文章编号：1003-0077(2015)04-0132-11

基于词向量预训练的不平衡文本情绪分类

林怀逸[1]，刘 箴[1]，柴玉梅[2]，刘婷婷[1]，柴艳杰[1]

（1. 宁波大学 信息科学与工程学院，浙江 宁波 315211；
2. 郑州大学 信息工程学院，河南 郑州 450001）

摘 要：深度学习中处理不平衡问题的方法多为代价敏感和采样。该文在词向量迁移的基础上提出预训练任务选择方法。用利于小类别区分的预训练词向量来初始化目标模型，并结合均衡过采样充分利用样本信息保持模型在大类别上的精度，使模型提取的文本特征在大小类别上具有公平性，从特征层面实现了平衡效果。实验结果表明，在文本情绪分类任务中，对比过采样方法，该方法在大部分无严重过拟合情况下有更好的平衡效果。当存在较严重过拟合时，该方法在目标分类数为三时平衡效果显著，并通过实验验证了预训练方法可与代价敏感方法相结合提升平衡性能。

关键词：不平衡分类；情绪分类；均衡过采样；预训练词向量

中图分类号：TP391 **文献标识码**：A

Imbalanced Emotion Classification Based on Word Vector Pre-training

LIN Huaiyi[1], LIU Zhen[1], CHAI Yumei[2], LIU Tingting[1], CHAI Yanjie[1]

(1. Faculty of Information Science & Technology, Ningbo University, Ningbo, Zhejiang 315211, China;
2. College of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China)

Abstract: The main methods to deal with imbalance problems in deep learning are focused on cost function and sampling technique. Based on word vector migration, this paper proposes a pre-training task selection method and initializes the target model with a pre-trained word vector that facilitates minority classes differentiation. Combined with balanced oversampling, the sample information is used to maintain the accuracy of the model in majority classes, so that the text features extracted by the model are balanced. Compared with the oversampling method, the experimental results show that the proposed method has a better balanced effect in most cases where text emotion classification result have no serious over-fitting. When there is a serious over-fitting, the method has a significant balance effect in three-type classification task. Experiments also verify that pre-training methods can be combined with cost-sensitive methods to improve the balance performance.

Keywords: imbalance classification; emotion classification; balanced over-sampling; word vector pretraining

0 引言

在社交媒体不断发展的大背景下，人们对媒体资讯发表个人看法，使用社交平台分享个人观点，利用社交工具排解心情，在个人网络空间倾倒苦水，这些内容均涵盖大量用户情感信息。从宏观角度，应用统计方法对网民评论和观点进行情感分析，能够监控和导向事件引起的舆情。从个人角度，对用户情感的掌握则能够有效了解用户心理健康状况并预测用户可能对社会或他人产生的影响。因此近年来情感分类任务受到广泛关注和研究。其中情绪分类是情感分类任务之一，目的是确定文本所带有的用户情绪表达。

当前许多研究报告所提出的深度学习方法在文本情绪分类任务上都卓有成效，但大都以类别数据

收稿日期：2018-10-11 定稿日期：2018-12-25

基金项目：NSFC—通用技术基础研究联合基金(U1636111)；国家自然科学基金(61761166005)；宁波市科技计划项目(2017C50018，2016D10016)

平衡为前提假设。实际应用中由于数据来源不同(采集的主题不同，平台不同等)，各类别样本之间往往存在数量分布不平衡的问题。深度学习中，数据不平衡使模型偏离预期，对大类别分类精度高，对小类别则分类精度低。

处理数据不平衡问题的目的是使模型接近其在数据平衡情况下的性能。主要解决方法包含代价敏感[1-3]、特征选择[4-5]以及数据采样，采样方法可细分为过采样原数据集中小类别样本[6-8]以及下采样大类别样本[9-12]，而特征选择主要应用于机器学习领域，通过选择使各类别相对公平的特征获得相对平衡的模型精度。

深度学习具有特征学习能力，因此常用的不平衡处理方法为代价敏感、采样及集成学习方法。例如，Wang[13]等提出平均错分误差损失函数及平均平方错分误差损失函数，使模型能够公平地从少类别中捕捉分类误差，殷昊[14]等结合集成学习和LSTM模型解决不平衡问题。

近年来预训练词向量的方法[15-17]被广泛应用于自然语言处理(NLP)任务中并取得良好效果[18]，说明词向量初始化能够影响模型特征的学习，从而改变模型精度，且研究表明针对任务微调词向量能够进一步提高模型精度[19]。因此本文通过预训练词向量的方法影响文本特征的选择，结合机器学习中特征选择思想，获得对各类别公平的文本特征实现模型精度的平衡。

1　相关工作

基于深度学习的方法被广泛应用于文本分类领域，常用模型为CNN(Convolutional Neural Network)和LSTM(Long Short-Term Memory)。研究表明，相比随机初始化，将Word2Vec方法预训练所得词向量对CNN模型初始化能够提高模型精度[18]。本文利用预训练词向量方法平衡模型在各分类上的精度，因此选择CNN模型作为分类模型来验证方法的有效性。

预训练的目的是获得能够提取普适特征的模型，并应用于各种不同任务的网络结构中。在自然语言领域，主要用于预训练语言模型，并应用于序列标注、文本分类、文本相似度、文本生成等各种任务中。预训练技术主要有以神经网络语言模型为基础的Word Embedding[16]技术，该技术广泛应用于各类自然语言任务中。ELMo(Embeddings from Language Models)[20]则利用双层双向LSTM构建的语言模型作为预训练目标解决一词多义问题，并在文本分类、阅读理解等6个自然语言处理任务中取得不同程度上的能力提升。同时，近期取得突破性进展的BERT(Bidirectional Encoder Representations from Transformers)[21]使用Transformers特征提取单元，构建双向语言模型结合预训练和微调二阶段框架，在11项自然语言任务中均取得目前最高精度。而本文认为对于存在数据不平衡的任务，在第二阶段，模型训练微调预训练模型之前，可针对性地对模型进行进一步预训练，使最终模型精度较为平衡。

深度学习中输入词向量与模型所有参数之间具有复杂的作用关系，Ignacio Cases[22]等证明了自然语言处理任务中，只要合理配置和优化模型，Word2Vec方法对比随机初始化及预训练添加语义信息的方法一定有更好的性能。在此基础上对词向量微调的研究包括Yang X等[23]提出一种有监督的词向量微调框架在无监督获得的词向量中额外添加有效信息，提高了模型在各自然语言处理任务中性能，另外提出结合多种窗口下的词向量以及词汇语义的微调方法，在情感相似度预测、类比推理以及完形填空任务中均取得更佳效果[24]，Uysal A K[25]等对比各种Word Embedding和在其上经过情感微调的词向量，发现使用后者的模型在IMDB、Sentiment 140以及Nine Public Sentiments三个情感相关任务中均有较高精度。上述研究说明针对任务微调Word2Vec所得词向量能够提高模型的分类精度。因此，本文通过微调词向量调整模型各类精度实现平衡效果。

关于如何微调词向量提高过拟合类别精度，研究表明词的表示依赖于所应用的任务[26]，Kim[18]和Pham D H[27]等发现情感任务中，模型训练后具有相似情感表达的词汇，在空间上的欧氏距离更小且相比静态词向量模型精度更高。说明情感任务中不同情感表达的词汇在词向量空间中的欧式距离越大，模型对这些情感的区分能力越强，精度越高。因此，本文对期望提高精度或加快拟合的类别进行预训练，通过预训练方法微调词向量，提高模型在这些类别上的精度。该微调词向量的方法相比其他方法更为简单，且在大部分情况下无需引入额外信息。

同时，本文方法结合均衡过采样，保留了大类别样本的所有信息，避免了下采样方法中由于摒弃大量有效信息而降低模型泛化能力的问题，维持了模

型在大类别上的精度，从而实现比过采样更好的平衡效果。

2 基于词向量预训练的不平衡情绪分类方法

2.1 情绪分类数据集

本文使用的数据整理自“自然语言处理与中文计算会议”(NLP&CC)情绪分析任务，样本数据分布情况如表1所示。其中分为无情绪、喜好、开心、惊讶、厌恶、悲伤、愤怒和恐惧八类情绪，另除无情绪类别外其他情绪可归为积极和消极两类。本文在该数据集上划分、采样形成多组存在不平衡问题的子数据集，用于验证不同情况下方法的有效性。

表1 情绪分类数据集中的样本数量分布

无情绪	积极			消极			
7 490	3 513			2 997			
	喜好	开心	惊讶	厌恶	悲伤	生气	恐惧
	1 879	1 310	324	1 239	1 018	595	145

2.2 词向量迁移与预训练任务选择方法

本文提出的词向量预训练方法流程如图1所示。其中，目标任务指实验既定的分类任务，预训练任务指在既定任务数据集中，使用本文提出的预训练任务选择方法选取部分数据进行的分类任务。预训练词向量指执行预训练任务后分类模型中的词向量矩阵。词向量的迁移指使用预训练词向量初始化目标任务的分类模型。整体流程为在既定任务中选择预训练任务并训练模型获得预训练词向量，该词向量再用于初始化目标任务模型，最终训练目标任务模型使其在各类别上分类精度平衡。

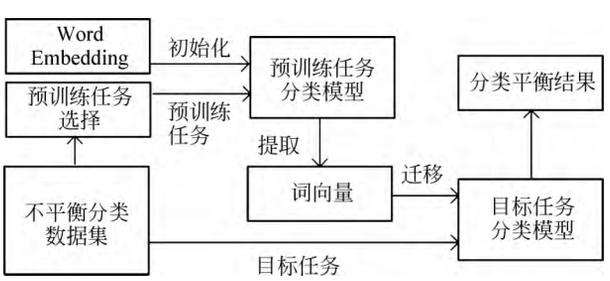

图1 词向量迁移流程图

由于特定模型预训练所得词向量矩阵在其他模型上不一定能达到期望的平衡效果，所以预训练任务和目标任务的分类模型均采用CNN。CNN模型结构如图2所示。

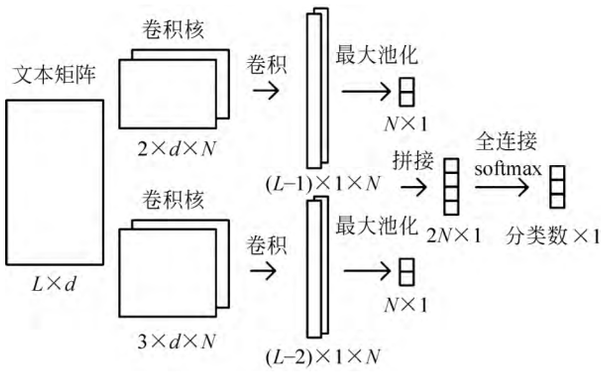

图2 文本CNN网络结构图

其中，文本矩阵由词的one-hot形式经过词向量矩阵映射得到。假设词向量表示为 $\boldsymbol{v}_n \in \mathbb{R}^d$，其中下标 n 表示文本中第 n 个词汇，d 表示词向量的维度，则文本矩阵由词向量按词序拼接组成表示为式(1)。

$$\boldsymbol{v}_{1:L} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \cdots, \boldsymbol{v}_L] \tag{1}$$

其中，L 表示文本固定长度。当实际文本长度大于 L 时，截断使其长度变为 L，当长度小于 L 时，使用表示未知词的词向量进行补齐。其中，未知词的词向量指各维度初始化为0的词向量，且该词向量在训练过程中由训练算法进行调整。获得文本矩阵后进行卷积操作。假设卷积核为 $w \in \mathbb{R}^{dx}$，其中 x 为卷积核宽度，该卷积核对 $\boldsymbol{v}_{i:i+x-1}$ 进行一次卷积操作获得特征值 c_i 表示为式(2)。

$$c_i = f(w * \boldsymbol{v}_{i:i+x-1} + b) \tag{2}$$

其中，$*$ 为对应元素乘积求和，b 为常数偏置项，f 表示非线性激活函数ReLU。所得 c_i 为文本第 i 个词起的一个x-gram特征[28]的特征值，再利用最大池化操作提取该文本最显著的x-gram特征，并与其他 n-gram特征拼接作为文本特征置于全连接层进行分类。

若使用Word2Vec方法对模型进行初始化。假设任务有三个类别 $\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3$，且各类别中样本均有相同语法结构，三个不同类别样本中最显著的x-gram文本区域为 $[\boldsymbol{v}_i, \cdots, \boldsymbol{c1\_v}_n, \cdots, \boldsymbol{v}_{i+x-1}]$，$[\boldsymbol{v}_j, \cdots, \boldsymbol{c2\_v}_n, \cdots, \boldsymbol{v}_{j+x-1}]$，$[\boldsymbol{v}_k, \cdots, \boldsymbol{c3\_v}_n, \cdots, \boldsymbol{v}_{k-x+1}]$ 其中对应位置词向量语法特性相同，且 $\boldsymbol{c1\_v}_n, \boldsymbol{c2\_v}_n, \boldsymbol{c3\_v}_n$ 为不同情感表达词汇，在训练所得词向量空间中，相似语法特性的词之间空间距离较近[29]，因此上述区域经过相同卷积核作用所得文本特征数值 c 相近，模型利用该文本特征不容易区分文本的情绪类别。但模型训练后 $\boldsymbol{c1\_v}_n$，

$c2\_v_n$，$c3_v_n$在空间上被分离，相同情绪表达的词向量在空间中距离相近，且相比静态词向量模型精度更高[18]。

因此，本文认为不同情绪表达词在空间中的距离影响模型提取的文本特征，而文本特征决定了模型在各情绪类别上的精度。对于过拟合程度较轻的情况，本文通过预训练方法微调情绪表达词在空间中的位置，使模型在训练开始时文本特征就能对过拟合类别之间有较好区分，提高过拟合类别的分类精度，并结合均衡过采样利用大类别样本数量优势，维持模型在大类别上的精度，实现模型精度的平衡。

对于严重过拟合情况，由于样本严重失衡，当模型对大类别拟合较好时，小类别由于过度训练产生严重过拟合[30]。因此，本文选择部分大类别数据作为预训练任务，加速模型对大类别的拟合，减少过拟合类别样本的重复训练次数，缓解过拟合现象，提高过拟合类别精度，实现平衡效果。

其中均衡过采样还避免了训练时大类别主导词向量分布的调整，严重破坏预训练词向量的分布，影响本文提出方法的预期效果。

本文首先对数据集分组，从最小类别开始将该类别样本数 3 倍以内的类别归入该分组，再从剩余类别中重复上述操作直至无剩余数据。此时认为分组间具有数据不平衡问题，而组内不平衡问题较弱。若所有类别样本被分至同一组，则缩小倍数重新分组。当降至 2 倍时仍然仅有单一分组则认为数据相对平衡，分组伪代码如表 2 所示。

表 2　不平衡分类数据分组方法

Input $ClassCount = [n_1, n_2, \cdots, n_c]$
Output $GroupList$
$Times = 3$
While $Times >= 2$
　$AscCount \leftarrow$ ASC($ClassCount$)
　Empty$GroupList$
　While len($AscCount$) > 0
　　Empty $Group$
　　$MinCount = AscCount.$firstElement
　　For $n_x \leftarrow$ Each *element* in $AscCount$
　　　If $n_x < Times * MinCount$ **then** Add n_x into $Group$
　　　else break
　　Add $Gourp$ into $GroupList$
　　Delete $n_1, \cdots, n_{x-1}$ in $AscCount$
　If len($GroupList$) ! $= 1$ **then** return $GroupList$ **end**
　else $Times \leftarrow Times - 1$; $AscCount \leftarrow$ ASC($ClassCount$)
return $GroupList$ **end**

分组后假设目标模型在各类别上的精度与类别样本数量呈正比，针对情绪分类数据中典型不平衡情况提出如下预训练任务选择方法。

当分组数为 2 时，样本不平衡导致的过拟合现象较轻微，认为模型分类精度不平衡的主要原因是模型在训练所得的词向量分布空间下提取的特征不利于小类别样本的区分。由于随机初始化模型在训练开始时所提取的特征对各类别的区分度相对公平，但样本数量不平衡导致训练后模型提取的特征更利于大类别的分类。因此我们选择小分组作为预训练任务，在初始化时使模型提取的特征更利于小类别之间或者与其他类别的区分，并通过均衡过采样较好保持词向量的分布，使模型最终提取的特征对各分类更为公平达到平衡模型精度的目的。若所选分组中只有单一类别时，采样大分组中最小类别使其与单一类别样本数相当，二者合并作为预训练任务；若目标任务为二分类任务，则收集或利用已有数据获取任务外其他类别样本，数量与小类别样本数相当后二者合并作为预训练任务。

当分组数大于 2 时，样本不平衡较为严重，认为此时小分组上的精度不平衡问题由过拟合所主导，严重降低模型在小类别上的泛化能力。而样本数居中的分组中，精度较低则由模型特征偏向所导致。因此通过下采样中间分组作为预训练任务的方法，加速中间类别拟合，减少小分组中样本重复训练，提高特征对中间类别的区分度，从而提高模型在中小类别上的精度。同时，为避免特征严重偏向，分组中各类别样本采样数与邻近小类别相当。若该分组类别数远小于其他分组或者分组数为偶数，则在全部类别中选取样本数量居中的部分类别同上下采样，作为预训练任务。类别数则由其余分组类别数均值决定，采样数量为所选类别的邻近小类别样本数。

上述方法均失效时，认为过拟合情况较为严重主导了所有中小分类的精度下降。因此，选择过拟合类别外样本数量上相邻的两个较大类别进行采样，使他们的样本数均与邻近的小类别样本数相当后作为预训练任务。通过该方法在初始化时形成更利于模型区分大类别样本的词向量分布，加速模型收敛，减少中小类别重复训练次数，降低过拟合程度。其中，为避免特征出现严重偏向仅选取两个相邻较大类别且进行采样。若过拟合分类外的类别数量为 1，首先采样该单一类别，再下采样各过拟合类别合并成新类，或者收集数据作为新类。最终，将大类别采样数据和新类合并作为预训练任务。

上述各方法中采样的目的包含平衡预训练中各类别样本数量，避免加剧过拟合和特征的严重偏向。而采样数量的选择取决于是否有严重过拟合类别数据参与预训练，若有则选取最小类别样本数作为采样数量。否则，选择邻近的小类别样本数，邻近小类别指样本数量小于各所选类别样本数的最大类别。

方法伪代码如表3所示。

表3 预训练选择方法

```
Input GroupList output from 表2结果
Output Pretrain
将 GroupList 中各 Group 升序排列，即 ASC(Group)
If len(GroupList) == 2
    # 取最小分组
    Pretrain ← min sample Group in GroupList
    If len(Pretrain) == 1
        # 下采样大分组中最小类别，加入预训练中
        #下采样数量为 Pretrain 中单一类别的样本数
        bgm ← USampling(min sample Class in big Group)
        Pretrain ← Pretrain + bgm
        # 若为二分类
        If Classes in GroupList == 2
            #收集任务外类别数据，数量与小类别样本数相当
            NewClass ← Collect Data without classes in GroupList
            Pretrain ← Pretrain + NewClass
If len(GroupList) > 2
    # 取中等数量分组
    MiddleGroupList ← middle sample Group(s) in GroupList
    If len(MiddleGroupList) == 1
        MiddleGroup ← Group in MiddleGroupList
        ClassNum ← len(MiddleGroup)
        #中间分组外其他分组中均平类别数
        ACNum ← AverageLen(No MiddleGroup Groups)
        If ClassNum == 1 or (ACNum - ClassNum) >3
            #下采样数量为邻近小类别样本数
            MiddleClasses ← USampling(Middle ACNum Classes)
            Pretrain ← MiddleClasses
        Else
            #下采样数量为邻近小类别样本数
            Pretrain ← USampling(MiddleGroup)
    Else
        ACNum ← AverageLen(Groups in GroupList)
        #下采样数量为邻近小类别样本数
        MiddleClasses ← USampling(Middle ACNum Classes)
        Pretrain ← MiddleClasses
When serious overfitting
    OFC ← overfitting classes
    LC ← ASC(AllClasses - OFC)
    If len(LC) == 1 # 过拟合类别外仅有一个类别
        #下采样数量与 OFC 中最小类样本数相当
        #收集样本数与邻近小类别相当
        NewClass ← USampling(Each Class in OFC)
                   orCollect Data without classes in OFC
        #下采样数量与 NewClass 样本数相当
        Pretrain ← USampling(LC[0]) + NewClass
    Else
        #下采样数量为邻近小类别样本数
        Pretrain ← USampling(Each class in LC[0:2])
return Pretrain
```

3 实验

3.1 实验内容

实验主要探究情绪分类任务中，本文方法在所涵盖情况中的有效性和平衡性能，并在具有平衡效果的各处理方式下，分别对比了代价敏感和集成学习方法的平衡性能。

情况1：分组数为2，大分组由单一类别构成。

情况2：分组数为2，小分组由单一类别构成。

情况3：分组结果中分组数为2，且各分组中类别数不止一个。

情况4：不平衡2分类。

情况5：分组结果中分组数大于2。

情况6：任务出现严重过拟合现象。

情况7：在情况6的基础上，目标任务中除过拟合类别外仅剩一个类别。

根据所要考察情况共设置8组实验，分别为：

（1）无情绪、积极和消极情绪分类，目的在于验证情况1下方法的有效性，并探究不同精度预训练模型对目标模型平衡性能的影响。

（2）厌恶、悲伤和惊讶情绪分类，目的在于验证情况2下方法的有效性。并与代价敏感和集成学习在平衡性能上进行对比，同时添加预训练与代价敏感相结合的实验，说明本文方法在代价敏感方法上同样具有进一步提升平衡性能的可能。

（3）从愤怒、悲伤、厌恶、开心和喜好情绪开始，验证情况3下方法的有效性。添加无情绪类别并逐

步增加样本数量，探究大类别样本数的增加对方法平衡性能的影响。目标任务中愤怒、悲伤和厌恶情绪类别样本数分别为 321，319 和 393，采样自表 1 数据集，目的在于构造情况 3。

(4) 无情绪和愤怒情绪的不平衡情绪二分类，目的在于验证情况 4 下方法的有效性。

(5) 愤怒、悲伤、厌恶、开心、喜好和无情绪分类，其中愤怒、悲伤和厌恶情绪类别数量分别为 321，319 和 393，采样自表 1 数据集，构造情况 5 并验证方法有效性。并与代价敏感和集成学习对比平衡性能。

(6) 愤怒、恐惧、悲伤和惊讶情绪分类，分别设置三组实验，探究严重过拟合时，情况 3 处理方法在目标模型上的平衡性能，并验证情况 6 下方法的有效性。

(7) 在实验(6)的基础上添加开心情绪类别，进一步探讨情况 6 下方法的平衡性能。

(8) 愤怒、惊讶和恐惧情绪分类，该组实验在满足情况 1 的条件下出现严重过拟合。首先，探究情况 1 方法的平衡性能，并验证情况 7 下方法的有效性，且对比了有无均衡过采样的情况，探究了均衡过采样对预训练的影响。最后还对比了代价敏感和集成学习的平衡性能。

实验中代价敏感使用代价敏感矩阵方法。矩阵数值根据各类别样本数量比例确定，集成学习则根据最小类别样本数 m，下采样大类别样本获取平衡数据。假设最大类别样本数为 n，则获得 $\lceil n/m \rceil$ 组数据分别训练子模型，二级模型则拼接各子模型特征，后接全连接层输出类别概率分布。

实验中测试集各类别样本数为最小类别样本数量的 10%。实验(1)测试集各类别样本数均为 299 个；实验 (2)中为 32 个；实验 (3)(4)(5)为 59 个；实验(6)(7)(8)中为 14 个。

实验中预训练任务使用的词向量由实验数据集使用 Python 中的 Gensim 工具包进行训练所得，模型为 CBOW(Continuous bag-of-words)，词向量维度为 128。所用情绪分类模型 CNN 的参数设置如表 4 所示。

表 4　CNN 模型参数表

参数项	参数值
优化算法	Adam
学习率	0.001
文本矩阵宽度	50
卷积核宽度	2,3,4,5
各宽度卷积核数	100
Dropout 概率	0.5

为缓解过拟合问题，对拼接后的文本特征层进行 dropout 操作，并使用 L2 正则化方法选取全连接层权值矩阵和偏置作为模型复杂程度的衡量，衡量指标为 $\frac{1}{2}\|w\|_2^2$，将二者指标求均值加入到模型的损失中。

3.2　实验评价指标

由于研究内容为数据不平衡问题，因此主要关注模型各类别上的详细评价及综合评价，所以，使用 P (Precision，准确率)、R (Recall，召回率)以及 $F1$ (F1-score)作为评价指标。假设 TP 为当前类别样本被正确分类的数量，FP 表示其他类别样本被划分到当前类的数量，FN 表示将当前类别样本被错分到其他类的数量。

准确率表示分到当前类别的样本中确实属于当前类的样本数所占的比例，如式(3)所示。

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{3}$$

召回率表示对当前类别样本进行正确分类的概率，如式(4)所示。

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{4}$$

$F1$ 根据准确率和召回率对模型在该分类上的性能做出综合评价，如式(5)所示。

$$F1 = \frac{2\times P\times R}{P+R} \tag{5}$$

3.3　实验结果与分析

实验表中 Word2Vec 指在所有数据集上使用 Word2Vec 方法训练获得词向量矩阵。表中第一行括号内容表示预训练任务类别。各类别名称后括号中的分数表示下采样比例，百分数则表示 R 值，整数值表示类别样本数。每组评价指标(P、R、$F1$)各为一组实验数据。

实验(1)中无情绪类别为大分组中单一类别。如表 5 所示，第二组实验中积极和消极 R 值分别提

高2.54%和2.74%，第三组分别提高1.74%和4.88%。但相比第一组实验，由于无情绪类别R值在第二、三组中分别下降了7.56%、11.50%，影响了消极情绪的P值，模型平均$F1$分别下降了0.63%和1.52%，因此模型虽然提升了过拟合类别的精度，使最大类别R值差缩小10.1%，但平衡作用一般。另外对比二、三两组实验发现R值更为平衡的预训练模型平衡性能更优。

表5 无情绪、积极和消极情绪分类实验结果(五次实验平均)

情绪分类	均衡过采样+Word2Vec			均衡过采样+Pretrain(积极77%，消极78%)			均衡过采样+Pretrain(积极81%，消极72%)		
	P/%	R/%	$F1$/%	P/%	R/%	$F1$/%	P/%	R/%	$F1$/%
积极 3 214	64.61	56.59	60.24	64.89	**59.13**	**61.85**	63.40	**58.33**	**60.72**
无情绪 7 191	57.15	73.71	64.34	57.23	66.15	61.35	57.03	62.21	59.35
消极 2 698	71.05	59.00	64.38	66.22	**61.74**	63.88	65.04	**63.88**	64.37
平均	64.27	63.10	62.99	62.78	62.34	62.36	61.82	61.47	61.47

实验(2)中惊讶情绪为小分组中单一类别。结果如表6所示，情况2下本文方法能够提高参与预训练的类别的R值，但与实验前提假设不符，类别R值与类别数量不呈正比，预训练使悲伤情绪类别R值进一步提高4.18%，最大R值差提高了1.05%。对比代价敏感矩阵的方法则平均R值相差0.7%，平均$F1$相差1.33%。但本文认为预训练方法同样可作用于代价敏感方法提高平衡性能，二者结合第4组实验发现平均R值进一步提高了1.39%，平均$F1$提高了1.36%。本文不对预训练与代价敏感结合的方法进行探讨，故下文不再对二者结合进行实验。最后，集成学习方法在该组数据中的平衡性能与均衡过采样相当。

表6 厌恶、悲伤和惊讶分类实验结果(三次实验平均)

情绪分类	均衡过采样+Word2Vec			均衡过采样+Pretrain(预训练惊讶和悲伤(1/5))			代价敏感			代价敏感+Pretrain(预训练惊讶和悲伤(1/5))			集成学习(5子模型)		
	P/%	R/%	$F1$/%	P/%	R/%	$F1$/%	P/%	R/%	$F1$/%	P/%	R/%	$F1$/%	P/%	R/%	$F1$/%
厌恶:1 207	63.09	65.62	64.31	63.07	60.42	61.65	65.04	65.62	64.17	**65.10**	58.33	61.41	56.38	65.62	60.58
悲伤:986	58.38	72.91	64.74	61.05	**77.09**	**67.89**	63.08	63.54	62.15	**69.66**	60.42	**64.30**	65.24	63.54	64.12
惊讶:292	63.81	44.79	52.14	61.91	**47.92**	**53.24**	63.21	58.33	60.45	59.08	**72.92**	**65.11**	62.28	53.12	57.31
平均	61.76	61.11	60.39	**62.01**	**61.80**	**60.92**	63.77	62.50	62.25	**64.61**	**63.89**	**63.61**	61.30	60.77	60.67

实验(3)中愤怒、悲伤和厌恶为小分组，结果如表7所示。四组实验的平均R值变化为+0.9%，+0.85%，−0.28%，+0.66%，平均$F1$变化为+1.38%，+1.23%，−0.53%，+0.62%。无严重过拟合情况下平衡性能较好。但随过拟合加剧平衡性能衰退。

实验(4)中新类由恐惧和惊讶情绪样本组成，结果如表8所示，愤怒情绪R值提高了7.80%，模型平均R值提高了1.36%，且平均$F1$提升了1.46%，说明引入额外数据同过拟合类别预训练的方法能够有效平衡情况4下模型精度。

表7 类别逐渐增加的不平衡情绪分类实验(三次实验平均)

情绪分类	均衡过采样+Word2Vec			均衡过采样+Pretrain(愤怒、悲伤和厌恶)		
	P/%	R/%	$F1$/%	P/%	R/%	$F1$/%
愤怒 321	58.73	44.63	50.03	55.48	41.24	47.08
悲伤 319	71.77	33.33	45.34	58.70	**46.89**	**51.49**
厌恶 393	43.57	32.77	35.92	**46.95**	**38.42**	**42.05**
开心 1251	37.48	64.97	47.32	**41.18**	59.32	**48.61**
喜好 1820	51.22	63.28	56.44	48.73	57.63	52.70
平均	52.55	47.80	47.01	50.21	**48.70**	**48.39**

续表

情绪分类	均衡过采样＋Word2Vec			均衡过采样＋Pretrain（愤怒、悲伤和厌恶）		
	P/%	R/%	$F1$/%	P/%	R/%	$F1$/%
愤怒 321	45.54	35.03	39.53	**46.66**	**36.16**	**40.19**
悲伤 319	59.75	37.29	45.29	55.75	**42.37**	**48.12**
厌恶 393	34.86	22.03	26.83	**42.92**	**35.03**	**38.41**
开心 1251	36.28	45.20	40.16	**40.88**	35.03	37.56
喜好 1820	48.29	57.06	52.22	42.45	**58.19**	48.65
无情绪 1061	36.59	55.93	44.14	**36.72**	50.85	42.59
平均	43.55	42.09	41.36	**44.23**	**42.94**	**42.59**
愤怒 321	55.55	39.55	45.96	48.65	**40.68**	44.02
悲伤 319	66.51	37.29	46.82	59.35	**42.94**	**49.32**
厌恶 393	43.96	27.12	30.71	35.08	22.60	27.35
开心 1251	37.49	48.02	41.95	**41.72**	38.42	39.70
喜好 1820	46.39	49.72	47.87	45.16	49.15	46.62
无情绪 2123	33.98	54.80	40.99	**34.64**	**61.02**	**44.11**
平均	47.31	42.75	42.38	44.10	42.47	41.85
愤怒 321	51.69	41.24	43.62	49.88	36.16	41.37
悲伤 319	49.29	36.72	41.94	57.19	**41.24**	**47.35**
厌恶 393	38.26	30.51	33.82	37.66	27.12	30.74
开心 1251	37.42	36.16	36.36	37.02	**40.11**	**38.30**
喜好 1820	47.79	49.72	48.54	48.12	**50.85**	**49.20**
无情绪 3184	37.53	58.19	45.45	37.62	**61.02**	**46.47**
平均	43.66	42.09	41.62	**44.58**	**42.75**	**42.24**

表 8　二分类不平衡情绪分类实验结果（五次实验平均）

情绪分类	均衡过采样＋Word2Vec			均衡过采样＋Pretrain（愤怒、新类）		
	P/%	R/%	$F1$/%	P/%	R/%	$F1$/%
无情绪 7 431	76.36	88.14	81.75	**81.13**	83.05	**81.90**
愤怒 536	86.18	72.54	78.64	82.87	**80.34**	**81.41**
平均	81.27	80.34	80.19	**82.00**	**81.70**	**81.65**

实验(5)中开心和喜好为中间分组，结果如表 9 所示。厌恶情绪作为最小类别 R 值提高约 6.21%，部分类别 R 值下滑 1.13%至 5.09%不等。但最大类间 R 值差缩小了 7.34%，且平均 R 值和平均 $F1$ 仅分别下降 0.28%和 0.07%，平衡效果一般。而代价敏感方法和集成学习方法的平均 R 值比均衡过采样方法分别高 1.69%和 2.54%，$F1$ 分别高 2.36%和 2.45%。

实验(6)结果如表 10 所示。第二组数据使用情况 3 处理方法，加剧了恐惧和惊讶类别的过拟合现象。因此，第三组实验以情况 6 方法将愤怒和悲伤样本分别下采样后作为预训练任务，结果恐惧类别 R 值提高 5.71%。但惊讶类别 R 值下降了 2.86%，模型平均 R 值下降了 1.78%，平均 $F1$ 下降了 1.88%，无平衡效果。

实验(7)在实验(6)基础上添加开心情绪类别，结果如表 11 所示。加剧了惊讶和恐惧类别的过拟合，本文方法已无平衡效果。

表 9　三分组的不平衡情绪分类实验结果（三次实验平均）

情绪分类	均衡过采样＋Word2Vec			均衡过采样＋Pretrain（开心和喜好）			代价敏感			集成学习（24 子模型）		
	P/%	R/%	$F1$/%	P/%	R/%	$F1$/%	P/%	R/%	$F1$/%	P/%	R/%	$F1$/%
愤怒:321	44.26	37.85	40.66	**48.34**	36.16	**42.93**	**48.90**	**45.76**	**47.14**	**46.56**	37.29	**41.13**
悲伤:319	55.13	39.55	44.88	53.57	34.46	42.92	51.30	**48.02**	**49.40**	50.60	**54.80**	**52.44**
厌恶:393	37.54	27.12	31.43	**40.83**	**33.33**	**34.50**	32.55	**33.90**	**33.16**	**38.50**	**37.29**	**37.85**
开心:1251	42.19	42.37	42.05	39.06	40.68	41.07	41.13	**42.94**	41.97	**48.12**	35.03	40.34
喜好:1820	49.56	52.54	50.84	49.22	**54.24**	**51.18**	**50.06**	**54.80**	**51.47**	**52.75**	43.50	47.38
无情绪:7431	38.58	60.45	46.68	**40.29**	59.32	43.52	**50.98**	44.63	**47.57**	**42.59**	**67.23**	**52.10**
平均	44.54	43.32	42.76	**45.22**	43.04	42.69	**45.82**	**45.01**	**45.12**	**46.52**	**45.86**	**45.21**

表10 愤怒、恐惧、悲伤和惊讶分类实验结果(五次实验平均)

情绪分类	均衡过采样+Word2Vec			均衡过采样+Pretrain(惊讶和恐惧)			均衡过采样+Pretrain(愤怒(3/5)和悲伤(1/3))		
	P/%	R/%	F1/%	P/%	R/%	F1/%	P/%	R/%	F1/%
愤怒 581	51.41	68.57	58.40	53.82	**71.43**	61.34	51.25	67.14	57.82
恐惧 131	64.92	30.00	36.99	50.95	20.00	28.40	36.35	**35.71**	34.51
悲伤 1004	52.53	72.86	60.42	45.64	**77.14**	57.20	55.14	64.29	57.56
惊讶 324	66.42	41.43	48.47	56.78	32.86	40.55	62.46	38.57	46.87
平均	58.82	53.21	51.07	51.80	50.36	46.87	51.30	51.43	49.19

表11 开心、愤怒、恐惧、悲伤和惊讶分类实验结果(五次实验平均)

情绪分类	均衡过采样+Word2Vec			均衡过采样+Pretrain(愤怒3/5和悲伤1/3)		
	P/%	R/%	F1/%	P/%	R/%	F1/%
愤怒 581	63.05	81.43	70.91	59.33	85.71	69.98
恐惧 131	33.33	5.72	9.71	46.67	4.28	7.68
悲伤 1 004	38.28	70.00	49.44	41.04	72.86	52.43
开心 1 296	58.17	81.43	67.71	58.21	81.43	67.88
惊讶 310	66.09	24.29	35.25	60.67	17.15	26.49
平均	51.78	52.57	46.60	53.18	52.29	44.89

实验(8)结果如表12所示,一至四组实验表明代价敏感在平衡性能上优于均衡过采样和集成学习,但在大类别分类上的R值较为有限。原因是由类别数量确定的代价敏感矩阵弱化了大类别的错分代价使模型更侧重于小类别样本的判断。同时,结合实验(2)和实验(5)可以发现集成学习在样本数量较少的情况下性能较为有限。第五组实验利用情况1处理方法,恐惧类别R值仅提升1.43%,但加剧了惊讶类别过拟合,R值下降10%。因此,第六组实验使用情况7处理方法将恐惧和惊讶数据分别采样合并为新类与愤怒类别采样数据合并作为预训练任务。结果表明相比第四组实验目标模型在恐惧和惊讶情绪上的R值分别提升了2.86%和4.28%,模型平均R值提升了约4.29%。同时,平均$F1$提高了4.37%。最后,对比第六、第七两组实验,说明在无均衡过采样的情况下,大类别将主导词向量分布的调整严重破坏预训练词向量的分布,使单独使用预训练词向量方法无法取得预期效果,且加剧了模型R值的偏向。

表12 愤怒、惊讶和恐惧分类实验结果(五次实验平均)

情绪分类	Word2Vec			代价敏感			集成学习(5子模型)			均衡过采样+word2vec		
	P/%	R/%	F1/%	P/%	R/%	F1/%	P/%	R/%	F1/%	P/%	R/%	F1/%
愤怒:581	51.79	81.43	63.21	**68.45**	62.86	**65.02**	56.77	70.00	62.58	55.59	75.71	63.61
恐惧:131	66.64	42.86	51.72	56.30	**77.14**	**64.53**	61.75	58.57	59.86	62.00	57.14	58.57
惊讶:310	53.70	41.43	46.60	**71.31**	**48.57**	**56.86**	51.11	41.43	45.52	60.22	42.86	49.60
平均	57.38	55.24	53.84	**65.36**	**62.86**	**62.14**	56.54	56.66	55.98	59.27	58.57	57.26

表12续 愤怒、惊讶和恐惧分类实验结果(五次实验平均)

情绪分类	均衡过采样+Pretrain(惊讶和恐惧)			均衡过采样+Pretrain(愤怒(1/3)和合并类(184个))			Pretrain(愤怒(1/3)和合并类(184个))		
	P/%	R/%	F1/%	P/%	R/%	F1/%	P/%	R/%	F1/%
愤怒:581	56.49	82.86	66.59	**67.60**	**81.43**	**72.98**	56.03	87.14	67.71
恐惧:131	60.37	58.57	58.19	60.63	**60.00**	**58.62**	67.00	21.43	29.66
惊讶:310	71.29	32.86	43.03	**68.69**	**47.14**	**53.30**	60.17	61.43	59.09
平均	62.72	58.09	55.94	**65.64**	**62.86**	**61.63**	61.06	56.67	52.15

4 结论

本文在深度学习中从特征层面将词向量预训练和均衡过采样相结合，提供了一种解决不平衡问题的简单有效的思路。实验证明在大部分非严重过拟合的任务中，该方法能够改变模型在各类别上精度实现模型精度平衡的效果。实验说明提出的方法，能够使模型提取的文本特征更具公平性，提升模型在过拟合类别上的精度，获得各类别精度较平衡的单一模型。因此，还可与随机初始化以及Word2Vec方法初始化的模型结合应用于集成学习框架上。同时，本文还验证了该方法与代价敏感方法结合进一步提高平衡性的可能，但限于工作进展程度，未能进一步探讨结合方式。同时，本文以模型各类别上精度与对应类别样本数量呈正比为前提假设，但实际情况中还与模型特征提取能力相关，是否能在考虑模型对各类别分类能力的同时选择合理的预训练任务还有待进一步研究解决。此外，本文方法仅提供一种新的解决平衡问题的思路，由于研究水平有限，方法适用范围无法定量测量，涵盖情况也较为有限，有待进一步扩展和一般化。

参考文献

[1] Zhang C, Wang G, Zhou Y, et al. A new approach for imbalanced data classification based on minimize loss learning[C]//Proceedings of IEEE 2nd International Conference on Data Science in Cyberspace. 2017: 82-87.

[2] Mirza B, Kok S, Lin Z, et al. Efficient representation learning for high-dimensional imbalance data[C]//Proceedings of IEEE International Conference on Digital Signal Processing. 2017.

[3] Bian J, Peng X G, Wang Y, et al. An efficient cost-sensitive feature selection using chaos genetic algorithm for class imbalance problem[J]. Mathematical Problems in Engineering, 2016, 2016(6): 1-9.

[4] 张延祥，潘海侠. 一种基于区分能力的多类不平衡文本分类特征选择方法[J]. 中文信息学报，2015，29(4)：111-119.

[5] Moayedikia A, Ong K L, Boo Y L, et al. Feature selection for high dimensional imbalanced class data using harmony search[J]. Engineering Applications of Artificial Intelligence, 2017, 57(C): 38-49.

[6] Kubat M, Matwin S. Addressing the Curse of imbalanced training sets: One sided selection[C]//Proceedings of the 14th International Conference on Machine Learning, Nashville, Tennesse. Morgan Kaufmann. 1997: 179-186.

[7] Japkowicz N. The Class Imbalance Problem: Significance and strategies[C]//Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning Las Vegas, Nevada. 2000: 111-117.

[8] 胡峰，王蕾，周耀. 基于三支决策的不平衡数据过采样方法[J]. 电子学报，2018，46(1)：135-144.

[9] Lewis D D, Catlett J. Heterogenous uncertainty sampling for supervised learning[C]//Proceedings of Eleventh International Conference on International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 1994: 148-156.

[10] Ling C X. Data mining for direct marketing: problems and solutions[C]//Proceedings of International Conference on Knowledge Discovery & Data Mining. AAAI Press, 1998: 217-225.

[11] 王中卿，李寿山，朱巧明，等. 基于不平衡数据的中文情感分类[J]. 中文信息学报，2012，26(3)：33-38.

[12] 熊冰妍，王国胤，邓维斌. 基于样本权重的不平衡数据欠抽样方法[J]. 计算机研究与发展，2016，53(11)：2613-2622.

[13] Wang S, Liu W, Wu J, et al. Training deep neural networks on imbalanced data sets[C]//Proceedings of International Joint Conference on Neural Networks. IEEE, 2016: 4368-4374.

[14] 殷昊，李寿山，贡正仙，等. 基于多通道LSTM的不平衡情绪分类方法[J]. 中文信息学报，2018，32(1)：139-145.

[15] Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning[C]//Proceedings of International Conference on Machine Learning. ACM, 2008: 160-167.

[16] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of International Conference on Neural Information Processing Systems. Curran Associates Inc. 2013: 3111-3119.

[17] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation [C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. 2014: 1532-1543.

[18] Kim Y. Convolutional neural networks for sentence classification [C]//Proceedings of EMNLP2014,

2014：1746-1751.

[19] Mikolov T，Grave E，Bojanowski P，et al. Advances in Pre-training distributed word representations[C]// Proceedings of Proceedings of the International Conference on Language Resources and Evaluation，2018.

[20] Matthew E Petersy，Neumanny M，Iyyery M，et al. Deep contextualized word representations[J]. arXiv. arXiv：1802.05365，2018.

[21] Devlin J，Chang M，Lee K，et al. BERT：Pre-training of deep bidirectional transformers for language understanding[J]. arXiv. arXiv：1810.04805，2018.

[22] Ignacio Cases，Minh-Thang Luong，Christopher Potts. On the effective use of pretraining for natural language inference[J]. arXiv. arXiv：1710.02076，2017.

[23] Yang X，Mao K. Supervised fine tuning for word embedding with integrated knowledge[J]. arXiv. arXiv：1505.07931，2015.

[24] Yang X，Mao K，et al. Task independent fine tuning for word embeddings[J]. IEEE/ACM Transactions on Audio Speech & Language Processing，2017，25(4)：885-894.

[25] Uysal A K，Yi L M. Sentiment Classification：Feature selection based approaches versus deep learning[C]//Proceedings of IEEE International Conference on Computer and Information Technology. IEEE，2017：23-30.

[26] Labutov I，Lipson H. Re-embedding words[C]//Proceedings of Meeting of the Association for Computational Linguistics. 2013：489-493.

[27] Pham D H，Nguyen T T T，Le A C. Fine-Tuning word embeddings for aspect-based sentiment analysis[C]//Proceedings of TSD2017：Text，Speech，and Dialogue 2017：500-508.

[28] Kalchbrenner N，Grefenstette E，Blunsom P. A convolutional neural network for modelling sentences[C]//Proceedings of ACL，Baltimore and USA，2014.

[29] Mikolov T，Chen K，Corrado G，et al. Efficient estimation of word representations in vector space[C]// Proceedings of In ICLR，2013.

[30] He H，Garcia E A. Learning from imbalanced aata[J]. IEEE Transactions on Knowledge & Data Engineering，2009，21(9)：1263-1284.

林怀逸(1990—)，**硕士，主要研究领域为自然语言处理。**

E-mail：1475168072@qq.com

刘箴(1965—)，**通信作者，博士，研究员，主要研究领域为人工智能和虚拟现实。**

E-mail：liuzhen@nbu.edu.cn

柴玉梅(1964—)，**硕士，教授，主要研究领域为机器学习、自然语言处理。**

E-mail：ieymchai@zzu.edu.cn