5\,000$)、THUCNews ($N^l = 300,\ N^u = 5\,000$)和20 Newsgroups ($N^l = 200,\ N^u = 2\,000$)上,实验结果如图4和图5所示.



(a) AGNews数据集



(b) THUCNews数据集



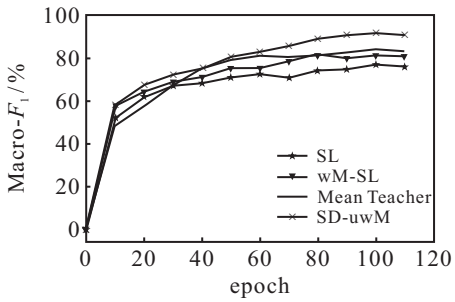
(c) 20 Newsgroups数据集

图4 选用LSTM在3种数据集上各模型Macro-F_1值随迭代次数变化比较

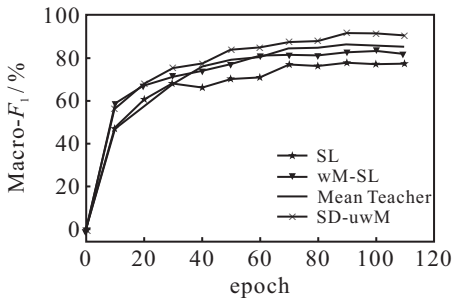
由图4和图5可见,随着迭代次数的增加,虽然SL、wM-SL、Mean Teacher和SD-uwM模型的指标大体呈上升趋势,并收敛至某个上限,但SD-uwM模型分类性能明显优于SL、wM-SL和Mean Teacher模型. 如图5(a)中,选用LSTM在AGNews上,对比Mean Teacher、wM-SL和SL, SD-uwM的Macro-F_1达到了90.3%,分别提高了8%、 9.9%和14.5%. 这是由于SD-uwM利用u-wordMixup方法以无监督一致性损失为目标进行未标注样本增强,能够减少过度拟合,从而提高分类性能.

3.2.3 未标注样本对SD-uwM模型的影响

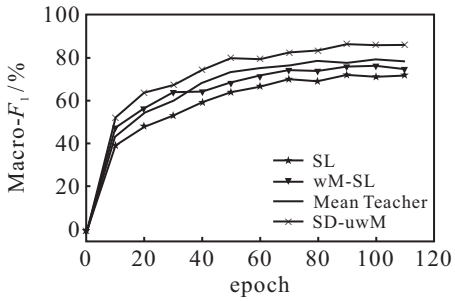
为验证未标注样本对SD-uwM模型的影响,在AGNews和THUCNews上固定有标注样本数N^l = 300, 20 Newsgroups上固定有标注样本数N^l = 200,不断增加未标注样本,对比SD-uwM模型与SL、wM-SL、Mean Teacher模型的分类结果如图6和图7所示.



(a) AGNews数据集



(b) THUCNews数据集



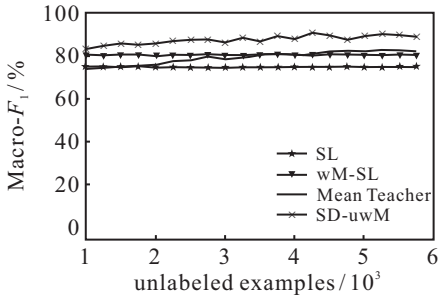
(c) 20 Newsgroups数据集

图5 选用TextCNN在3种数据集上各模型Macro-F_1值随迭代次数变化对比

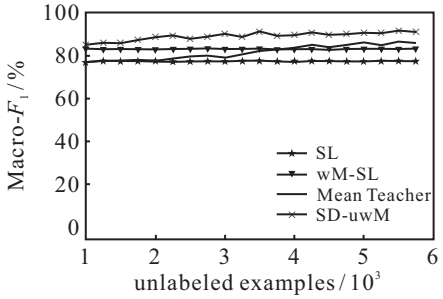
由图6和图7可见,随着未标注样本的增加,SD-uwM和Mean Teacher的指标呈上升趋势,但SD-uwM的分类结果明显优于Mean Teacher、 wM-SL和SL. 如图7(b)中,选用LSTM在THUCNews上,对比Mean Teacher、 wM-SL和SL, SD-uwM的Macro-F_1达到了91.4%,分别提高了5.3%、8.2%和13.9%. 可以看出,SD-uwM模型利用u-wordMixup方法对未标注样本进行增强,并结合了Mean Teacher模型思想,因此能够提高文本分类性能.

3.2.4 SD-uwM模型的时间性能分析

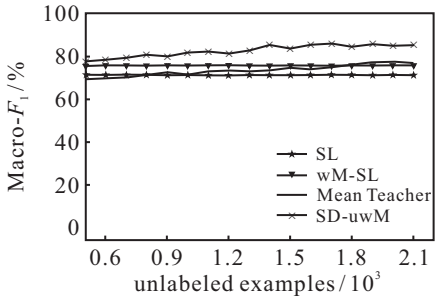
通常的半监督学习算法在选择未标注样本时,需要计算未标注样本与有标注样本相似度矩阵,这将增加时间复杂度. 而本文的SD-uwM模型是随机采样,不需要计算两种样本的相似度. SD-uwM模型与典型半监督学习方法Co-training的时间性能比较如表2所示.



(a) AGNews数据集

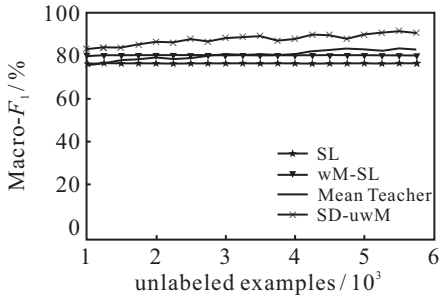


(b) THUCNews数据集

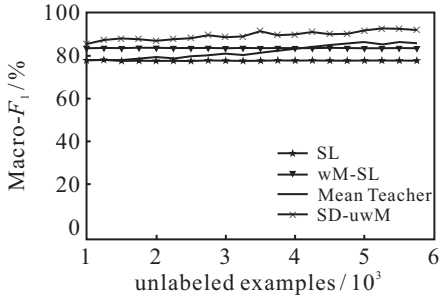


(c) 20 Newsgroups数据集

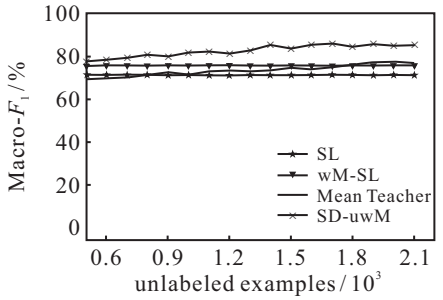
图 6　选用LSTM在3种数据集上各模型Macro-F_1值随未标注样本数变化对比



(a) AGNews数据集



(b) THUCNews数据集



(c) 20 Newsgroups数据集

图 7　选用TextCNN在3种数据集上各模型Macro-F_1值随未标注样本数变化对比

表 2　SD-uwM和Co-training时间性能比较

dataset	model	accuracy (%)	time (s)	ratiooftraintime (SD-uwM/Co-training)
20 Newsgroups $N^l=200, N^u=2\,000$	SD-uwM	**86.4±1.3**	**0.01**	1/3 000
	Co-training	83.3±1.2	30	
THUCNews $N^l=300, N^u=4\,000$	SD-uwM	**90.5±1.3**	**0.02**	1/2 200
	Co-training	88.4±1.2	44	

由表2可见,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,当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$O(N^lN^u)$.

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,同时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 T, Maeda S I, Koyama M, et al. Virtual adversarial training: A regularization method for supervised and semi-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(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, 等. 有监督主题模型的SLDA-TC文本分类新方法[J]. 电子学报, 2019, 47(6): 1300-1308.
(Tang H L, Dou Q S, Yu L P, et al. SLDA-TC: A novel text categorization approach based on supervised topic model[J]. Acta Electronica Sinica, 2019, 47(6): 1300-1308.)

[4] 唐焕玲, 鲁明羽, 邬俊. 基于投票信息熵的AdaBoost改进算法[J]. 控制与决策, 2010, 25(4): 487-492.
(Tang H L, Lu M Y, Wu J. Improved AdaBoost algorithm based on vote entropy[J]. Control and Decision, 2010, 25(4): 487-492.)

[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 J Y, Zhou Z C, Yang H, et al. Adaptive loss semi-supervised feature selection based on Hessian regularization[J]. Control and Decision, 2021, 36(8): 1862-1870.)

[6] Berthelot David, Carlini Nicholas, Goodfellow Ian, 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 Eleventh 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 deepsemi-supervised learning[J]. Advances in Neural Information Processing Systems, 2016, 29: 1163-1171.

[10] Clark K, Luong M T, Manning C D, et al. Semi-supervised sequence modeling with cross-view training[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, 2018: 1914−1925.

[11] Wei J, Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks[C]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, 2019: 6381-6387.

[12] Thyagarajan A. Siamese recurrent architectures for learning sentence similarity[C]. The 30th Aaai Conference on Artificial Intelligence. AAAI Press, 2016: 2786-2792.

[13] Jiao X Q, Yin Y C, Shang L F, et al. TinyBERT: Distilling BERT for natural language understanding[C]. Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg, 2020: 4163-4174.

[14] Jiao X Q, Yin Y C, Shang L F, et al. Mixup: Beyond empirical risk minimization[J/OL]. 2017, arXiv: 1710.09412.

[15] Guo H, Mao Y, Zhang R. Augmenting data with mixup for sentence classification: An empirical study[J/OL]. 2019, arXiv: 1905.08941.

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

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

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, 硕士生, 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, 硕士生, 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.