融合类别先验 Mixup 数据增强的罪名预测方法

线岩团[1,2] 陈文仲[1,2] 余正涛[1,2] 张亚飞[1,2] 王红斌[1,2]

摘 要 罪名预测是人工智能技术应用于司法领域的代表性任务. 该任务根据案情描述和事实预测被告人被判的罪名. 由于各类罪名样本数量高度不平衡, 分类模型训练时分类器易偏向高频罪名类别, 从而导致低频罪名预测性能不佳. 针对罪名预测类别不平衡问题, 提出融合类别先验 Mixup 数据增强策略的罪名预测模型, 改进低频罪名预测效果. 该模型利用双向长短期记忆网络与结构化自注意力机制学习文本向量表示, 在此基础上, 通过 Mixup 数据增强策略在向量表示空间中合成伪样本, 并利用类别先验使合成样本的标签偏向低频罪名类别, 以此来扩增低频罪名训练样本. 实验结果表明, 与现有方法相比, 该方法在准确率、宏精确率、宏召回率和宏 F1 值上都获得了大幅提升, 低频罪名预测的宏 F1 值提升达到 13.5%.

关键词 类别先验 Mixup, 罪名预测, 类别不平衡分类, 低频罪名

引用格式 线岩团, 陈文仲, 余正涛, 张亚飞, 王红斌. 融合类别先验 Mixup 数据增强的罪名预测方法. 自动化学报, 2022, **48**(8): 2097−2107

DOI 10.16383/j.aas.c200908

Category Prior Guided Mixup Data Argumentation for Charge Prediction

XIAN Yan-Tuan[1,2] CHEN Wen-Zhong[1,2] YU Zheng-Tao[1,2] ZHANG Ya-Fei[1,2] WANG Hong-Bin[1,2]

Abstract Charge prediction is a typical task of artificial intelligence technology applied in the field of justice. The task is to predict the charges of the accused based on the description of the case and the fact section. Due to the highly class-imbalanced of charges, classifiers usually result in poor prediction of low-frequency charges. To address the class-imbalanced problem, we propose a Mixup data augmentation strategy that combines category prior knowledge to improve the prediction performance of low-frequency charges. In this paper, we first learn the representation of text vector using the bi-directional long short-term memory model and structured attention mechanism. Then, we apply the proposed Mixup data augmentation strategy to generate synthetic samples in the text representation space. To emphasize data augmentation of the low-frequency charge samples, the category prior is employed to bias the synthetic labels to the low-frequency category. The experimental results on real-work data sets demonstrate that our method achieves significant and consistent improvements compared to other state-of-the-art baselines on the accuracy, macro precision, macro recall, and macro F1 value. Specifically, our model outperforms other baselines by more than 13.5% macro F1 value in the low-frequency charges.

Key words Category prior guided mixup, charge prediction, class imbalanced classification, low-frequency charge

Citation Xian Yan-Tuan, Chen Wen-Zhong, Yu Zheng-Tao, Zhang Ya-Fei, Wang Hong-Bin. Category prior guided mixup data argumentation for charge prediction. *Acta Automatica Sinica*, 2022, **48**(8): 2097−2107

罪名预测是法律判决预测任务中具有代表性的子任务, 也是法律辅助系统的重要组成部分[1].

罪名预测通常被看作针对案件事实的文本分类问题. 早期研究工作通常利用统计机器学习方法实现罪名预测[2−4]. 随着深度学习在自然语言处理领域的广泛应用, 基于深度学习方法的罪名预测模型大量涌现.

2018 中国 “法研杯”司法人工智能挑战赛发布中文司法判决预测数据集, 共包含 260 余万条数据, 数据源于 “中国裁判文书网”公开的刑事法律文书[5]. 针对中文的司法判决预测任务, 目前有较多的研究工作均在此数据集上展开.

Zhong 等[6] 将多种判决预测任务之间的依赖视为有向无环图, 提出了拓扑多任务学习框架, 并将多种判决任务间的依赖关系融入分类模型, 改进了罪名预测效果. Yang 等[7] 借助多任务间的拓扑结

收稿日期 2020-10-31 录用日期 2021-03-02
Manuscript received October 31, 2020; accepted March 2, 2021
云南省基础研究计划 (202001AT070046), 国家重点研发计划 (2018YFC0830104, 2018YFC0830105, 2018YFC0830100) 和国家自然科学基金 (61966020) 资助
Supported by Science and Technology Plan Projects of Yunnan Province (202001AT070046), National Key Research and Development Program Foundation of China (2018YFC0830104, 2018YFC0830105, 2018YFC0830100), and National Natural Science Foundation of China (61966020)
本文责任编委 刘洋
Recommended by Associate Editor LIU Yang
1. 昆明理工大学信息工程与自动化学院 昆明 650500 2. 昆明理工大学云南省人工智能重点实验室 昆明 650500
1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500

构, 通过多角度前向预测和反向验证提高了多任务审判预测性能. 王文广等[8] 提出了融合层次注意力网络和卷积神经网络的多任务罪名预测模型. 已有研究表明, 将罪名预测与其他相关判决预测任务联合建模, 为模型提供更多的监督信息, 可以改进罪名预测效果.

Jiang 等[9] 采用深度强化学习方法抽取文本中的论据, 并利用论据增强分类来提高罪名预测的准确率. 刘宗林等[10] 在罪名预测和法条推荐联合模型中融入罪名关键词提升了罪名预测性能. Xu 等[11] 采用图神经网络学习易混淆法条之间的差异, 并设计注意力机制充分利用这些差异从事实描述中抽取出明显特征去区分易混淆罪名. 已有的罪名预测工作大多从多任务学习和外部知识融入的角度开展罪名预测研究, 未考虑罪名预测的数据分布问题.

由于各类案件发生概率的差异较大, 罪名预测数据存在着严重的类别不平衡问题. 以 Hu 等[12] 构建的罪名预测数据集为例, Criminal-L 训练集共包含 149 类罪名, 将各罪名按其样本占比降序排列, 其中前 10 类高频罪名对应的样本占比约为 78%, 而最后 100 类罪名的样本仅占约 3%, 这是典型的 “长尾数据”. 各类罪名在数量上的高度不平衡易导致模型在训练时偏向于高频罪名而忽略低频罪名, 造成在罪名预测时低频罪名易被错误分类的问题, 从而严重影响模型性能.

针对罪名预测的类别不平衡问题, Hu 等[12] 在人工标注法律属性的基础上, 构建联合罪名预测和法律属性预测的多任务分类模型, 提高了低频罪名的预测性能. He 等[13] 在胶囊网络基础上, 提出融合文本序列信息和空间信息的罪名预测模型, 并引入 Focal Loss 损失函数, 有效提高了低频罪名的预测效果.

和已有的多任务方法[12] 与改进损失函数的方法[13] 不同, 本文从数据增强角度研究罪名预测的类别不平衡问题. 本文借鉴图像分类中的混合样本数据增强方法[14−15], 在文本的表示空间中扩增训练样本, 并提出融合罪名先验概率的标签合成策略, 使合成样本偏向低频罪名类别, 从而达到扩增低频罪名训练样本的目的. 在表示空间中合成偏向低频罪名的训练样本, 既扩增了训练样本的数量, 又丰富了特征的多样性, 有助于平滑模型的分类面, 提高模型的泛化能力.

本文采用 Lin 等[16] 提出的结构化自注意力句子嵌入方法构建罪名预测模型, 并在模型训练过程中融入类别先验混合样本数据增强策略, 提升模型性能. 实验结果表明, 本文提出的融入类别先验 Mixup 数据增强的罪名预测方法可以在不增加人工标注和辅助任务的前提下, 有效改进罪名预测模型性能, 显著提高低频罪名的预测效果. 本文提出方法的源代码可从网址 https://github.com/xianyt/ proir_mixup_charge 下载.

本文方法的主要贡献如下:

1) 本文将 Mixup 数据增强方法引入罪名预测任务中, 利用文本表示空间中的插值操作合成训练样本. 合成样本增加了训练样本的多样性, 有效提高了罪名预测模型的泛化能力.

2) 本文针对罪名不平衡问题, 提出了类别先验引导的 Mixup 数据增强策略. 该策略在文本表示空间中生成倾向于低频罪名的合成样本, 扩增了低频罪名样本, 有效缓解了罪名不平衡问题, 提高了低频罪名的预测效果.

3) 与基线模型相比, 本文方法在 Hu 等[12] 构建的 3 个不同规模的罪名预测数据集上都取得了最好的预测效果. 模型在宏准确率、宏召回率和宏 F1 值上都有显著提升, 低频罪名宏 F1 值提升达到 13.5%.

1 相关工作

已有的罪名预测研究工作主要从多任务联合学习[12] 和外部知识融入[10] 的角度来提升模型性能, 并利用辅助任务和改进的损失函数来缓解罪名预测任务面临的类别不平衡问题. 与已有工作不同的是, 本文从数据增强角度来改进罪名预测方法, 提升罪名预测性能.和已有罪名预测方法相比, 本文方法没有引入辅助任务,也不需要额外的数据标注工作; 另外, 本文提出的数据增强策略不依赖于特定的文本编码器, 可以应用于已有的罪名预测模型.

Zhang 等[14] 提出的 Mixup 方法是一种应用于图像分类的数据增强策略.该方法是从训练集中随机抽取图像样本, 并通过线性混合来合成新的图像样本, 有效改进了小样本图像分类的性能[14]. 由于文本是一种离散表示, 所以 Mixup 方法无法直接应用于文本分类任务. Verma 等[15] 提出的 Manifold Mixup 方法在图像的嵌入空间中利用随机混合图像的向量表示来生成编码空间中的伪样本; 相比 Mixup 方法, Manifold Mixup 能够提供更高层的监督信息, 使模型具有更好的泛化能力. 受 Manifold Mixup 方法启发, 本文提出了融合类别先验 Mixup 方法, 与 Manifold Mixup 方法中对样本向量表示和标签采用相同混合因子的做法不同, 本文方法针对文本表示和分类标签采用不同的混合因子, 利用罪名的先验概率来生成偏向低频类别的伪样本, 以此来缓解罪名不平衡问题.

目前, Mixup 方法在自然语言处理领域仅有少量的研究工作. Guo 等[17] 将 Mixup 数据增强方法应用于句子分类任务, 提出了词级和句子级的 Mixup 策略, 提升了句子分类的性能, 将 Mixup 数据增强方法应用于句子分类任务, 提出了词级和句子级的 Mixup 策略, 提升了句子分类的性能. Chen等[18] 将 Mixup 方法应用于半监督文本分类任务, 改进了分类效果. 目前还未见针对不平衡文本分类问题的 Mixup 方法. 所以, 本文面向罪名预测任务, 研究不平衡文本分类的 Mixup 数据增强策略具有明显的创新性.

2　罪名预测模型

本文提出的罪名预测方法在深度学习文本分类模型基础上, 引入 Mixup 数据增强策略, 并利用罪名先验概率生成偏向低频罪名的伪样本, 以此缓解罪名预测中的类别不平衡问题.

本文提出的罪名预测模型包括编码层、类别先验引导 Mixup 层和分类层 3 层. 图 1 展示了本文提出的罪名预测模型的总体结构. 最下方的编码层用于学习罪名描述文本的向量表示, 该层包括 3 个子层, 分别是词嵌入层、双向长短时记忆网络编码层[19] (Bi-directional long short-term memory, Bi-LSTM) 和结构化注意力层[9]. 在训练模型时, 本文方法在编码层与分类层间加入类别先验引导 Mixup 层, 该层通过随机混合的文本向量表示和对应的分类标签生成伪样本和伪标签. 伪样本向量表示和文本向量表示被送入分类层. 分类层通过全连接层和 Softmax 激活函数计算罪名预测值, 并针对伪样本和普通样本计算分类损失.

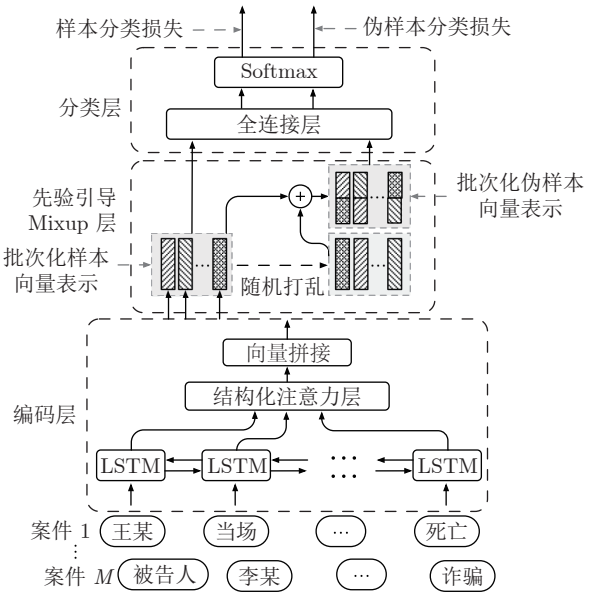


图 1　罪名预测模型的总体结构图

Fig. 1　Overview of proposed charge prediction model

本文选择 Bi-LSTM 作为文本编码器主要有 3 个方面的考虑. 首先, Bi-LSTM 是一种被广泛应用的序列编码器, 可以有效对长文本进行建模. Bi-LSTM 适合用于对篇章级的案件描述进行编码, 其有效性已在多个罪名预测模型中得到验证[6−7,13]; 其次, 在实验过程中作者发现, Bi-LSTM 与结构化注意力机制结合可能很好地获取多个侧面的文本分类特征. 最后, 相比于双向编码器表示模型 (Bidirectional encoder representation from transformers, BERT)[20] 等预训练语言模型, Bi-LSTM 结构简单易于训练, 可应用于大规模文本分类问题. 而且, 在类别严重不平衡的罪名预测任务上, Bi-LSTM 模型训练过程中过拟合现象不明显. 第 4.4 节对比了不同文本编码器对罪名预测性能的影响.

案情描述和事实文本中的词序列 $\boldsymbol{x} = [w_1, w_2, \cdots, w_n]$ 经过词嵌入编码后得到词序列的低维向量表示 $\boldsymbol{E} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n]$, 其中, n 表示文本长度, w_i 表示文本中的第 i 个词, $\boldsymbol{e}_i \in \mathbf{R}^d$ 表示第 i 个词的词向量, d 表示词向量的维度.

Bi-LSTM 层以词序列的向量表示为输入计算词语在上下文中的向量表示:

$$\begin{cases} \overrightarrow{\boldsymbol{h}}_i = \overrightarrow{\text{LSTM}}(\overrightarrow{\boldsymbol{h}}_{i-1}, \boldsymbol{e}_i) \\ \overleftarrow{\boldsymbol{h}}_i = \overleftarrow{\text{LSTM}}(\overleftarrow{\boldsymbol{h}}_{i+1}, \boldsymbol{e}_i) \end{cases} \tag{1}$$

式中, $\overrightarrow{\text{LSTM}}$ 和 $\overleftarrow{\text{LSTM}}$ 分别表示正向和逆向长短期记忆网络 (Long short-term memory, LSTM), $\overrightarrow{\boldsymbol{h}}_i$ 和 $\overleftarrow{\boldsymbol{h}}_i$ 分别表示对应的 LSTM网络的输出.

为了获得具有上下文语义的词语表示,本文将正向和逆向的 LSTM 输出 $\overrightarrow{\boldsymbol{h}}_i$、$\overleftarrow{\boldsymbol{h}}_i$ 和 $\boldsymbol{e}_i$ 拼接作为第 i 个词在序列中的隐状态表示:

$$\boldsymbol{h}_i = \text{concat}(\overrightarrow{\boldsymbol{h}}_i, \overleftarrow{\boldsymbol{h}}_i, \boldsymbol{e}_i) \tag{2}$$

通过拼接 $\boldsymbol{h}_i$ 序列可得到词序列的隐状态表示 $\boldsymbol{H} \in \mathbf{R}^{n\times(2u+d)}$, 其中, u 表示隐状态的维度.

$$\boldsymbol{H} = [\boldsymbol{h}_1; \boldsymbol{h}_2; \cdots; \boldsymbol{h}_n] \tag{3}$$

本文采用结构化自注意力层来计算文本多个侧面的向量表示. 该层的注意力权重矩阵 $\boldsymbol{A} \in \mathbf{R}^{r\times n}$ 由 2 层感知机计算得到,

$$\boldsymbol{A} = \text{softmax}(\boldsymbol{W}_{s2}\text{tanh}(\boldsymbol{W}_{s1}\boldsymbol{H}^{\text{T}})) \tag{4}$$

式中, $\boldsymbol{W}_{s1} \in \mathbf{R}^{d_a\times(2u+d)}$ 和 $\boldsymbol{W}_{s2} \in \mathbf{R}^{r\times d_a}$ 是注意力层的参数, d_a 和 r 为模型的超参数, d_a 表示注意力层隐状态的维度, r 是注意力机制的个数.

文本表示矩阵 $\boldsymbol{Z} \in \mathbf{R}^{r\times(2u+d)}$ 由词序列的隐状

态表示 $\boldsymbol{H}$ 和注意力权重矩阵 $\boldsymbol{A}$ 的乘积得到,

$$\boldsymbol{Z}=\boldsymbol{A}\boldsymbol{H} \tag{5}$$

文本的向量表示 $\boldsymbol{z}$ 由矩阵 $\boldsymbol{Z}$ 中的 r 个向量拼接得到, 其维度为 $r\times(2u+d)$.

在训练过程中, 类别先验引导 Mixup 层通过随机混合批次内的文本向量表示 $\{\boldsymbol{z}_i\}_i^M$ 得到扩增的文本向量表示 $\{\tilde{\boldsymbol{z}}_j\}_j^M$, 其中 M 是一个批次内的样本数据量, 具体方法将在第 3 节中详细阐述.

最后, 分类层通过线性层和 Softmax 激活函数预测各罪名的概率,

$$\hat{\boldsymbol{y}}=\text{softmax}(\boldsymbol{W}\boldsymbol{z}+\boldsymbol{b}) \tag{6}$$

式中, $\boldsymbol{W}\in\mathbf{R}^{K\times r(2u+d)}$ 和 $\boldsymbol{b}\in\mathbf{R}^K$ 分别是线性层的权重矩阵和偏置, K 表示罪名类别数.

3 类别先验 Mixup 数据增强方法

3.1 Mixup 数据增强策略

Mixup 数据增强方法的主要思想是通过混合随机抽取的 2 个图像和对应标签来生成伪样本来扩增训练数据[14]. 在此基础上, Verma 等[15] 提出在嵌入空间中生成伪样本 $(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})$ 的 Manifold Mixup 方法,

$$\begin{aligned}\tilde{\boldsymbol{x}}&=\lambda g_k(\boldsymbol{x}_i)+(1-\lambda)g_k(\boldsymbol{x}_j)\\ \tilde{\boldsymbol{y}}&=\lambda\boldsymbol{y}_i+(1-\lambda)\boldsymbol{y}_j\end{aligned} \tag{7}$$

式中, $g_k(\cdot)$ 表示神经网络编码器中从输入到第 k 层的前向过程, $\lambda\in[0,1]$ 为混合因子, 由 Beta 分布采样得到. 该方法在图像的嵌入空间中合成伪样本, 利用更高层次的表示为模型提供更多的监督信号, 从而有效提高了模型的泛化能力.

算法 1. 类别先验 Mixup 训练算法

输入. 数据集 $\mathcal{D}=\{(x_i,y_i)\}_{i=1}^N$, 文本编码器 f, 罪名先验 $\{p_k\}_{k=1}^K$, 批次大小 M.

输出. 模型参数 θ.

1) 初始化模型参数 θ;

2) while θ 不收敛时, do;

3) $\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^M\leftarrow\text{SampleData}(\mathcal{D},M)$ // 生成小批次样本;

4) $\{\boldsymbol{z}_i\}_{i=1}^M\leftarrow f\left(\{\boldsymbol{x}_i\}_{i=1}^M\right)$ // 将文本编码为向量;

5) $\{z_j,\boldsymbol{y}_j\}_{j=1}^M\leftarrow\text{Shuffle}\left(\{z_i,\boldsymbol{y}_i\}_{i=1}^M\right)$ // 随机打乱训练数据;

6) for $\{(\boldsymbol{z}_i,\boldsymbol{y}_i),(\boldsymbol{z}_j,\boldsymbol{y}_j)\}$ in PairData $\left(\{\boldsymbol{z}_i,\boldsymbol{y}_i\}_{i=1}^M,\{\boldsymbol{z}_j,\boldsymbol{y}_j\}_{j=1}^M\right)$ do // 随机配对样本;

7) $\lambda\leftarrow\text{Beta}(\alpha,\alpha)$ // 采样样本的混合因子 l;

8) $\lambda_y\leftarrow\text{LabelMixFactor}(p_i,p_j,\lambda)$ // 按式 (9) ~ (10) 计算标签的混合因子;

9) $\tilde{\boldsymbol{z}}_i\leftarrow\lambda\boldsymbol{z}_i+(1-\lambda)\boldsymbol{z}_j$ // 按式 (8) 合成样本;

10) $\tilde{\boldsymbol{y}}_i\leftarrow\lambda_y\boldsymbol{y}_i+(1-\lambda_y)\boldsymbol{y}_j$ // 按式 (8) 合成样本;

11) end for;

12) $\{\widehat{\boldsymbol{y}}_i\}_{i=1}^M\leftarrow\text{Predict}\left(\{\boldsymbol{z}_i\}_{i=1}^M\right)$ // 式 (6);

13) $\{\widehat{\boldsymbol{y}}_j\}_{j=1}^M\leftarrow\text{Predict}\left(\{\tilde{\boldsymbol{z}}_j\}_{j=1}^M\right)$ // 式 (6);

14) $\mathcal{L}(\boldsymbol{\theta})\leftarrow$ 按式 (11) 计算损失 // 计算损失函数;

15) $\boldsymbol{\theta}\leftarrow\boldsymbol{\theta}-\delta\nabla_\theta\mathcal{L}(\boldsymbol{\theta})$ // 更新模型参数;

16) end while.

3.2 类别先验 Mixup

本文借鉴 Manifold Mixup 方法的思想, 在文本的向量表示空间中合成伪样本. 在此基础上, 提出了融合类别先验的 Mixup 数据增强策略. 该策略在合成样本的表示和标签时采用不同的混合因子, 并通过各类别罪名的先验概率计算标签的混合因子, 以便使伪样本的标签偏向低频罪名. 本文提出的融合类别先验 Mixup 数据增强策略的公式可表示为:

$$\begin{cases}\tilde{\boldsymbol{z}}=\lambda f(\boldsymbol{x}_i)+(1-\lambda)f(\boldsymbol{x}_j)\\ \tilde{\boldsymbol{y}}=\lambda_y\boldsymbol{y}_i+(1-\lambda_y)\boldsymbol{y}_j\end{cases} \tag{8}$$

式中, $f(\cdot)$ 为将文本编码为向量的神经网络, $\lambda\in[0,1]$ 为样本的混合因子, 由 Beta (α,α) 分布采样得到, α 为超参数, $\lambda_{\text{y}}\in[0,1]$ 为标签的混合因子, $(\boldsymbol{x}_{\text{i}},\boldsymbol{y}_i)$ 和 $(\boldsymbol{x}_j,\boldsymbol{y}_j)$ 是从同一个训练批次中随机抽取的样本对.

为了能在训练过程中通过 Mixup 方法扩增低频罪名训练样本, 本文通过融合各类别罪名的先验概率来指导 Mixup 为低频罪名标签赋予更大的混合因子, 使得合成的伪样本更偏向于少样本类别. 为此, 首先根据类别先验概率计算类别混合因子 λ_p:

$$\lambda_p=1-\tanh\left(\frac{p(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)+p(\boldsymbol{x}_j)}\right) \tag{9}$$

式中, $p(\boldsymbol{x}_i)$ 和 $p(\boldsymbol{x}_j)$ 分别为样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 所对应类别的先验概率. 各类别罪名的先验概率根据训练集中的各类别罪名的占比计算得到. 如果 $\boldsymbol{x}_i$ 为低频罪名, 则意味着其先验概率低, 那么按式 (9) 为其分配较大的 λ_p, 以使得伪样本的标签偏向低频罪名; 反之, 为其分配较小的 λ_p.

在得到 λ_p 后, 本文将之与采样得到的样本混合因子 λ 进行平均得到标签混合因子 λ_y:

$$\lambda_y=\frac{2\lambda_p\lambda}{\lambda_p+\lambda} \tag{10}$$

通过引入类别先验, 使得合成样本既扩增了训练样本, 同时缓解了模型过于偏向高频罪名的问题.

本文将 Mixup 数据增强策略引入深度学习罪名预测模型中. 在训练过程中通过式 (8)、式 (9) 和

式 (10) 随机混合一个批次内的文本向量表示及其标签来获得伪样本, 并利用交叉熵分别计算样本和伪样本的损失, 模型损失 $\mathcal{L}(\boldsymbol{\theta})$ 公式如下:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^{M}\left(-\sum_{k=1}^{K}\boldsymbol{y}_{ik}\ln(\hat{\boldsymbol{y}}_{ik})\right)+ \frac{1}{M}\sum_{j=1}^{M}\left(-\sum_{k=1}^{K}\tilde{\boldsymbol{y}}_{jk}\ln(\hat{\boldsymbol{y}}_{jk})\right) \tag{11}$$

式中, 第 1 项为样本分类损失, 第 2 项为伪样本分类损失. M 为一个批次的样本数量, K 为罪名类别数, $\boldsymbol{y}_{ik}\in\{0,1\}$为样本 i在类别 k上的标签, $\tilde{\boldsymbol{y}}_{jk}\in[0,1]$ 为伪样本 j 在类别 k 上的伪标签, $\hat{\boldsymbol{y}}_{ik}$ 和 $\hat{\boldsymbol{y}}_{jk}$ 分别为样本 i 和 j 在类别 k 上的预测值.

融合类别先验 Mixup 数据增强的罪名预测模型的训练过程参见算法 1. 算法输入中的文本编码器对应第 2 节的编码层, 罪名先验概率由训练集中各罪名的样本数量预先估计得到.

4　实验及结果分析

为了验证所提出方法的有效性, 本文将之与现有罪名预测方法进行对比实验, 并分析了相关实验结果.

4.1　数据集与评价指标

本文采用 Hu 等[12] 构建的罪名预测数据集验证本文方法的有效性. 该数据集主要针对低频罪名和易混淆罪名预测任务构建, 不包含多被告、数罪并罚的情形. 该数据集包含小、中、大 3 个不同规模的子数据集, 分别命名为 Criminal-S、Criminal-M、Criminal-L, 数据集统计信息参见表 1.

图 2 展示了 Criminal 数据集中 3 个不同规模子数据集训练样本的高频、中频和低频罪名的分布情况. 图 2 中的高频、中频和低频罪名根据 Criminal-S 数据集的样本数量统计得到, 其中低频罪名为样本数少于 10 的罪名 (共 49 类), 高频罪名为样本数多于 100 的罪名 (共 49 类), 其余的作为中频罪名 (共 51 类). 由图 2 可以看出, 3 个不同规模数据集的罪名分布均呈现出典型的 “长尾分布”特征, 其中, 49 类高频罪名样本占比约为 97%, 中频罪名样本占比仅为 2.6% 左右, 而 3 个子数据集的低频罪名更加稀少, 均少于 1%. 从图 2 还可发现, 3 个不同规模的样本分布差异主要集中在低频罪名上.

为进一步比较 3 个子数据集在类别不平衡上的差异, 本文在图 3 展示了样本数量最少的 75 个罪名的样本分布情况. 本文统计了 Criminal-S 数据集中各罪名的样本数量, 并将罪名按样本从多到少排列作为图 3 的横坐标. 图 3 的纵坐标为各罪名样本的占比. 从图 3 可发现, 3 个数据集中频样本上的分布基本一致, 但是在低频样本分布上具有明显差异. 在 Criminal-S 数据集中低频罪名的分布比较稳定, 最低占比稳定在 0.013% 左右. Criminal-M 数据集的低频罪名占比在 0.005% ~ 0.013% 之间波动, 而 Criminal-L 数据集的低频罪名占比在 0.003% ~ 0.013% 之间波动. 相比之下, Criminal-L 数据集类别不平衡程度最严重, Criminal-M 数据集次之, 而 Criminal-S 数据集类别不平衡程度最轻.

在评价指标方面, 本文与文献 [12−13] 同样采用准确率 (Accuracy, Acc.)、宏精确率 (Macro-pre-

表 1　数据集统计信息

Table 1　The statistics of different datasets

数据集	Criminal-S	Criminal-M	Criminal-L
Train	61 589	153 521	306 900
Test	7 702	19 189	38 368
Valid	7 755	19 250	38 429

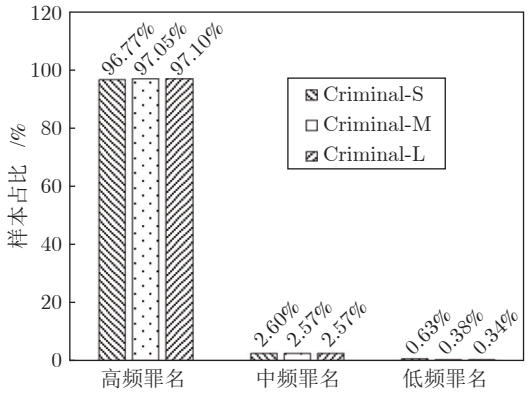


图 2　训练集罪名样本分布

Fig. 2　Charge distribution of the training set

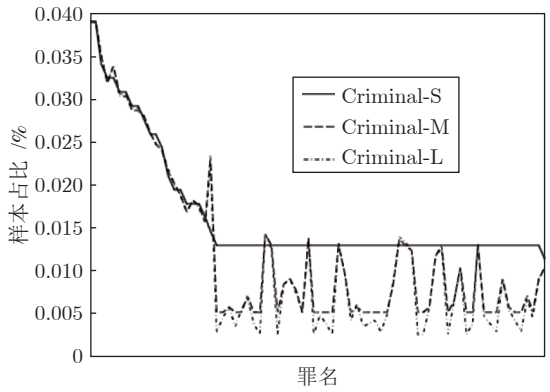


图 3　训练集罪名部分样本分布

Fig. 3　Charge distribution of the training set

cision, MP)、宏召回率 (Macro-recall, MR) 和宏 F1 值 (Macro F1) 作为模型性能的评价指标.

4.2 模型实现细节

本文采用 Pytorch 实现提出的模型和算法. 犯罪事实描述的最大词序列长度设为 500, 词频低于 5 的词被视为未知词. 词嵌入维度 d 设为 100, 并采用文献 [12] 的预训练词向量初始化嵌入层参数. Bi-LSTM 层的隐状态维度 u 设为 300. 嵌入层和 Bi-LSTM 层的 dropout 值分别设置为 0.3 和 0.1. 结构化自注意力机制的头数 r 设为 24, 注意力层隐状态维度 d_a 设为 128. 样本混合因子 λ 由参数 $\alpha = 150$ 的 Beta 分布采样得到, 标签混合因子 λ_y 由式 (9) 和式 (10) 计算得到.

模型采用 Adam 梯度下降算法[21] 训练, 初始学习率设为 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$. 最大训练轮次设为 50, 批次大小设为 256. 训练过程采用提前停止策略, 并根据验证集损失函数最小选择最优模型.

为减小案件描述中不同金额、重量、酒精含量、年龄等对模型词汇表的影响, 本文在数据预处理时对犯罪事实描述中的金额、重量、酒精含量、年龄等的数字部分进行了替换处理. 例如, 将 “2018 年”替换为 “×年”, “1000 元”替换为 “×元”.

4.3 对比实验

本文采用以下几种典型的文本分类模型和当前性能最优的罪名预测方法作为基线模型:

1) TFIDF+SVM: 该方法采用词频逆文档频率 (Term frequency-inverse document frequency, TF-IDF)[22] 抽取文本特征, 特征维度为 2000, 并采用支持向量机 (Support vector machine, SVM)[23] 作为分类器.

2) 卷积神经网络 (Convolutional neural networks, CNN): 该方法利用多个不同尺度的卷积网络构建文本分类器[24].

3) 长短期记忆网络: 采用双层 LSTM 作为案件事实编码器, 并利用最大池化获取分类特征[19].

4) 事实−法条注意力模型 (Fact-law attention model, Fact-Law Att): Luo 等[25] 提出的融合法条相关性与注意力机制的多任务罪名预测模型.

5) 小样本属性模型 (Few-shot attributes model, Few-Shot Attri): Hu 等[12] 提出的融合法律属性与罪名预测的联合模型, 该方法通过引入法律属性分类任务改进低频罪名预测性能.

6) 序列增强型的胶囊模型 (Sequence enhanced capsule model, SECaps): He 等[13] 提出的融合文本序列信息和空间信息的罪名预测模型, 并引入 Focal Loss 损失函数, 进而改进低频罪名的预测效果.

除 TFIDF+SVM 模型外, 其余对比模型词嵌入维度设为 100. LSTM 模型的隐状态维度设为 100. CNN 模型的滤波器宽度为 (2, 3, 4, 5), 每个滤波器的大小为 25. 基线模型实验结果引用自文献 [12–13].

本文实现了 2 个引入 Mixup 数据增强策略的模型: LSTM-Att-Manifold-Mixup 表示引入 Manifold Mixup 数据增强策略的罪名预测模型, LSTM-Att-Prior-Mixup 为融合类别先验 Mixup 的罪名预测模型. 本文方法与基线模型的对比实验结果见表 2.

由表 2 的实验结果可以看出, 本文方法与基线模型相比, 在 3 个数据集上均取得了最好的预测结果, 准确率、宏精确率、宏召回率和宏 F1 值均显著优于基线模型. 与现有最优模型 SECaps 相比, 本文模型 LSTM-Att-Prior-Mixup 在 Criminal-M 数据集上的性能提升最为明显, 准确率提升了 0.9%, MP 值提高了 9.5%, MR 值提高了 11.8%, F1 值提高了 10.5%. 对比实验结果表明, 类别先验 Mixup 数据增强方法能有效提高罪名预测模型的性能.

与 LSTM-Att-Manifold-Mixup 方法相比, 引入类别先验的 LSTM-Att-Prior-Mixup 方法在召回率和 F1 值上具有明显提升, 但其对准确率和 MP 值的影响并不明显, 在 Criminal-L 的准确率和宏精确率略有下降. 但总的来说引入类别先验有助于提高罪名预测的总体性能, 而不会对模型准确率造成过多不利影响. LSTM-Att-Prior-Mixup 方法针对小规模数据集 Criminal-S 的提升最显著, 相比 SECapsF1 提高了 6.8%, 而对于 Criminal-M 和 Criminal-L 数据集提升效果有所减弱, 其主要原因可能是随训练样本的增多, Mixup 方法合成样本的作用在减弱, 甚至成为一种不利于模型训练的噪声, 从而影响模型的准确率和 MP 值.

为进一步验证本文方法对低频罪名分类性能的改进作用, 本文针对不同频率罪名开展对比实验, 根据训练集中罪名的出现频率将罪名划分为 3 类, 出现次数不高于 10 的罪名被看作低频罪名, 出现次数高于 100 的罪名被看作高频罪名, 其余的作为中频罪名. 针对 Criminal-S 数据集的不同频率罪名预测实验结果见表 3.

由表 3 可以看出, 本文方法在高频、中频和低频罪名上的宏 F1 值均优于基线模型. 本文模型对低频罪名预测性能的提升尤为显著, 相比 SECaps 模型宏 F1 值提升达到 13.5%. 实验结果表明, 本文提出的数据增强策略不仅能大幅改进低频罪名

表 2　罪名预测对比实验结果
Table 2　Comparative experimental results

模型	Criminal-S				Criminal-M				Criminal-L			
	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Acc.	MP	MR	F1
TFIDF+SVM	85.8	49.7	41.9	43.5	89.6	58.8	50.1	52.1	91.8	67.5	54.1	57.5
CNN	91.9	50.5	44.9	46.1	93.5	57.6	48.1	50.5	93.9	66.0	50.3	54.7
CNN-200	92.6	51.1	46.3	47.3	92.8	56.2	50.0	50.8	94.1	61.9	50.0	53.1
LSTM	93.5	59.4	58.6	57.3	94.7	65.8	63.0	62.6	95.5	69.8	67.0	66.8
LSTM-200	92.7	66.0	58.4	57.0	94.4	66.5	62.4	62.7	95.1	72.8	66.7	67.6
Fact-Law Att	92.8	57.0	53.9	53.4	94.7	66.7	60.4	61.8	95.7	73.3	67.1	68.6
Few-Shot Attri	93.4	66.7	69.2	64.9	94.4	69.2	69.2	67.1	95.8	75.8	73.7	73.1
SECaps	94.8	71.3	70.3	69.4	95.4	71.3	70.2	69.6	96.0	81.9	79.7	79.5
LSTM-Att	95.2	75.1	74.4	73.5	95.9	75.9	76.6	75.2	96.6	**86.2**	79.5	80.8
LSTM-Att-Manifold-Mixup	95.3	73.7	75.3	73.3	96.3	80.1	79.1	79.5	**96.8**	85.8	81.5	82.3
LSTM-Att-Prior-Mixup	**95.3**	**76.7**	**78.2**	**76.2**	**96.3**	**80.8**	**82.0**	**80.1**	96.6	84.5	**84.9**	**83.3**

表 3　不同频率罪名预测宏 F1 值
Table 3　Macro F1 value of different frequency charges

模型	低频 (49类)	中频 (51类)	高频 (49类)
Few-Shot Attri	49.7	60.0	85.2
SECaps	53.8	65.5	89.0
LSTM-Att	54.1	65.0	90.1
LSTM-Att-Manifold-Mixup	64.2	66.5	89.5
LSTM-Att-Prior-Mixup	**67.3(↑13.5%)**	**67.8(↑2.3%)**	**90.0(↑1.0%)**

表 4　易混淆罪名预测宏 F1 值
Table 4　Macro F1 value for confusing charges

模型	F1值
LSTM-200	79.7
Few-Shot Attri	88.1
SECaps	90.5
LSTM-Att	91.8
LSTM-Att-Manifold-Mixup	**92.3**
LSTM-Att-Prior-Mixup	**92.1(↑1.6%)**

的分类效果, 对高频和中频罪名预测性能也有一定的促进作用. 其主要原因是合成样本有助平滑模型分类的决策面, 而类别先验引导的 Mixup 数据增强策略合成的数据有效增强了低频罪名的训练数据, 从而提高了模型对低频罪名的泛化能力.

为验证本文方法对易混淆罪名预测性能的改进, 本文选取 Criminal-S 数据集中 4 组典型的易混淆罪名开展实验, 它们分别是 “放火罪”与 “失火罪”、“抢夺罪”与 “抢劫罪”、“行贿罪”与 “受贿罪”、“盗伐林木罪”与 “滥伐林木罪”. 表 4 为现有方法与本文方法针对易混淆罪名的宏 F1 值.

由表 4 可以看出, 与基线模型相比, 本文方法对易混淆罪名的预测宏 F1 值获得了明显提高. 相比性能最好的 SECaps 模型, 本文方法在易混淆罪名上的宏 F1 值提升了 1.6%. 文本方法在易混淆罪名上与 LSTM-Att-Manifold-Mixup 模型性能相当, 宏 F1 值仅相差 0.2%. 实验结果表明, 在文本的嵌入空间中合成伪样本, 可以改进模型的泛化能力, 提升易混淆罪名预测结果.

4.4　不同编码器对比实验

为了验证本文提出的数据增强方法对不同编码器的适应性, 本文将模型中的文本编码器替换为 BERT 预训练语言模型, 并针对 Criminal-S 数据集进行了对比实验.

考虑到司法文本的领域特性, 本文采用清华大学人工智能研究院自然语言处理与社会人文计算研究中心提供的刑事文书 BERT 预训练语言模型[26]作为模型的编码层. 在实验中, 本文实现了两个基于 BERT 罪名预测模型, 其中 BERT-CLS 表示采用 [CLS] 对应向量作为文本表示的罪名预测模型; BERT-Att 表示在 BERT 输出的基础上采用结构化自注意力机制获取文本表示的罪名预测模型. 在微调 BERT 模型时, 作者根据实验发现将学习速率设为 1×10^{-4}, 并根据验证集的 F1 值选择最优模型时获得的性能最好. 此外, 由于受限于 GPU 的显存容量, BERT 模型训练的批次大小设为 32. 表 5 展示了不同编码器与不同 Mixup 数据增强策略结合后, 模型对测试集的预测 F1 值.

由表 5 的实验结果对比可以看出, 以 BERT 作为编码器的模型预测性能均低于采用 Bi-LSTM 作为编码器的模型. 在实验过程中发现采用 BERT 作为编码器的罪名预测模型存在严重的过拟合问题.

表 5 不同编码器对比实验结果
Table 5 Comparative experimental results of different encoder

模型	Criminal-S			
	Acc.	MP	MR	F1
BERT-CLS	93.4	65.6	63.1	63.2
BERT-CLS-Manifold-Mixup	93.6	69.2	69.5	67.6
BERT-CLS-Prior-Mixup	93.8	70.6	72.9	70.6
BERT-Att	93.6	68.5	69.7	67.2
BERT-Att-Manifold-Mixup	94.1	70.8	73.0	70.9
BERT-Att-Prior-Mixup	94.4	71.4	73.3	71.1
LSTM-Att-Prior-Mixup	**95.3**	**76.7**	**78.2**	**76.2**

在训练过程中, BERT 模型在训练集上的准确率上升很快, 在第 7 ~ 8 轮时模型对训练集的准确率达到 1, 但此时验证集的准确率为 94% 左右. 出现这一现象的原因可能是 BERT 模型参数量巨大, 在微调时模型过于偏向高频罪名, 从而导致模型的总体性能较差.

对比不同的 BERT 模型, BERT-Att 的性能要优于 BERT-CLS. 实验结果表明在预训练语言模型的基础上, 采用结构化注意机制有助于模型学习到更好的分类特征.

在数据增强策略方面, BERT-CLS 模型和 BERT-Att 模型在引入数据增强策略后, 模型性能均获得明显提升. 与 Manifold-Mixup 方法相比, 本文提出的类别先验 Mixup 数据增强策略可获得更好的预测性能.

实验结果表明, 本文提出的类别先验 Mixup 数据增强策略可适用于不同的文本分类模型, 同时有助于改进模型对类别不平衡文本分类数据的性能.

4.5 消融实验

类别先验 Mixup 数据增强策略和结构化自注意力机制是本文方法的重要组成部分. 为验证它们对罪名预测模型性能的影响, 本文进行了 2 组消融实验. 第 1 组实验从模型训练过程中移除类别先验 Mixup 数据增强策略, 该实验在表 6 标记为 LSTM-Att. 第 2 组实验在移除类别先验 Mixup 数据增强策略基础上, 将结构化自注意力层替换为最大池化层, 该实验标注记为 LSTM-Maxpool.

由表 6 的消融实验结果可发现, 移除类别先验 Mixup 数据增强策略后, 模型性能明显下降, 模型针对 Criminal-S 和 Criminal-M 两个数据集的准确率和宏精确率有所下降, 而 3 个数据集的 MR 平均下降了 4.9%, F1 值平均下降了 3.4%. 实验结果表明, 本文提出的类别先验 Mixup 数据增强策略对缓解罪名不平衡具有重要作用, 数据增强策略可显著提高模型的召回率和 F1 值, 而不会对模型的准确率和宏精确率造成过多的影响.

本文将结构化自注意力层替换为 Max-pooling 层后, 模型性能大幅下降, 准确率平均下降了 0.9%, MP 平均下降了 22.8%, MR 平均下降了 28.4%, F1 值平均下降了 25.8%, 该实验结果表明, 从文本中获取丰富的分类特征对于罪名预测模型的性能提升具有重要影响. 相比于最大池化层, 结构化自注意力机制能够更加有效地捕获不同侧面案情的文本特征, 从而大幅提高模型的性能.

从消融实验结果可看出, 在利用结构注意力获取有效罪名分类特征的基础上, 引入本文提出的类别先验 Mixup 数据增强策略可进一步提高罪名预测性能.

4.6 超参数的影响

本节讨论模型主要超参数对罪名预测性能的影响.

样本混合因子 λ 决定了样本的合成比例, 对伪样本的分布具有重要影响. 本文对比了不同 Beta 分布超参数对模型性能的影响. 图 4 展示了不同超参数 α 下模型对 Criminal-S 数据集的预测结果, 横坐标为超参数 α, 纵坐标为模型性能指标.

由图 4 的实验结果可以看出, 随着超参数 α 的增大模型的性能也逐步提高. 其原因是当 α 值较小时, 采样得到的 λ 值偏向于 0 或 1, 导致伪样本向量表示偏向其中一个样本, 影响了伪样本的多样性. 当 α 值增大时, 采样得到的 λ 值趋向于 0.5, 则样本表示在合成样本中的占比趋向于平均, 则合成样本

表 6 消融实验罪名预测结果
Table 6 Results of ablation experiments

模型	Criminal-S				Criminal-M				Criminal-L			
	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Acc.	MP	MR	F1
LSTM-Att-Prior-Mixup	**95.3**	**76.7**	**78.2**	**76.2**	**96.3**	**80.8**	**82.0**	**80.1**	**96.6**	84.5	**84.9**	**83.3**
LSTM-Att	95.2	75.1	74.4	73.5	95.9	75.9	76.6	75.2	**96.6**	**86.2**	79.5	80.8
LSTM-Maxpool	93.5	44.2	41.1	41.3	95.6	58.0	54.1	54.9	96.3	71.2	64.8	65.9

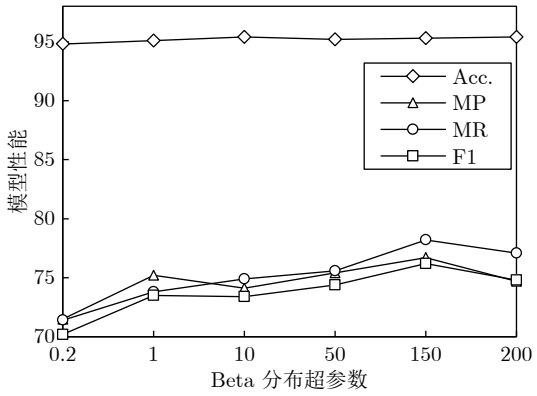


图 4　Beta 分布超参数的影响

Fig. 4　Impact of Beta distribution parameters

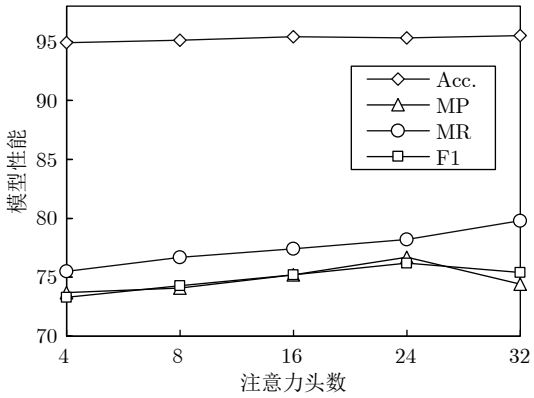


图 5　注意力头数的影响

Fig. 5　Impact of head number in attention Layer

在表示空间中分布更加均匀. 这样的数据分布有助于平滑模型的分类决策面, 提高模型的泛化能力. 当 α 超过 150 后模型性能有所下降, 其原因可能是采样得到的 λ 值接近 0.5 且方差很小, 这也会影响合成样本的多样性, 从而对模型训练造成不利影响.

当 α 超过 150 后模型性能有所下降, 其原因可能是采样得到的 λ 值接近 0.5 且方差很小, 这也会影响合成样本的多样性, 从而对模型训练造成不利影响.

结构化自注意力层的超参数 r 决定了文本表示的维度. 图 5 展示了不同 r 值对应的模型性能, 横坐标为 r 的值, 纵坐标为模型的性能指标.

由图 5 结果可以看出, 随着注意力头数 r 的增大, 文本表示包含的特征越来越丰富, 模型的性能也不断提升. 当 r 大于 24 后, 模型性能有所下降, 其原因可能是当 r 增大时, 模型的复杂度也随之增大, 导致模型过拟合, 降低了模型的泛化能力.

4.7　案例分析

本文通过可视化模型的注意力权重来分析罪名预测模型的分类依据, 并通过对比 LSTM-Att 模型和 LSTM-Att-Prior-Mixup 的注意力差异对数据增强策略对模型的影响进行分析.

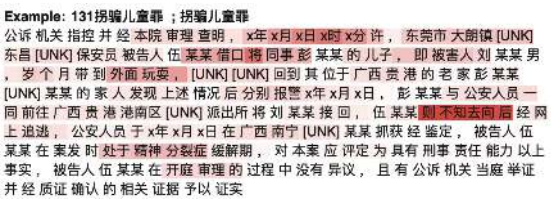
图 6 展示了模型对低频罪名 “拐骗儿童罪”案例的注意力权重分布情况. 从总体上看, 2 个模型都关注到了案情描述中比较重要的词语, 比如 “不知去向” “借口”等. 但是, 相比于 LSTM-Att-Prior-Mixup 模型, LSTM-Att 模型的注意力更加分散, 它还关注了许多与罪名分类无关的词语, 如 “港南区” “评定”等. 可能正是由于这些注意力的分散导致 LSTM-Att 模型将该案件错分为 “非法拘禁罪”.

图 7 展示了模型对易混淆罪名 “行贿罪”案例的注意力权重分布情况. 与低频罪名案例中的情况类似, 两个模型都关注到了案情描述中比较重要的词语, 比如 “行贿” “收受” “x 万”等与行贿、受贿紧密相关的词语. 然而, 相比之下 LSTM-Att 模型的注意力更加分散, 从而导致模型将该案件错分为 “受贿罪”.

综上所述, 类别先验 Mixup 有助于模型学习到更优的注意力机制, 使得模型关注的词语更加集中, 从而提高了模型对低频罪名和易混淆罪名的预测能力.

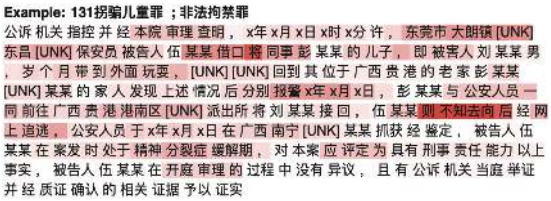
5　结束语

本文将 Mixup 数据增强策略引入到罪名预测



(a) LSTM-Att 模型注意力可视化

(a) Visualization of attention mechanism of LSTM-Att Model

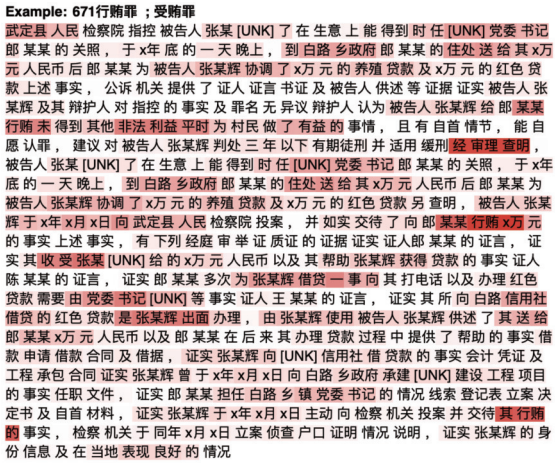


(b) LSTM-Att-Prior-Mixup 模型注意力可视化

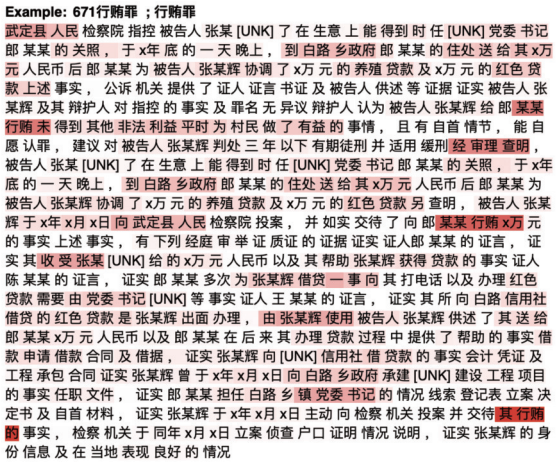
(b) Visualization of attention mechanism of LSTM-Att-Prior-Mixup Model

图 6　低频罪名案例

Fig. 6　Sample of low frequency charge



(a) LSTM-Att 模型注意力可视化
(a) Visualization of attention mechanism of LSTM-Att Model



(b) LSTM-Att-Prior-Mixup 模型注意力可视化
(b) Visualization of attention mechanism of LSTM-Att-Prior-Mixup Model

图 7 易混淆罪名案例
Fig. 7 Sample of confusing charge

任务中, 并针对罪名不平衡问题提出了类别先验 Mixup 数据增强策略, 有效缓解了类别不平衡带来的影响, 提高了低频罪名和易混淆罪名的分类性能; 相比已有方法, 本文提出的类别先验 Mixup 数据增强方法简单有效, 无需额外的人工标注, 也不需要引入辅助任务.

本文主要关注于改进低频罪名预测性能, 并针对单罪名预测问题验证了所提方法的有效性, 而数罪并罚情况下的 Mixup 数据增强策略将在下一步工作中进行研究.

References

1 Zhong H X, Xiao C J, Tu C C, Zhang T Y, Liu Z Y, Sun M S. How does NLP benefit legal system: A summary of legal artificial intelligence. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual Event: 2020. 5218–5230

2 Kort F. Predicting supreme court decisions mathematically: A quantitative analysis of the 'Right to counsel' cases. *The American Political Science Review*, 1957, **51**(1): 1–12

3 Mackaay E, Robillard P. Predicting judicial decisions: The nearest neighbour rule and visual representation of case patterns. *De Gruyter*, 1974, **41**: 302–331

4 Liu C L, Chang C T, Ho J H. Case instance generation and refinement for case-based criminal summary judgments in chinese. *Journal of Information Science and Engineering*, 2004, **20**(4): 783–800

5 Xiao C J, Zhong H X, Guo Z P, Tu C C, Liu Z Y, Sun M S, et al. CAIL2018: A large-scale legal dataset for judgment prediction, arXiv preprint, 2018, arXiv: 1807.02478

6 Zhong H X, Guo Z P, Tu C C, Xiao C J, Liu Z Y, Sun M S. Legal judgment prediction via topological learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: 2018. 3540–3549

7 Yang W, Jia W, Zhou X J. Legal judgment prediction via multi-perspective bi-feedback network. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: 2019. 4085–4091

8 Wang Wen-Guang, Chen Yun-Wen, Cai Hua, Zeng Yan-Neng, Yang Hui-Yu. Judicial document intellectual processing using hybrid deep neural networks. *Journal of Tsinghua University (Science and Technology)*, 2019, **59**(7): 505–511
(王文广, 陈运文, 蔡华, 曾彦能, 杨慧宇. 基于混合深度神经网络模型的司法文书智能化处理. 清华大学学报(自然科学版), 2019, **59**(7): 505–511)

9 Jiang X, Ye H, Luo Z C, Chao W H, Ma W J. Interpretable rationale augmented charge prediction system. In: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New-Mexico, USA: 2018. 149–151

10 Liu Zong-Lin, Zhang Mei-Shan, Zhen Ran-Ran, Gong Zuo-Quan, Yu Nan, Fu Guo-Hong. Multi-task learning model for legal judgment predictions with charge keywords. *Journal of Tsinghua University (Science and Technology)*, 2019, **59**(7): 497–504
(刘宗林, 张梅山, 甄冉冉, 公佐权, 余南, 付国宏. 融入罪名关键词的法律判决预测多任务学习模型. 清华大学学报(自然科学版), 2019, **59**(7): 497–504)

11 Xu N, Wang P, Chen L, Pan L, Wang X Y, Zhao J Z. Distinguish confusing law articles for legal judgment prediction. In: Proceedings of the 58th Annual Meeting of the Association-for-Computational-Linguistics, Virtual Event: 2020. 3086–3095

12 Hu Z K, Li X, Tu C C, Liu Z Y, Sun M S. Few-shot charge prediction with discriminative legal attributes. In: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New-Mexico, USA: 2018. 487–498

13 He C Q, Peng L, Le Y Q, He J W, Zhu X Y. SECaps: A sequence enhanced capsule model for charge prediction. In: Proceedings of the 28th International Conference on Artificial Neural Networks. Munich, Germany: Springer Verlag, 2019. 227–239

14 Zhang H, Cisse M, Dauphin Y N, David L P. Mixup: Beyond empirical risk minimization. arXiv preprint, 2017, arXiv: 1710.09412

15 Verma V, Lamb A, Beckham C, Najafi A, Mitiagkas I, Lopez-Paz D, et al. Manifold mixup: Better representations by interpolating hidden states. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, CA, USA: 2019. 11196–11205

16 Lin Z H, Feng M W, Santos C N, Yu M, Xiang B, Zhou B, et al. A structured self-attentive sentence embedding. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France: 2017. 1–15

17 Guo H Y, Mao Y Y, Zhang R C. Augmenting data with mixup for sentence classification: An empirical study. arXiv preprint, 2019, arXiv: 1905.08941

18 Chen J A, Yang Z C, Yang D Y. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Virtual Event: 2020. 2147–2157

19 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, **9**(8): 1735–1780

20 Devlin J, Chang M W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceeding of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, USA: 2019. 4171–4186

21 Kingma D P, Ba J L. Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: 2015. 1–15

22 Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 1988, **24**(5): 513–523

23 Suykens J A K, Vandewalle J. Least squares support vector machine classifiers. *Neural Processing Letters*, 1999, **9**(3): 293–300

24 Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 19th conference on Empirical Methods in Natural Language. Doha, Qatar: 2014. 1746–1751

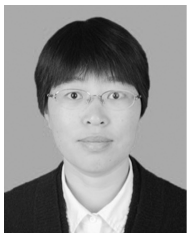
25 Luo B F, Feng Y S, Xu J B, Zhang X, Zhao D Y. Learning to predict charges for criminal cases with legal basis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: 2017. 2727–2736

26 Zhong H X, Zhang Z Y, Liu Z Y, Sun M S. Open chinese language pre-trained model zoo [Online], available: https://github.com/thunlp/openclap, March 6, 2021.

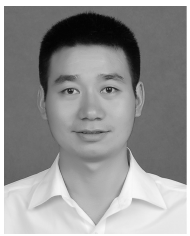
线岩团 昆明理工大学信息工程与自动化学院副教授. 主要研究方向为自然语言处理, 信息抽取和机器翻译.
E-mail: xianyt@kust.edu.cn
(**XIAN Yan-Tuan** Associate professor at the School of Information Engineering and Automation, Kunming University of Science and Technology. His research interest covers natural language processing, information extraction and machine translation.)

陈文仲 昆明理工大学信息工程与自动化学院硕士研究生. 主要研究方向为自然语言处理和信息检索.
E-mail: Chen_WenZhong@163.com
(**CHEN Wen-Zhong** Master student at the School of Information Engineering and Automation, Kunming University of Science and Technology. His research interest covers natural language processing and information retrieval.)

余正涛 昆明理工大学信息工程与自动化学院教授. 主要研究方向为自然语言处理, 信息检索, 机器翻译和机器学习. 本文通信作者.
E-mail: ztyu@hotmail.com
(**YU Zheng-Tao** Professor at the School of Information Engineering and Automation, Kunming University of Science and Technology. His research interest covers natural language processing, information retrieval, machine translation, and machine learning. Corresponding author of this paper.)



张亚飞 昆明理工大学信息工程与自动化学院副教授. 主要研究方向为自然语言处理和模式识别.
E-mail: zyfeimail@163.com
(**ZHANG Ya-Fei** Associate professor at the School of Information Engineering and Automation, Kunming University of Science and Technology. Her research interest covers natural language processing and pattern recognition.)



王红斌 昆明理工大学信息工程与自动化学院副教授. 主要研究方向为自然语言处理和信息抽取.
E-mail: wanghongbin@kust.edu.cn
(**WANG Hong-Bin** Associate professor at the School of Information Engineering and Automation, Kunming University of Science and Technology. His research interest covers natural language processing and information extraction.)