doi: 10.3969/j.issn.0490-6756.2020.01.011

一种解决命名实体识别数据集类别标记失衡的方法

许丽丹[1], 刘嘉勇[1], 何 祥[2]
(1.四川大学网络空间安全学院, 成都 610065; 2.四川大学电子信息学院, 成都 610065)

摘 要: 命名实体识别研究中常见的公开数据集普遍存在数据类别标记不平衡的问题,限制了基于统计学习模型方法性能的进一步提高.针对上述问题,提出了基于遗传算法的数据类别标记平衡方法.该方法基于原始数据集中已有的标记数据,通过修改遗传算法中的指标适应度函数和基因组合规则,合成类别分布均衡的文本用以扩充原始数据集,降低标记数据不平衡性从而改善命名实体识别的效果.为验证该方法的有效性,采用 Bi-LSTM-CRF 模型分别基于 CoNLL 2003 及 JNLPBA 数据集设计了该方法与平衡欠采样、随机过采样方法的对比实验.从实验中发现,提出的方法在 CoNLL2003 数据集上模型召回率提高 3.26%, F_1 值提高 1.70%;在 JNLPBA 数据集上召回率提高 2.44%, F_1 值提高 1.03%.实验结果表明,提出的方法能够有效地缓解类别标记失衡问题达到提高命名实体识别效果的目的.

关键词: 命名实体识别; 类别失衡; 数据合成; 统计学习模型; 遗传算法;
中图分类号: TP391 **文献标识码**: A **文章编号**: 0490-6756(2020)01-0082-07

Method for solving class imbalance of named entity recognition dataset

XU Li-Dan[1], *LIU Jia-Yong*[1], *HE Xiang*[2]
(1. College of Cybersecurity, Sichuan University, Chengdu 610065, China;
2. College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: The public data sets in named entity recognition research are often class label imbalanced, which limits the further performance improvement based on statistical learning model methods. Aiming at the above problems, a data class label balancing method based on genetic algorithm is proposed, which modifies the fitness function and gene combination rules tried to balance the dataset by generating new samples to augment the original dataset. In order to verify the validity, the proposed method was compared with the balanced undersampling method and the random oversampling method by using the Bi-LSTM-CRF model on the CoNLL 2003 and JNLPBA datasets respectively. The results show that the proposed method increased the recall rate by 3.26% and the F_1 value by 1.70% on the CoNLL2003 dataset, and the recall rate by 2.44% and the F_1 value by 1.03% on the JNLPBA dataset. The experimental results demonstrate that the proposed method can effectively alleviate the class imbalance and improves the effect of named entity recognition.

Keywords: Named entity recognition; Class imbalance; Data synthesis; Statistical learning model; Genetic algorithm

收稿日期: 2019-05-11
基金项目: 中国科学院网络测评技术重点实验室开放课题基金"面向非结构化数据的威胁情报知识图谱构建"(NST-18-001)
作者简介: 许丽丹(1994-), 女, 硕士生, 研究方向为网络数据分析与信息安全. E-mail: xulidandan@163.com
通讯作者: 何祥. E-mail: rivsearcher@gmail.com

1　引　言

命名实体识别是自然语言处理领域中一项重要任务，经过几十年的研究发展已取得了显著成果[1-5]. 但是目前已有的研究较少关注命名实体识别数据集数据类别标记不平衡这类数据分布问题[6]. 数据类别标记不平衡[7]是指一个类别的数据量与属于另一类别的数据量差距较大，且小样本量类呈现出的信息更具价值. 数据类别标记不平衡会影响统计学习模型的效果，导致模型更关注多数类别数据，忽略少数类别数据[7-16]. 在分类问题中解决数据类别标记不平衡的方法主要有以下三类：(1) 采样方法，包括欠采样和过采样方法[7]；(2) 改进统计学习算法，包括 one-class 学习方法[8]，修改算法的代价[9]以及集成学习[10]；(3) 特征选择，通过收集最佳特征子集[11]以实现最佳性能.

针对序列数据，Tomanek 等人通过主动学习，在数据标记阶段尽可能平衡地标记数据[12]. Douzas 等人使用条件对抗生成模型为少数类别数据生成更多数据以缓解数据失衡[13]，但这种方法模型复杂、计算代价大. Gliozzo 假设频繁出现的单词一般不提供文本具体信息，也不会是实体. 可以过滤掉这些词以减小非实体类单词与实体类单词比值[14]. Maragoudakis 等人使用 Tomek 连接法在训练阶段减少训练集中不必要的负面样本[15]. Akkasi 等人[6]提出了平衡欠采样方法，该方法保留句子逻辑短语结构以及句子间相关性，并在四个生物医学数据集上分别实验比较随机欠采样、SWF[14]和平衡欠采样三种欠采样方法的效果并证明了调节数据集中实体类单词和实体类单词比例可以改善模型效果. 然而这种方法直接剔除数据集中非实体单词或语句，可能破坏文本的短语结构，丢失有价值的数据.

针对已有方法可能造成的信息丢失问题，本文通过改造遗传算法[17]用于序列文本合成，提出了一种基于遗传算法的数据类别标记平衡方法，本文简称 CBM-GA(class balance method based on genetic algorithm). 该方法保留所有原始语料，利用遗传算法基因重组、繁衍行为特点，充分挖掘类别平衡文本特征，尽可能维持语句中实体短语结构的同时，合成新文本以扩充原始数据集. 实验表明，本文提出的方法有效缓解数据集类别标记不平衡问题，提高命名实体识别任务性能.

2　模型架构

2.1　类别平衡方法框架

本文提出的数据类别平衡方法 CBM-GA 框架如图 1 所示，首先对原始数据集进行筛选，获得合适的父代样本集，通过适应度函数评估样本并按照其适应度函数值排序获得有序父代样本集. 从有序父代样本集中随机选择样本形成父代样本对，经过交叉和变异操作生成新样本，再次使用适应度函数评估新样本，抽取适应度函数值高的新样本集合更新下一轮文本合成操作所需的父代样本集，如此循环 N 次. 最后将第 N 轮生成新样本数据集和原始数据集合并产生扩充数据集，用于命名实体识别.

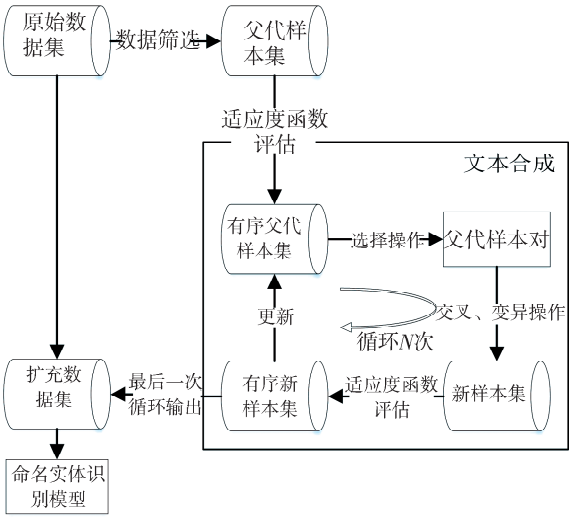

图 1　CBM-GA 框架图
Fig. 1　Method framework

2.2　类别平衡方法原理

2.2.1　*数据筛选*　现有用于命名实体识别研究的公开数据集中实体单词个数远小于非实体单词个数，且包含大量无实体单词、类别严重失衡的语句[6]. 直接使用这样的数据集进行文本合成会增加后续操作代价，降低文本合成效果. 因此需要进行文本筛选，初步获取类别分布相对平衡的文本集合作为父代样本集.

如何评估文本类别标记平衡的程度是首先需要解决的问题，Akkasi 等人[6]用非实体单词个数和实体单词个数比值来评估文本类别标记平衡性，如式(1)所示.

$$R=\frac{W_O}{W_E} \tag{1}$$

式(1)中，W_O 表示语句中非实体单词个数；分母

W_E表示语句实体单词个数. 本文将 R 值作为文本平衡性的衡量指标,R 值越大文本类别越不平衡.

统计分析常用数据集 CoNLL2003[18]、JNLPBA[19]中语句的非实体单词个数与实体单词个数比值 R 和语句单词总数的关系,我们可以得出以下结论:文本单词个数越多,R 值越大,文本类别越不平衡,两者近似反比例关系,如图 2 所示.

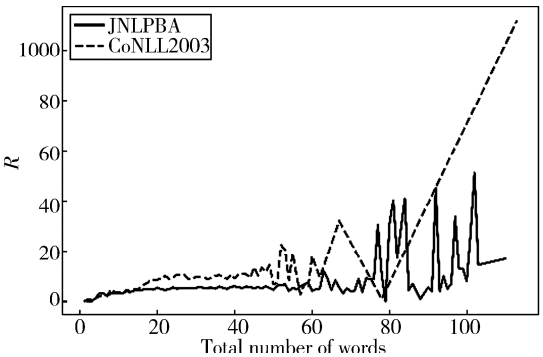

图 2 两个常用数据集文本 R 值与文本单词个数关系图

Fig. 2 Relationship between the text R value of two common datasets and the number of text words

图 2 中横轴表示语句的单词总数,纵轴表示语句中非实体单词个数与实体单词个数的比值. 实线、虚线分别表示 JNLPBA、CoNLL2003 数据集的文本 R 值与文本长度的关系,二者整体呈上升趋势. 当文本单词个数超过 60,文本的 R 值波动幅度变大;当文本单词个数超过 100,JNLPBA 数据集的 R 值均处于剧烈波动状态,CoNLL2003 数据集 R 值快速增长,表明此时文本类别较不平衡,使用这样的数据进行后续操作将不利于合成类别均衡的新文本.

因此,为了减少后续文本合成操作的计算代价,提高合成效率,数据筛选过程中需要剔除类别严重失衡的数据. 结合上述分析,数据筛选具体流程如下.

(1) 计算每个语句中实体单词个数,剔除不含实体单词的语句;

(2) 剔除单词个数超过 100 的语句.

2.2.2 *适应度评估* 适应度函数,在遗传算法中用以评估给定解决方案与所需问题最佳解决方案的接近程度[17]. CBM-GA 方法使用适应度函数作为评估指标以评估样本的平衡性,并将类别标记平衡条件引入适应度函数中,使得算法寻优过程中尝试构造类别分布平衡的新样本. 定义适应度函数 f 如式(2)所示.

$$f=sigmoid(Rs-R)+\lambda\times sigmoid(R-Rr) \tag{2}$$

式(2)中,R 为式(1)定义的语句非实体单词和实体单词的比值;R_s是合成文本 R 值上限值;R_r是合成文本 R 值下限值;λ 是 R_r权重系数. 式(2)借助 sigmoid 函数控制合成样本的 R 值在$[R_r, R_s]$范围内.

由图 2 分析可知文本单词个数不应过长,因此向适应度函数中添加单词个数小于 100 的限制,如式(3)所示.

$$L=\frac{relu(l-100)}{l} \tag{3}$$

式(3)中,L 表示语句单词个数限制函数;l 表示当前样本单词个数. relu 函数限制合成文本的长度. 适应度函数引入单词个数限制,如式(4)所示.

$$f=sigmoid(Rs-R)+\lambda\times sigmoid(R-Rr)+\mu\times\frac{relu(l-100)}{l} \tag{4}$$

式(4)中,μ 是 L 的权重系数,用以控制类别平衡条件和单词个数限制的重要性差异.

2.2.3 *文本合成* CBM-GA 方法文本合成过程主要是通过一系列选择、交叉和变异操作合成数据标记平衡的数据. 其中,选择操作从有序父代样本集合中抽取父代样本构建父代样本对集合;交叉操作将父代样本对进行组合以生成新样本;变异操作调整新样本的单词顺序,向样本中添加更多随机性.

CBM-GA 方法将文本视为染色体,将单词及其实体标记类型视为染色体上的基因. 如图 3 所示. 每个框内上行表示单词,下行表示对应实体类型标记[17],二者共同构成一个完整的基因. 基因构成染色体,又称之为样本. 将数据集中每个文本视为样本,数据集视为样本群体.

为了便于后续分析,定义样本染色体 $x=(x_1, x_2, \cdots, x_n)$,其中,$x_i$ 是包含单词及其对应实体类型标记的基因,n 是染色体长度.

1) 选择操作

在选择步骤中,遍历有序父代样本集并按照比例选择方法[20]随机选择两个父代样本构成父代样本对,从而构成父代样本对集合.

EU	rejects	German	call	to	boycott	British	lamb	.
B-ORG	O	B-MISC	O	O	O	B-MISC	O	O

图 3 染色体示意图

Fig. 3 Chromosome schematic diagram

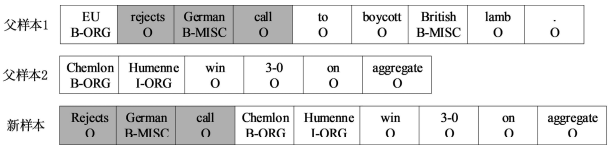

图 4　交叉过程示意图

Fig. 4　Crossover schematic diagram

2) 交叉操作

在交叉阶段，对父代样本对集合中每个父代样本对采用随机交叉策略进行单词序列相互交换，以产生新样本，如图 4 所示.

为避免实体短语被切分而导致命名实体模糊、歧义等问题，CBM-GA 方法修改基因组合规则，限制单词交换位置仅为开始，结尾以及 O 标记[18]的非实体单词位置以确保实体单词结构完整交换. 因此首先根据父样本 1 的实体标记类型，构建交换位置集合 l. 并随机从 l 中抽取交叉发生的开始、结束位置以得到单词序列如图中黑色框单词序列，将其与父样本 2 合并生成新样本 x'.

$$x' = x_1[l_{\text{start}} : l_{\text{end}}] + x_2 \tag{6}$$

式(6)中，x_1 表示父样本 1；x_2 表示父样本 2；l_{start}、l_{end}分别表示从交换位置集合 l 随机抽取的开始、结束位置；x'表示新样本.

3) 变异操作

本文采用随机交换单词位置的方式实现变异. 具体操作为：设定一个变异概率 α，对交叉得到的新样本 x'，随机产生一个[0,1]区间上的随机数 r，如果 $r<\alpha$，则随机交换新样本 x'两个单词的位置. 最终生成样本如式(7)所示.

$$x'' = \begin{cases} x' & , if \quad r \geqslant \alpha \\ \begin{pmatrix} x_1, x_2, \cdots, x_{i-1}, x_k, x_{i+1}, \cdots \\ , x_{k-1}, x_i, x_{k+1}, \cdots, x_n \end{pmatrix} & , if \quad r < \alpha \end{cases} \tag{7}$$

式(7)表明当 $r<\alpha$ 时，随机选定两个单词位置 i、k，交换 x'中两个位置的单词从而实现变异操作，反之则不进行变异.

综上分析，CBM-GA 方法完整表述如算法 1.

算法 1　CBM-GA 方法

输入　原始训练样本集合 S；参数 R_s、R_r、λ、μ、循环次数 N、变异概率 α、合成样本集合大小 L_a

输出　扩展的新样本集合 A

Begin

初始化迭代次数 n：$n=0$

(1) 遍历原始训练样本 S，过滤不含实体单词及文本单词个数超过 100 的语句，获得父代样本集 D；

(2) 遍历父代样本集 D，根据式(4)计算每个语句的适应值并排序，得到有序父代样本集 D_1；

(3) 使用比例选择法随机从 D_1 中抽取两个父代样本组成的父代样本对集合；

(4) 遍历父代样本对集合：

(a) 针对父代样本对(p_1，p_2)，分别从 p_1、p_2样本的开始、结尾及非实体单词的位置集合中随机选择位置变量以截取单词序列；

(b) 将单词序列合并成新样本 T；

(c) 以概率 α 交换 T 中任意两个位置的单词，并将 T 加入新样本集合 G；

(5) 遍历 G，根据式(4)计算每个新语句的适应值. 抽取适应值最高的 L_a 个样本的集合替换 D_1、G，迭代次数 n 加 1；

(6) 如果 $n<N$，重复步骤(3)～(5)操作；否则合并原始训练样本集合 S 和合成样本集合 G，得到扩展的新样本集合 A.

End

3　实验与结果分析

3.1　实验数据及环境

为了验证 CBM-GA 方法的有效性，排除单一数据集影响，实验采用命名实体识别研究中常用数据集 CoNLL2003[18]和 JNLPBA[19]进行实验. 数据集基本统计信息如表 1 所示.

表 1　数据集统计信息表

Tab. 1　Dataset statistics table

数据集	划分	语句总数	实体总数
CoNLL2003	训练集	14041	34043
	验证集	3250	8603
	测试集	3453	8112
JNLPBA	训练集	18546	109588
	测试集	3856	19392

如表 1 所示，CoNLL 2003 已分配好训练集、验证集和测试集；JNLPBA 仅划分了训练集和测试集. 本文从 JNLPBA 中随机抽取 1/3 数据作为验证集.

本文实验平台为 ubuntu16. 04 系统服务器，GPU 为 GeForce GTX 1070，显存 8 G. 实验模型使用 tensorflow 框架构建.

3.2 实验评价指标

Akkasi 等人提出平衡欠采样方法以缓解数据集的类别失衡问题,从而改善命名实体识别的效果,其借助命名实体识别评价指标作为最终的方法有效性衡量指标[6]. 本文沿用该评价指标来论证 CBM-GA 方法的有效性. 命名实体识别一般采用精准率(prec)、召回率(recall)和 F_1 值作为模型性能评估指标[1-5]. 本文使用的模型评价指标定义如下.

$$\text{Prec}=\frac{\text{预测为实体且实际为实体单词个数}}{\text{预测为实体单词总数}} \tag{8}$$

$$\text{Recall}=\frac{\text{预测为实体且实际为实体单词个数}}{\text{样本实际单词总数}} \tag{9}$$

$$F_1=2\frac{\text{Prec}\times\text{Recall}}{\text{Prec}+\text{Recall}} \tag{10}$$

3.3 实验参数设置与寻优

根据 2.2 节定义,CBM-GA 方法实现过程涉及参数:合成样本 R 值上限 R_s,合成样本 R 值下限 R_r,R_r 值权重 λ,L 的权重系数 μ,合成样本集合大小 L_a,变异概率 α 以及循环次数 N.

根据 Whitley 等人的经验[17],本文实验将 α 设为 0.01. 为给其它参数设置合理的取值,实验基于 Bi-LSTM-CRF 模型[2]使用贝叶斯优化[16]寻优方法来设置参数. 具体操作如下.

1) 分别以 1 为步长,设置 R_s,R_r 的取值范围为 [0,30];以 500 为步长,设置 L_a 的取值范围为 [500,2000];以 0.1 为步长,设置 λ 取值范围为 [0,1];以 1 为步长,设置循环次数 N 取值范围为 [1,10],构建参数集合;

2) 基于 Bi-LSTM-CRF 模型使用贝叶斯优化寻优方法以 F_1 值为指标,对参数集合进行寻优. 选取验证集实验结果中 F_1 值最大的参数作为后续实验参数.

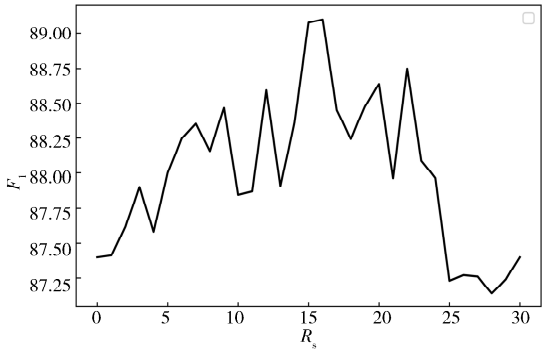

图 5 参数 R_s 选择实验结果
Fig. 5 Results of R_s selection

以 CoNLL 2003 数据集的 R_s 参数选取为例,以 1 为步长,设置 R_s 取值区间为 [0,30],绘制不同 R_s 对应 F_1 值分布图,如图 5 所示.

根据图 5 可知,当 R_s 为 15 时,合成的数据集进行命名实体识别 F_1 值最高,因此选定 $R_s=15$. 由此思路选择其他参数,最终参数取值如表 2 所示.

表 2 数据集参数表
Tab. 2 Parameter table

数据集	Rs	Rr	λ	μ	La	N
JNLPBA	8	1	0.7	1	1000	2
CoNLL2003	15	1	0.7	1	1000	2

3.4 实验测试结果

实验使用 Bi-LSTM-CRF 模型作为基准模型. 为了验证 CBM-GA 方法的有效性和优异性,分别设计了 2 组对比实验. 且为了避免偶然因素影响,实验结果均为 5 次重复实验结果. 具体如下.

为了验证 CBM-GA 方法的有效性,设计基准模型和 CBM-GA 方法对比实验.

1) 基准实验:使用基准模型分别对 CoNLL2003、JNLPBA 建模;

2) CBM-GA 方法实验:使用基准模型分别对 CBM-GA 方法作用后的两数据集建模;

针对 CoNLL 2003 数据集的实验结果如表 3 所示,CBM-GA 方法相比基准模型在保持精确率几乎不变的情况下,召回率提升 3.26%,F_1 值提高 1.70%;针对 JNLPBA 数据集的实验结果如表 4 所示,CBM-GA 方法虽然造成精确率的小幅下降,但其召回率提高了 2.44%,最终 F_1 值增加了 1.03%.

为了验证 CBM-GA 方法表现优于已有平衡欠采样、随机过采样方法,设计以下对比实验.

1) 随机过采样方法实验:使用随机过采样方法扩充原始 CoNLL 2003、JNLPBA,获取与 CBM-GA 方法相同规模的扩充数据集,并使用基准模型其建模;

2) 平衡欠采样方法实验. 使用平衡欠采样方法处理原始两数据集获得新样本集合,并随机采样新样本集合扩充原始数据集以获取与 CBM-GA 方法相同规模的扩充数据集,使用基准模型其建模;

优异性验证实验结果如表 5 和表 6 所示,针对 CoNLL 2003 数据集,CBM-GA 方法召回率比平衡欠采样高 2.98%,比随机过采样方法高 3.29%,F_1 值均超出 1.76%以上;针对 JNLPBA,CBM-GA 方法召回率比平衡欠采样高 1.78%,比

随机过采样方法高 2.25%，F_1 值均超出 0.97% 以上.

综上实验分析验证了 CBM-GA 方法可以有效提高模型召回率，改善命名实体识别效果，相比已有方法表现更优异.

表 3　CoNLL 2003 数据集上有效性验证结果

Tab. 3 Validation results of effectiveness on CoNLL 2003

方法	数据集 CoNLL 2003		
	精确率/%	召回率/%	F_1
基准	88.38	86.44	87.40
CBM-GA	88.51	89.70	89.10

表 4　JNLPBA 数据集上有效性验证结果

Tab. 4　Validation results of effectiveness on JNLPBA

方法	数据集 JNLPBA		
	精确率/%	召回率/%	F_1
基准	69.49	74.65	71.97
CBM-GA	69.32	77.09	73.00

表 5　CoNLL 2003 数据集上优异性验证结果

Tab. 5　Validation results of superiority on CoNLL 2003

方法	CoNLL 2003		
	精确率/%	召回率/%	F_1
随机过采样	88.30	86.41	87.34
平衡欠采样	87.37	86.72	87.04
CBM-GA	88.51	89.70	89.10

表 6　JNLPBA 数据集上优异性验证结果

Tab. 6　Validation results of superiority on JLPBA

方法	JNLPBA		
	精确率/%	召回率/%	F_1
随机过采样	69.42	74.84	72.03
平衡欠采样	68.45	75.31	71.71
CBM-GA	69.32	77.09	73.00

3.5　结果分析

进一步分析实验结果，以 CoNLL 2003 数据集为例，绘制基准、CBM-GA 实验接收者操作特征曲线(receiver operating characteristic curve，ROC)[21] 如图 6 所示.

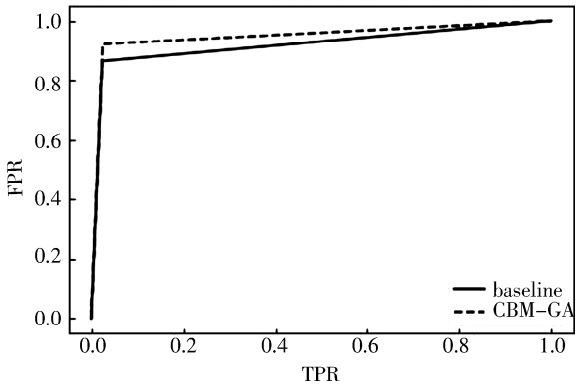

图 6　ROC 曲线

Fig. 6　ROC of baseline and CBM-GA

分别计算两条 ROC 曲线对应 AUC[22] 值如表 7 所示.

表 7　AUC 值

Tab. 7　AUC value

方法	AUC
基准	0.920
CBM-GA	0.936

图 6 和表 7 更一步证明 CBM-GA 模型通过缓解实体类和非实体类单词个数的不平衡问题，有效地改善了命名实体识别的效果.

从时间代价分析，CBM-GA 算法增广 CoNLL 2003 数据集需 2.42 min，增广 JNLPBA 数据集训练需 2.4 min，相比 Bi-LSTM-CRF 模型训练每 epoch 需 32 s，50 个 epoch 需要 26.7 min，CBM-GA 算法的运行成本是可以接受的.

4　结　论

数据类别标记失衡是普遍存在于开源数据集的问题，但目前关于命名实体识别任务上的数据标记失衡研究较少. 本文针对这一现状创新性地改造遗传算法，提出了保持文本实体短语结构的 CBM-GA 方法. 实验结果表明，CBM-GA 方法在文本数据预处理阶段有效缓解数据集类别标记失衡问题，改善模型召回率并进一步提高命名实体识别性能. 该方法可以应用在其它序列标注任务上如分词、机器翻译等中.

参考文献：

[1]　Lample G, Ballesteros M, Subramanian S, *et al*. Neural architectures for named entity recognition [J]. Assoc Comput Linguist, 2016, 1: 260.

[2] Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. Assoc Comput Linguist, 2016, 1: 1064.

[3] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs [J]. Assoc Comput Linguist, 2016, 4: 357.

[4] Gregoric A Z, Bachrach Y, Coope S. Namedentity recognition with parallel recurrent neural networks [J]. Assoc Comput Linguist, 2018, 2: 69.

[5] 张若彬, 刘嘉勇, 何祥. 基于 BLSTM-CRF 模型的安全漏洞领域命名实体识别[J]. 四川大学学报:自然科学版, 2019, 56: 469.

[6] Akkasi A, Varoğlu E, Dimililer N. Balancedundersampling: a novel sentence-based undersampling method to improve recognition of named entities in chemical and biomedical text [J]. Appl Intel, 2017, 48:1.

[7] Guo X, Yin Y, Dong C, *et al*. On the class imbalance problem[C]// Proceedings of the Fourth International Conference on Natural Computation. [S. l.]: IEEE, 2008.

[8] Wang S, Tang K, Yao X. Diversity exploration and negative correlation learning on imbalanced data sets [C]// Proceedings of the International Joint Conference on Neural Networks. [S. l.]: IEEE, 2009.

[9] Sun Y, Kamel M S, Wong A K C, *et al*. Cost-sensitive boosting for classification of imbalanced data [J]. Pattern Recogn, 2007, 40: 3358.

[10] Dai Q, Zhang T, Liu N. A new reverse reduce-error ensemble pruning algorithm [J]. Appl Soft Comput, 2015, 28: 237.

[11] He H, Ma Y. Imbalanced learning (foundations, algorithms, and applications) || ensemble methods for class imbalance learning [J]. IEEE, 2013, 61: 82.

[12] Tomanek K, Hahn U. Reducing class imbalance during active learning for named entity annotation [C]//Proceedings of the fifth international conference on knowledge capture. [S. l.]: ACM, 2009.

[13] Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks[J]. Expert Syst Appl, 2018, 91: 464.

[14] Gliozzo A M, Giuliano C, Rinaldi R. Instance pruning by filtering uninformative words: an information extraction case study [C]// Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics. Heidelberg: Springer, 2005.

[15] Maragoudakis M, Kermanidis K, Garbis A, *et al*. Dealing with imbalanced data using bayesian techniques[C]// Proceedings of the 5th International Conference on Language Resources and Evaluation. [S. l.]:[s. n.], 2006.

[16] Snoek J, Larochelle H, Adams R P. Practical bayesian optimization of machine learning algorithms [C]// Proceedings of the advances in neural information processing systems. [S. l.]:[s. n.], 2012.

[17] Whitley D. A genetic algorithm tutorial[J]. Stat Comput, 1994, 4: 65.

[18] Tjong Kim Sang E F, De Meulder F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition [J]. Assoc Comput Linguist, 2003, 4:142.

[19] Kim J D, Ohta T, Tsuruoka Y, *et al*. Introduction to the bio-entity recognition task at JNLPBA [C]// Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. [S. l.]:[s. n.], 2004.

[20] Fogel D B. Evolutionary algorithms in theory and practice [J]. Complexity, 1997, 2: 26.

[21] Zou K H, O'Malley A J, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models [J]. Circulation, 2007, 115: 654.

[22] Fawcett T. An introduction to ROC analysis [J]. Pattern Recogn Lett, 2006, 27: 861.

引用本文格式:

中 文: 许丽丹, 刘嘉勇, 何祥. 一种解决命名实体识别数据集类别标记失衡的方法[J]. 四川大学学报:自然科学版, 2020, 57: 82.

英 文: Xu L D, Liu J Y , He X. Method for solving class imbalance of named entity recognition dataset [J]. J Sichuan Univ: Nat Sci Ed, 2020, 57: 82.