

# 基于预训练模型与知识蒸馏的法律判决预测算法

潘瑞东<sup>1</sup>, 孔维健<sup>1,2†</sup>, 齐洁<sup>1,2</sup>

(1. 东华大学 信息科学与技术学院, 上海 201600; 2. 东华大学 数字化纺织服装技术教育部工程研究中心, 上海 201620)

**摘要:** 针对法律判决预测中罪名预测和法条推荐子任务, 提出了基于BERT (Bidirectional Encoder Representation from Transformers) 预训练模型与知识蒸馏策略的多任务多标签文本分类模型. 为挖掘子任务间的关联, 提高预测准确率, 运用BERT预训练模型进行多任务学习, 建立了BERT<sub>12</sub>multi文本分类模型; 针对罪名、法条类别中的样本不均衡问题, 采用分组的焦点损失 (Focal Loss) 以增强模型对于罕见罪名及法条的辨别能力; 为降低模型计算复杂度并且提高模型推理速度, 提出了一种以教师模型评价为参考的知识蒸馏策略, 通过动态平衡蒸馏中的蒸馏损失和分类损失, 将BERT<sub>12</sub>multi压缩为浅层结构的学生模型. 综上, 构建出可以处理不均衡样本且具有较高推理速度的多任务多标签文本分类模型BERT<sub>6</sub>multi. 在CAIL2018数据集上的实验表明, 采用预训练模型及分组Focal Loss可显著提高法律判决预测的性能; 通过融入教师模型评价, 知识蒸馏得到的学生模型推理速度提高近一倍, 并且在罪名预测及法条推荐任务中获得86.7%与83.0%的F1-Score (Micro-F1与Macro-F1的均值).

**关键词:** 法律判决预测; 预训练模型; 焦点损失; 多任务学习; 模型压缩; 知识蒸馏

中图分类号: TP273

文献标志码: A

DOI: 10.13195/j.kzyjc.2020.0985

开放科学(资源服务)标识码(OSID):



## Legal Judgment Prediction based on Pre-training Model and Knowledge Distillation

Pan Rui-dong<sup>1</sup>, Kong Wei-jian<sup>1,2†</sup>, Qi Jie<sup>1</sup>

(1. School of Information Science and Technology, Donghua University, Shanghai 201600, China 2. Engineering Research Center of Digitized Textile and Fashion Technology of Ministry Education, Donghua University, Shanghai 201620, China)

**Abstract:** Based on the BERT pre-training model and knowledge distillation, a multi-task and multi-label text classification model is proposed for two sub-tasks of the legal judgment prediction, namely, charge prediction and law article recommendation. To find the correlation between two sub-tasks and improve the performance of prediction, a text classification model named BERT<sub>12</sub>multi is formulated by multi-task learning based on a BERT pre-training model. The hierarchical Focal Loss is introduced to improve the ability of distinguishing the charges and law articles, which are sampled imbalanced. In order to reduce the computing complexity and increase the speed of the inference, we propose a knowledge distillation strategy based on the evaluation of the teacher model. The strategy compresses BERT<sub>12</sub>multi into a student model with a shallow structure by balancing between the classification loss and the distillation loss dynamically. Hence, a multi-task and multi-label text classification model with higher inference speed named BERT<sub>6</sub>multi is introduced, which can deal with the imbalance problem of samples. Experiments on the CAIL2018 dataset show that the pre-training model and hierarchical Focal Loss can improve the performance of the prediction algorithm effectively. Combined with our knowledge distillation strategy, the inference speed of the student model is nearly doubled. The F1-Scores (mean value of Micro-F1 and Macro-F1) for charge prediction and law article recommendation are 86.7% and 83.0%.

**Keywords:** legal judgment prediction; pre-training model; Focal Loss; multi-task learning; model compression; knowledge distillation

收稿日期: 2020-07-18; 修回日期: 2020-11-17.

基金项目: 国家自然科学基金项目(61773112、61603088)、流程工业综合自动化国家重点实验室开放课题基金资助项目.

†通讯作者. E-mail: kongweijian@dhu.edu.cn.

## 0 引言

法律判决预测是人工智能技术尤其是自然语言处理方法在司法领域的典型应用,也是推动司法智能化的主要途径.在司法过程中应用法律判决预测技术,并不直接代替法官判决,而是通过提供定罪、量刑的参考,辅助法判决,提升法官工作效率.提高法律判决预测算法的准确性,既可以有效帮助法官从大量事务性工作中解放,也可以应用于法律咨询,以较低的人力成本,给予没有法律背景知识人群相应的法律指导与援助.

由于罪名及法条数目众多且相对固定,一般将法律判决预测任务作为文本分类任务进行处理.但罪名预测及法条推荐任务的类别数目显著多于一般文本分类任务,且存在如“抢劫罪”与“抢夺罪”等定义差别较小的易混淆类别.故罪名预测和法条推荐相较情感分类等常见文本分类任务更加复杂,对文本分类模型有更高的性能要求.

早期的判决预测通常基于已有案件判决结果的统计分析,一些工作将词频统计与机器学习方法集成到法律判决预测任务中<sup>[1-3]</sup>.如Sulea等<sup>[3]</sup>提出的基于支持向量机(squares support vector machine, SVM)<sup>[4]</sup>法律判决预测模型.

随着文本向量化技术如word2vec<sup>[5]</sup>、GloVe<sup>[6]</sup>、FastText<sup>[7]</sup>等的提出,相较于从文本描述中提取浅层文本特征和词频信息的词频统计算法(如TF-IDF<sup>[8]</sup>、Text-Rank<sup>[9]</sup>),采用词嵌入算法,可以将文本表示映射成固定维度的词向量编码,提供更准确的文本向量化表达.研究者们尝试将词嵌入算法结合深度学习模型来处理法律判决任务<sup>[10-13]</sup>.

Yang等<sup>[11]</sup>针对法律文本中的数值单位关键词,如酒精的含量、毒品的重量、盗窃的金额等,设计了具有多视角前向预测和后向验证的双反馈机制用于匹配文本中的数字搭配信息,提高了模型捕捉数字和关键词搭配信息的能力;刘宗林<sup>[12]</sup>等采用关键词抽取算法挖掘裁判文书中的罪名关键词,融入深度学习模型,提出了MTL-Fusion模型,有效提高了模型对于易混淆罪名的辨别能力;王文广等<sup>[13]</sup>将深度学习文本分类模型HAN<sup>[14]</sup>和DPCNN<sup>[15]</sup>应用于法律判决预测,并针对具体判决任务对模型进行融合改进,提出了一种基于混合深度神经网络模型HAC (Hybrid attention and CNN model),在各项判决预测任务中表现出色.

近年来,在自然语言处理领域的研究表明,与基于词嵌入的深度学习模型相比,在大规模未标记的语料上预训练,并针对具体的数据集和下游

任务进行微调的无监督预训练模型,可以显著提高深度学习模型在包括文本分类等各项自然语言处理任务中的表现.以BERT<sup>[16]</sup>为代表的预训练模型(如ELMo<sup>[17]</sup>,XLNet<sup>[18]</sup>等)提出以来,在医疗、广告推荐、搜索等诸多领域中的应用取得了重要进展.将预训练模型应用于法律判决预测任务,可以进一步提高预测的准确性和可靠性.

Chalkidis等<sup>[19]</sup>提出了基于BERT预训练模型的HIER-BERT模型用于处理超长案件文本.在欧洲人权法院数据集中罪名二分类、法条多标签分类、案件重要性预测三项法律判决预测子任务中的表现与其他机器学习和深度学习模型如(BiGRU<sup>[20]</sup>、HAN等)相比性能更优.

预训练模型往往有庞大的参数量,以BERT为例,由12层Transformer编码器<sup>[21]</sup>组成的BERT-base预训练模型有超过1亿的参数量,由24层Transformer编码器组成的BERT-large预训练模型有超过3.3亿的参数量.在实际应用中,预训练模型存在计算效率低,资源消耗大的缺陷.研究者主要通过知识蒸馏<sup>[22-24]</sup>等策略,压缩预训练模型体积.如Sun等<sup>[24]</sup>针对BERT等预训练模型提出了耐心的知识蒸馏策略(Patient Knowledge Distillation, PKD),通过教师网络的深层结构提取丰富信息,在蒸馏过程中加入来自教师模型中间层的监督,提高了学生模型的性能.

本文的主要贡献如下:

1) 针对法律判决整体复杂度较高、罪名预测及法条推荐任务关系紧密的特性,将预训练模型BERT应用于罪名预测及法条推荐任务中,并采用联合建模的方式,建立了多任务多标签文本分类模型BERT<sub>12</sub>multi.

2) 针对罪名及法条类别众多,且类别间样本不均衡、部分类别正负样本比例悬殊的问题,采用一种分组的Focal Loss<sup>[25]</sup>策略,对样本数量接近的类别设置相同的Focal Loss超参数,进一步提高了BERT<sub>12</sub>multi模型性能,在罪名预测、相关法条推荐任务中F1-Score分别为87.0%和83.2%.

3) 为提高模型实用性,针对预训练模型参数冗余导致BERT<sub>12</sub>multi推理速度较慢的问题,提出一种基于教师模型评价的动态权重知识蒸馏策略,将教师模型在训练数据<sup>[26]</sup>中的表现作为平衡学生模型蒸馏损失及分类损失的依据.经知识蒸馏获得的学生模型BERT<sub>6</sub>multi与教师模型BERT<sub>12</sub>multi的性能接近(F1-Score为86.7%和83.0%),但隐藏层数仅为教师模型的一半,计算量减少了近50%,显著提高了计算效率.

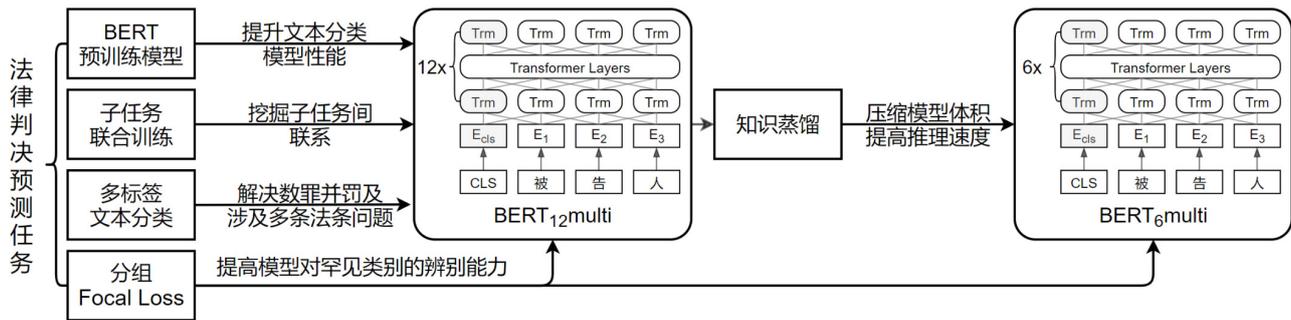


图1 算法流程图

## 1 问题描述

罪名预测及法条推荐任务定义如下:

**罪名预测:**根据刑事法律文书中的案情描述和事实部分,预测被告人被判的罪名

**法条推荐:**根据刑事法律文书中的案情描述和事实部分,预测本案涉及的相关法条

实验采用的数据集共涉及183项刑事罪名、202项刑法条例,存在被告人同时触犯数条法条,或者数罪并罚的情况.关于数据集的详细描述见4.1节.

## 2 基于预训练模型及知识蒸馏的多任务学习模型

依据前文及现有研究,法律判决预测相较一般文本分类任务具有较高复杂度,类别数目显著较多,存在数罪并罚的现象,且罪名预测与法条推荐子任务间具有关联性;罪名及法条存在类别间样本不平衡问题,常见罪名及法条样本数目远多于罕见罪名及法条.针对以上问题,提出了以下方案:

1) 采用BERT预训练模型进行多标签文本分类提高整体性能,并采用对罪名预测、法条推荐子任务联合训练的方式挖掘任务间的关联性.

2) 将罪名、法条类别按样本数量分组,采用分组的Focal Loss提高了模型对于少样本类别的分类能力.

3) 由于BERT预训练模型参数量较大,且推理速度较慢,尝试通过知识蒸馏压缩模型体积、提高模型推理速度.

任务流程如图1所示.

### 2.1 BERT<sub>12multi</sub>多任务多标签文本分类模型

过去对于法律判决预测的研究中,往往只考虑一项罪名或者法律条例,将罪名预测、法条推荐任务作为多类别文本分类任务.由于在一条案件的审判中往往涉及多条相关法律,且被告人存在数罪并罚的情况,将罪名预测、法条推荐任务作为多标签文本分类任务进行处理,同时对所有罪名、法条类别做二

元分类,计算案件文本属于每一类别的概率.

罪名预测、法条推荐之间具有高度相关性,如刑法第264条“盗窃公私财物,数额较大或者多次盗窃的,处三年以下有期徒刑、拘役或者管制,并处或者单处罚金……”中定义了盗窃罪,与盗窃罪有直接关系.考虑针对不同任务独立建模无法捕捉任务间关联性,及训练多个独立模型的资源代价,故对罪名预测、法条推荐任务进行多任务学习,采用联合建模的方式,在单个模型中完成两项法律判决任务的预测.

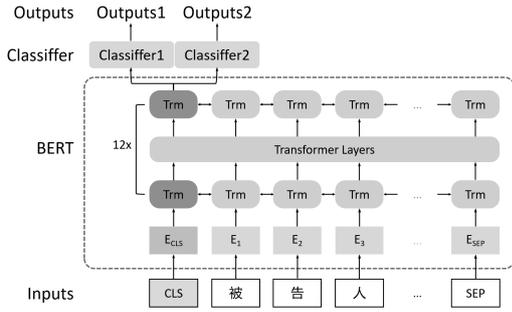
模型编码器端采用基于12层Transformer结构的BERT-base预训练模型,使用中文预训练权重BERT-wwm<sup>[27]</sup>初始化参数,解码器端为2个神经元数目分别为罪名数目及法条数目的全连接层组成的分类器,使用sigmoid函数对分类器每个神经元的输出进行二分类概率映射.在编码器与解码器间采用Dropout策略<sup>[28]</sup>,在模型训练时随机使一部分参数“失活”防止过拟合. Dropout策略仅在预训练模型微调训练时激活,在模型预测以及知识蒸馏时不采用.使用微调的方式对罪名预测、法条推荐任务进行建模.

将经过预处理后的案件文本补齐或截断为统一长度,依据BERT的预训练词表转换为索引序列矩阵,按批次输入BERT模型,取经BERT模型编码后[CLS]特殊符号位置处对应的输出向量作为输入文本的句向量输入分类器,经Dropout后,将文本句向量经过分类器映射为对应类别,即:

$$h_i = \text{BERT}(x_i; \theta) \quad (1)$$

$$\hat{y}_i = \sigma(\text{Dropout}(W \cdot h_i)) \quad (2)$$

$h_i$ 表示样本 $x_i$ 输入BERT模型,在[CLS]标记对应位置处的输出向量; $W$ 是分类器的权重矩阵; $\sigma$ 表示sigmoid激活函数, $\hat{y}_i$ 表示样本 $x_i$ 在各类别中的映射概率.为与知识蒸馏时采用6层Transformer编码层结构的学生模型进行区分,将此多任务学习模型命名为BERT<sub>12multi</sub>,模型结构如图2所示.

图2 BERT<sub>12multi</sub>模型

## 2.2 分组Focal Loss

在神经网络的训练中,对于二分类或多标签分类问题一般采取二分类交叉熵(Cross Entropy Loss)作为损失函数,即:

$$L_{CE}(x_i) = \begin{cases} -\log \hat{y}_i & y_i = 1 \\ -\log(1 - \hat{y}_i) & y_i = 0 \end{cases} \quad (3)$$

其中 $y_i$ 为样本 $x_i$ 标签结果,1表示标签为正样本,0表示标签为负样本, $\hat{y}_i$ 为模型预测结果.二分类交叉熵损失函数对于正负样本没有倾向性,当数据集中负样本远多于正样本时,普通交叉熵损失函数会使得分类结果趋向于负样本;另一方面,对于多标签或多分类问题,不同类别的样本分布不均衡会导致分类结果倾向于样本数较多的类别,使得样本数较少的类别召回率低于样本数较多类别,影响最终分类结果.

常见的罪名及法条占据数据中的大部分,部分罕见罪及法条样本稀少.本文采用的CAIL2018-Small数据集<sup>[26]</sup>涉及202项刑事罪名以及183条刑法条例,平均一个案件涉及1.36项罪名及1.58条刑法条例.对于所有类别,其正样本数都远小于负样本数,且类别间存在样本不均衡问题.罪名及法条类别分布呈现长尾特性,以罪名为例,如图3所示,“盗窃罪”等十项最高频罪名覆盖训练集约33%的案例,而“走私罪”等十项低频罪名仅仅覆盖训练集约0.1%的案例.针对正负样本类别不平衡问题,考虑采用Focal Loss<sup>[25]</sup>作为分类损失函数:

$$L_{FL}(\mathbf{x}_i) = \begin{cases} -\alpha(1 - \hat{y}_i)^\gamma \log \hat{y}_i & y_i = 1 \\ -(1 - \alpha)\hat{y}_i^\gamma \log(1 - \hat{y}_i) & y_i = 0 \end{cases} \quad (4)$$

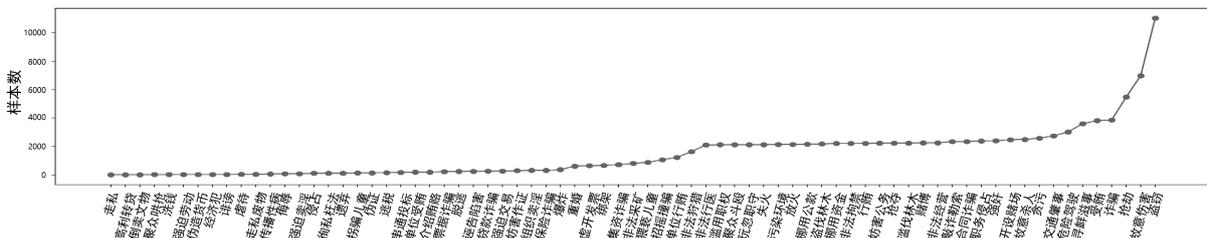


图3 部分罪名分布

$\alpha$ 与 $\gamma$ 为平衡类别正负类别损失的超参数,其中 $\alpha$ 用于平衡正负样本权重, $\gamma$ 用于调整简单样本的权重; $\alpha$ 设置越大,类别正样本权重越高, $\gamma$ 设置越大,对于困难样本权重越高;当 $\alpha$ 设置为0.5, $\gamma$ 设置为0时,Focal Loss与交叉熵效果一致.

由于法律判决预测为多标签文本分类,故将原二分类Focal Loss扩展为多标签分类:

$$l_{FL}(x_i \in c) = \begin{cases} -\alpha_c(1 - \hat{y}_{ic})^{\gamma_c} \log \hat{y}_{ic} & y_{ic} = 1 \\ -(1 - \alpha_c)\hat{y}_{ic}^{\gamma_c} \log(1 - \hat{y}_{ic}) & y_{ic} = 0 \end{cases} \quad (5)$$

$$L_{FL} = \frac{1}{C} \sum_{c=1}^C l_{FL}(x_i \in c) \quad (6)$$

$l_{FL}(x_i \in c)$ 表示样本在类别 $c$ 上的分类损失, $C$ 为类别总数,最终样本 $x_i$ 取所有类别上损失的均值为最终损失.原则上对于每一个类别都可以设置不同的 $\alpha_c$ 与 $\gamma_c$ .为避免超参数过多造成实验调参困难,依据样本数量将罪名与法条类别进行分箱,将样本数接近的类别设置相同超参数,实现分组Focal Loss,分组结果如表1所示.

表1 罪名及法条分组

	样本条件数	类别数目标	$\alpha$	$\gamma$	代表类别
罪名	< 100	59	0.7	3	走私、洗钱
	[100, 500)	56	0.7	2	串通投标
	[500, 2000)	25	0.5	2	绑架、重婚
	$\geq 2000$	62	0.3	2	盗窃、赌博
法条	< 100	46	0.7	3	第326条
	[100, 500)	55	0.7	2	第375条
	[500, 2000)	18	0.5	2	第210条
	$\geq 2000$	64	0.3	2	第310条

## 2.3 融入教师模型评价的知识蒸馏策略

BERT<sub>12multi</sub>中含有超过1亿个参数,在预测时会消耗大量资源,这阻碍了其在计算资源有限的实践中的应用.因此,尝试在保证性能的前提下,采用知识蒸馏策略<sup>[22]</sup>,将BERT<sub>12multi</sub>作为教师模型进行知识蒸馏,减少模型冗余的参数,提高模型的推理速度.

首先初始化一个较浅的学生模型BERT<sub>6</sub>multi, 其具备6层Transformer编码器, 其余参数如分类器、隐藏层维度等与BERT<sub>12</sub>multi保持一致, 并用BERT<sub>12</sub>multi前6层对学生模型参数初始化。

知识蒸馏通过让学生模型模仿教师模型的输出分布使学生模型获得与教师模型相近的性能. 在模型的输出中引入包含蒸馏温度参数 $T$ 的 $softmax$ 层以软化分布. 以教师模型 BERT<sub>12</sub>multi为例, 在罪名预测任务(记为task1)中, 其软化后的概率 $P_{task1}^t$ 可表示为:

$$\begin{aligned} P_{task1}^t(\hat{y}_{task1} | x_i) &= softmax\left(\frac{W_1^t h_i^t}{T}\right) \\ &= softmax\left(\frac{W_1^t \cdot BERT^t(x_i; \theta^t)}{T}\right) \end{aligned} \quad (7)$$

其中BERT<sup>t</sup>为BERT<sub>12</sub>multi中微调后的预训练编码器,  $\theta^t$ 为教师模型中的参数,  $T$ 是知识蒸馏中的温度参数, 较高的温度参数可以产生更多样化的概率分布.  $P_{task2}^t$ 为法条推荐任务中教师模型的输出概率分布, 除了类别数以及分类器参数以外, 与罪名预测一致. 教师模型的软标签 $\hat{y}_i^t$ 表示为:

$$\hat{y}_i^t = (P_{task1}^t(\hat{y}_{task1,i} | x_i), P_{task2}^t(\hat{y}_{task2,i} | x_i)) \quad (8)$$

衡量教师模型与学生模型在预测分布上差距的蒸馏损失可以定义为:

$$\begin{aligned} L_{DS} &= L_{DStask1} + L_{DStask2} \\ &= - \sum_{j \in [1,2]} \sum_{i \in [N]} \sum_{c_j \in C_j} [P_{taskj}^t(\hat{y}_{taskj,i} = c_j | x_i; \theta^t) \cdot \\ &\quad \log P_{taskj}^s(\hat{y}_{taskj,i} = c_j | x_i; \theta^s)] \end{aligned} \quad (9)$$

其中 $j$ 表示任务,  $c_j$ 和 $C_j$ 是相应任务中的类标签和类标签集,  $\theta^s$ 表示学生模型中的参数,  $P_{taskj}^s$ 表示学生模型在相同任务中输出的相应概率分布。

除了模仿老师模型的输出分布以外, 学生模型同时在任务中进行有监督微调, 本文中学生模型采取与教师模型相同的Focal Loss损失:

$$L_{FL}^s = L_{FLtask1}^s + L_{FLtask2}^s \quad (10)$$

因此, 知识蒸馏策略中目标函数可以表述为:

$$L_{KD} = (1 - \alpha)L_{FL}^s + \alpha L_{DS} \quad (11)$$

其中 $\alpha \in (0, 1)$ , 是平衡分类损失和蒸馏损失重要性的超参数, 在训练前固定.  $\alpha$ 越大, 表示知识蒸馏过程更依赖教师模型与学生模型之间概率分布的差距;  $\alpha$ 越小, 表示蒸馏更依赖训练样本标签的监督。

实验发现, 采用固定的在知识蒸馏过程中会造成一定的性能损失. 法律判决预测任务相较一般的文本分类任务类别数显著较多, 且存在如“抢劫罪”

与“抢夺罪”、“盗窃罪”与“侵占罪”等易混淆类别. 因此教师模型BERT<sub>12</sub>multi对于某些易混淆或数据稀少的罪名或法条类别, 即使在训练集中也存在多判、漏判、错判。

若采用固定的参数会导致学生模型学习到某些教师模型的错误信息, 故考虑将教师模型在训练数据中的表现作为依据, 对于教师模型预测准确的数据, 提高教师模型在知识蒸馏中的权重, 鼓励学生模型模仿教师模型的预测分布; 对于教师模型预测出现偏差的数据, 则降低教师模型的权重, 让学生模型接收更多来自数据标签的监督. 故采用一种动态的权衡函数, 动态平衡分类损失与蒸馏损失:

$$\begin{aligned} d(L_{FL}^t(x_i)) &= a \cdot Scaler(|L_{FL}^t(x_i)|) + b \\ &= a \cdot \frac{\max(|L_{FL}^t(x \in N)|) - |L_{FL}^t(x_i)|}{\max(|L_{FL}^t(x \in N)|) - \min(|L_{FL}^t(x \in N)|)} + b \end{aligned} \quad (12)$$

其中 $L_{FL}^t(x_i)$ 表示样本 $x_i$ 在教师模型上的分类损失函数;  $Scaler$ 为归一化函数;  $a$ 和 $b$ 为控制归一化后 $d(L_{FL}^t(x_i))$ 取值范围的超参数, 本文取 $a=0.5, b=0.2$ , 即控制 $d(L_{FL}^t(x_i)) \in [0.2, 0.7]$ . 在罪名预测以及法条推荐任务中分别计算 $d(L_{FLtask1}^t(x_i))_{task2}$ 和 $d(L_{FLtask2}^t(x_i))_{task2}$ .

采用动态权重的蒸馏损失可表示为:

$$L_{KD-dynamic} = (1 - d(L_{FL}^t(x_i))) L_{FL}^s + d(L_{FL}^t(x_i)) L_{DS} \quad (13)$$

参考Sun等<sup>[24]</sup>提出的PKD-Skip蒸馏策略, 在蒸馏过程中融入来自教师模型的中间隐藏层[CLS]向量的监督, 可以使学生模型可以有效学习教师模型隐藏层中的信息, 提高模型学生模型性能. 参考PKD-Skip蒸馏策略, 在知识蒸馏中提取出教师模型中间层表示, 在知识蒸馏中融入教师模型第2、4、6、8、10层隐藏层[CLS]的向量表示, 引入额外中间层分布损失:

$$L_{PT} = \sum_{i=1}^N \sum_{l=1}^L \left\| \frac{h_{i,l}^s}{\|h_{i,l}^s\|_2} - \frac{h_{i,I_{pt}(l)}^t}{\|h_{i,I_{pt}(l)}^t\|_2} \right\|_2^2 \quad (14)$$

其中 $L$ 表示学生网络中的层数,  $N$ 表示训练样本数,  $h_{i,I_{pt}(l)}^t$ 和 $h_{i,l}^s$ 表示教师模型和学生模型在对应隐藏层[CLS]位置处对应向量的表示. 引入中间层分布损失的最终蒸馏损失为:

$$L_{PKD-dynamic} = (1 - d(L_{FL}^t(x_i))) L_{FL}^s + d(L_{FL}^t(x_i)) L_{DS} + \beta L_{PT} \quad (15)$$

其中 $\beta$ 为中间层分布损失的权重. 学生模型BERT<sub>6</sub>multi及蒸馏过程如图4所示。

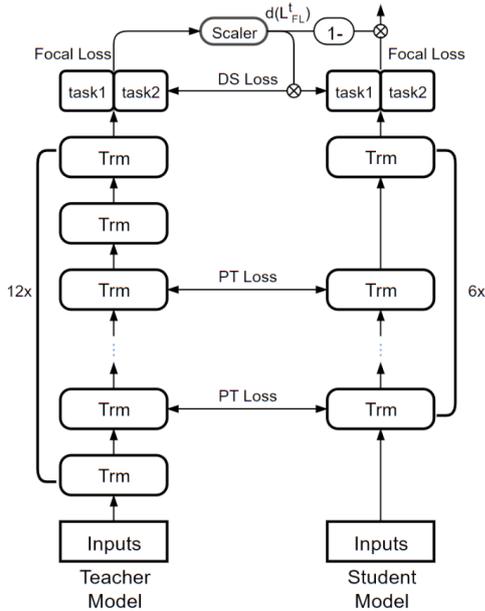


图4 知识蒸馏过程

### 3 实验

为比较模型的性能,采用CAIL2018-Small数据集作为验证数据集,数据集及评价指标介绍详见3.1、3.2节。为验证各环节对模型性能提升的贡献,及对对比知识蒸馏策略对模型性能及推理速度的影响,分阶段进行了如下实验:

1) 对比经典机器学习文本分类模型及常见深度学习文本分类模型联合、非联合训练与采用预训练模型联合、非联合训练的性能差异,验证引入预训练模型及联合训练对于任务性能的提升。

2) 对比采用交叉熵损失与Focal Loss条件下各深度学习模型性能差异,验证采用分组Focal Loss对于模型性能的影响。

3) 将不同蒸馏策略下的学生模型BERT<sub>6</sub>multi的表现与教师模型BERT<sub>12</sub>multi及其他法律判决预测深度学习模型进行了对比,并验证了知识蒸馏在体积及推理速度上的提升。

#### 3.1 实验数据

采用公开的CAIL2018-Small数据集,共包含19.6万条来自裁判文书网的法律文书,涉及202项刑事罪名以及183条刑法条例。故将罪名预测、法条推荐分类任务类别数目设置为202类和183类。

数据质量会显著影响模型的训练结果,针对CAIL2018-Small数据集进行数据清洗,去除数据中的无效及错误标记等异常数据,防止模型受到异常数据影响;再去除与判决结果无关的停用词;最后将文书中被告人的姓名,统一使用“被告人”代替。某条处理后的数据及标签如图5所示。

**预处理后数据:** 被告人在担任主任期间,受国土局委派到瑞城家居广场指挥部、中洲路指挥部、马坨弃土场等任工作人员,从事征地调查、征收、协助拆迁房屋测量、补偿等工作,利用职务便利,收受他人财物28900元,为他人谋取利益,其行为构成××罪。2012年,被告人受房地产开发有限公司总经理王某甲的委托与经营“大碗庄饭店”的王某就“大碗庄饭店”拆迁工作进行商谈,被告人虚构事实,隐瞒事实真相,骗取王某甲30万某,据为己有,其行为又构成××罪,应数罪并罚。  
**判决罪名:** 诈骗、受贿  
**涉及法条:** 385、383、266、386

图5 预处理后结数据及标签

#### 3.2 评价指标

参考2018年法研杯官方对于罪名预测和相关法条推荐子任务的评价标准,采用分类任务中的Micro-F1与Macro-F1作为评价指标,其计算方式为:

$$\text{Micro-F1} = \frac{2 \times \text{precision}_{\text{micro}} \times \text{recall}_{\text{micro}}}{\text{precision}_{\text{micro}} + \text{recall}_{\text{micro}}} \quad (16)$$

$$\text{Macro-F1} = \frac{\sum_{i=1}^N \text{f1}_i}{N} \quad (17)$$

采用Micro-F1与Macro-F1的均值F1-Score作为模型最终评价指标,即:

$$\text{F1-Score} = \frac{\text{Micro-F1} + \text{Macro-F1}}{2} \quad (18)$$

#### 3.3 模型参数

教师模型BERT<sub>12</sub>multi中Dropout比例设置为0.2,即在训练时使编码器输出的20%参数随机失活。知识蒸馏中的蒸馏温度参数 $T$ 设置为5,中间层损失权重 $\beta$ 设置为1000。教师模型BERT<sub>12</sub>multi和学生模型BERT<sub>6</sub>multi均采用Adam (adaptive moment estimation)<sup>[29]</sup>优化器进行训练,学习率设置为 $2 \times 10^{-5}$ ,文本长度限制为512字,batch size设置为128(受显存限制,采用梯度累积的方式,BERT<sub>12</sub>multi实际batch size为16,BERT<sub>6</sub>multi实际batch size为32,每累积至128更新1次梯度)。在CAIL2018-Small训练集上训练50轮,每轮在验证集上验证2次,取验证结果最佳的模型在测试集上进行测试。

参考文本分类技术现有研究,采用TF-IDF+SVM<sup>[4][8]</sup>机器学习模型,以及主流的基于词嵌入的深度学习文本分类模型TextCNN模型<sup>[30]</sup>、HAN模型<sup>[14]</sup>、DPCNN<sup>[15]</sup>模型作为对比模型,使用分词工具对案件文本分词后,固定句长为300词,采用word2vec<sup>[5]</sup>、GloVe<sup>[6]</sup>词嵌入算法,固定词嵌入维度为200,分别在全部数据集上训练词向量,并与在大型语料中通过Directional Skip-Gram (DSG)算法<sup>[31]</sup>获得的维度为200的开源词向量进行对比。依据各模型在不同词向量中的表现,最终在TextCNN模

型和HAN模型中使用word2vec词向量,DPCNN模型使用开源词向量. 针对罪名预测、法条推荐两项任务,各对比模型分别独立、联合建模,记录模型的最佳成绩.

同时将只使用分类损失和蒸馏损失的知识蒸馏策略<sup>[23]</sup>,以及耐心的知识蒸馏策略<sup>[24]</sup>与融入了教师模型评价的BERT<sub>6</sub>multi模型进行对比.除蒸馏策略外,教师模型及学生模型结构保持一致.

各对比模型batch size设置为128,与BERT<sub>12</sub>multi保持一致,其他参数如学习率、隐藏层维度、卷积核尺寸、蒸馏参数等在常用数值中采用网格搜索确定.

为进一步对比模型性能,结合近两年在法律判决预测领域的研究,及考虑到外国司法体系,裁判文书,判决习惯与国内差异,将BERT<sub>12</sub>multi、模型BERT<sub>6</sub>multi模型与现有中文法律判决模型MTL-Fusion<sup>[12]</sup>及HAC<sup>[13]</sup>进行对比.论文<sup>[12]</sup>中采用数据集与本文一致,论文<sup>[13]</sup>采用与本文相似的CAIL2018-Large数据集,故不对模型的参数进一步调整.

对于BERT<sub>12</sub>multi、BERT<sub>6</sub>multi及TextCNN、MTL-Fusion等对比模型,选取模型预测概率大于0.5的类别视为预测正样本,反之为负样本,即:

$$result(x_i \in c) = \begin{cases} 1 & \hat{y}_{ic} > 0.5 \\ 0 & \hat{y}_{ic} \leq 0.5 \end{cases} \quad (19)$$

### 3.4 实验结果

首先,将TextCNN等对比模型与采用BERT预训练模型作为编码器的BERT<sub>12</sub>multi进行对比,同时对联合训练的效果进行消融实验.如表2所示,其中multi表示采用联合建模方式对两项子任务进行多任务学习.各模型的Micro-F1指标较高,说明对各模型对于大部分样本可以做出准确的判断,Macro-F1相对较低,说明各模型误差主要源于部分样本数目较少的难分类类别;采用BERT预训练模型作为编码器在性能上,尤其是Macro-F1指标上显著优于机器学习模型或基于词嵌入的深度学习模型.其中BERT<sub>12</sub>multi性能最强,相比对比模型中非预训练模型在罪名预测任务及法条推荐任务中F1-Score的最佳成绩分别提升了0.042和0.040,说明预训练模型可以有效降低判决预测模型的整体误差;在BERT预训练模型及大部分对比模型中,采用联合建模方式对罪名预测及法条推荐任务中对于模型性能有一定提升效果,相比独立建模,各模型联合训练在两项子任务的F1-Score上均平均提高了约0.002.

其次,对采用分组Focal Loss的预测效果进行了实验,考虑对两项任务建立独立模型的计算资源消耗,实验中的对比模型均采用与BERT<sub>12</sub>multi一致的联合训练方式进行多任务学习.

表2 各模型评价指标对比1

任务名称	罪名预测任务			法条推荐任务		
模型名称	Micro-F1	Macro-F1	F1-Score	Micro-F1	Macro-F1	F1-Score
TF IDF+SVM for task1	0.792	0.63	0.711	-	-	-
TF IDF+SVM for task2	-	-	-	0.763	0.588	0.676
TextCNN for task1	0.868	0.739	0.803	-	-	-
TextCNN for task2	-	-	-	0.844	0.699	0.771
TextCNN multi	0.868	0.741	0.805	<b>0.844</b>	0.704	0.774
HAN for task1	0.856	<b>0.776</b>	0.816	-	-	-
HAN for task2	-	-	-	0.832	0.72	0.776
HAN multi	0.859	0.775	<b>0.817</b>	0.835	<b>0.723</b>	0.779
DPCNN for task1	0.866	0.753	0.809	-	-	-
DPCNN for task2	-	-	-	0.84	0.722	0.781
DPCNN multi	<b>0.868</b>	0.754	0.811	0.843	0.721	<b>0.782</b>
BERT for task1	0.886	0.825	0.856	-	-	-
BERT for task2	-	-	-	<b>0.868</b>	0.774	0.821
BERT <sub>12</sub> multi	<b>0.894</b>	<b>0.826</b>	<b>0.859</b>	0.867	<b>0.778</b>	<b>0.822</b>

表3 各模型评价指标对比2

任务名称		罪名预测任务			法条推荐任务		
模型名称	损失函数	Micro-F1	Macro-F1	F1-Score	Micro-F1	Macro-F1	F1-Score
TextCNN	Cross Entropy	0.868	0.741	0.805	0.844	0.704	0.774
HAN	Cross Entropy	0.859	0.775	0.817	0.835	0.723	0.779
DPCNN	Cross Entropy	0.868	0.754	0.811	0.843	0.721	0.782
BERT <sub>12</sub> multi	Cross Entropy	0.894	0.826	0.859	0.867	0.778	0.822
TextCNN	Focal Loss	0.869	0.752	0.811	0.846	0.727	0.787
HAN	Focal Loss	0.86	0.783	0.821	0.838	0.733	0.786
DPCNN	Focal Loss	0.872	0.761	0.817	0.845	0.732	0.789
BERT <sub>12</sub> multi	Focal Loss	<b>0.901</b>	<b>0.841</b>	<b>0.871</b>	<b>0.873</b>	<b>0.792</b>	<b>0.832</b>

表4 各模型评价指标对比3

任务名称 模型名称	罪名预测任务			法条推荐任务		
	Micro-F1	Macro-F1	F1-Score	Micro-F1	Macro-F1	F1-Score
MTL-Fusion	0.881	0.814	0.847	0.852	0.755	0.804
HAC	0.876	0.803	0.84	0.856	0.763	0.81
BERT <sub>12</sub> multi	0.901	0.841	0.871	0.873	0.792	0.832
BERT <sub>6</sub> multi(KD)	0.878	0.804	0.841	0.843	0.741	0.792
BERT <sub>6</sub> multi(PKD)	0.894	0.834	0.864	0.868	0.785	0.827
BERT <sub>6</sub> multi(Ours)	<b>0.898</b>	<b>0.836</b>	<b>0.867</b>	<b>0.871</b>	<b>0.788</b>	<b>0.83</b>

如表3所示,采用分组Focal Loss的各模型在Macro-F1指标上均有提高,其中BERT<sub>12</sub>multi模型在两项任务中的Macro-F1指标上提高了0.012和0.018,表明分组Focal Loss可以在不显著增加超参数数量的情况下,有效提高模型在类别样本不平衡数据集集中的性能,降低对于样本数量较少类别的判断误差。

各蒸馏策略训练的BERT<sub>6</sub>multi及现有法律判决预测模型MTL-Fusion、HAC在CAIL2018-Small中表现如表4所示。其中,KD表示采用固定参数的交叉熵损失和蒸馏损失的蒸馏策略,PKD表示在KD基础上,引入教师模型与学生模型中间层分布损失的耐心的知识蒸馏策略;Ours代表本文采用的融入教师模型评价的耐心知识蒸馏策略。

结合表2、3、4,在CAIL2018-Small数据集上,采用融入教师模型评价的耐心知识蒸馏策略获得的BERT<sub>6</sub>multi与BERT<sub>12</sub>multi性能接近,优于TF-IDF+SVM、TextCNN、MTL-Fusion等对比模型以及其他蒸馏策略获得的BERT<sub>6</sub>multi模型。

为对比学生模型与教师模型在推理速度上的差异,将BERT<sub>6</sub>multi与BERT<sub>12</sub>multi在GTX 1660Ti(6GB)、pytorch1.1、python3.7的环境下,对相同的1000条案件数据进行逐条推理获得平均单条数据推理时间,并统一batch size为64对测试集32421条数据进行验证,获得总验证时间。BERT<sub>6</sub>multi与BERT<sub>12</sub>multi参数量及推理速度对比如表5所示。

表5 BERT<sub>12</sub>multi与BERT<sub>6</sub>multi参数量及推理速度对比

	BERT <sub>12</sub> multi	BERT <sub>6</sub> multi	倍率
模型编码器层数	12	6	0.5x
模型编码器参数	102M	60M	0.59x
单条推理时间	0.057s	0.029s	0.51x
总验证时间	22min50s	10min55s	0.48x

从表5中可以看出,经知识蒸馏后的BERT<sub>6</sub>multi体积约为BERT<sub>12</sub>multi一半,模型推理速度提升约一倍。

## 4 结论

本文针对法律判决预测任务中罪名预测及法

条推荐两项子任务,基于无监督预训练模型BERT对两项子任务联合建模,提出了多任务判决预测模型BERT<sub>12</sub>multi。针对法律判决预测任务中类别正负样本不均衡、类别呈现长尾分布的特征,对样本数接近的类别进行分组,采用了一种分组Focal Loss策略,平衡样本不均衡的类别。在实验中,采用分组Focal Loss策略的BERT<sub>12</sub>multi性能优秀,但由于预训练模型参数量较大,使得计算资源占用较高,运行效率较低。针对法律判决预测任务类别数目众多、且存在易混淆类别的特点,提出了一种改进的知识蒸馏策略,将教师模型在训练数据中的表现作为学生模型从教师模型及标签数据中学习的依据,采用动态权重权衡标签损失与蒸馏损失,获得学生模型BERT<sub>6</sub>multi。

在CAIL2018数据集上的实验证明,采用预训练模型可以有效提高深度学习在法律判决预测任务中的表现;分组Focal Loss策略可以进一步提高BERT<sub>12</sub>multi在不均衡数据集上的性能,而不显著增加超参数;将训练好的预训练模型作为教师模型,采用融入教师模型评价的耐心知识蒸馏策略,可以有效缩小预训练模型的体积,加快预训练模型推理速度,进一步提高预训练模型的实用性。

## 参考文献(References)

- [1] Liu Y H, Chen Y L. A two-phase sentiment analysis approach for judgement prediction[J]. Journal of Information Science, 2018, 44(5): 594-607.
- [2] Lin W C, Kuo T T, Chang T J, et al. Exploiting machine learning models for chinese legal documents labeling, case classification, and sentencing prediction[J]. Proceedings of ROCLING, 2012: 140.
- [3] Sulea O M, Zampieri M, Malmasi S, et al. Exploring the use of text classification in the legal domain[J]. International Conference on Foundations of Intelligent Systems, 2006: 681-690.
- [4] Suykens J A K, Vandewalle J. Least squares support vector machine classifiers[J]. Neural processing letters, 1999, 9(3): 293-300.
- [5] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information

- processing systems. 2013: 3111-3119.
- [6] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.
- [7] Joulin A, Grave É, Bojanowski P, et al. Bag of Tricks for Efficient Text Classification[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017: 427-431.
- [8] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. *Information processing & management*, 1988, 24(5): 513-523.
- [9] Mihalcea R, Tarau P. TextRank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 404-411.
- [10] Kampas D, Chalkidis I. Deep learning in law: early adaptation and legal word embeddings trained on large corpora[J]. *Artificial Intelligence and Law*, 2019, 27(2): 171-198.
- [11] Yang W, Jia W, Zhou X, et al. Legal judgment prediction via multi-perspective bi-feedback network[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. AAAI Press, 2019: 4085-4091.
- [12] 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[J].*清华大学学报(自然科学版)*,2019,59(07):497-504.
- [13] 王文广, 陈运文, 蔡华, 等. 基于混合深度神经网络模型的司法文书智能化处理[J]. *清华大学学报(自然科学版)*, 2019, 59(7): 505-511.
- [14] Yang Z, Yang D, Dyer C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016: 1480-1489.
- [15] Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 562-570.
- [16] Kenton J D M W C, Toutanova L K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of NAACL-HLT. 2019: 4171-4186.
- [17] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]//Proceedings of NAACL-HLT. 2018: 2227-2237.
- [18] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[C]//Advances in neural information processing systems. 2019: 5754-5764.
- [19] Chalkidis I, Androutsopoulos I, Aletras N. Neural Legal Judgment Prediction in English[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4317-4323.
- [20] Cho K, van Merriënboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1724-1734.
- [21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
- [22] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network[J]. *stat*, 2015, 1050: 9.
- [23] Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [J/OL]. (2019-10-02). <https://arxiv.org/abs/1910.01108>
- [24] Sun S, Cheng Y, Gan Z, et al. Patient Knowledge Distillation for BERT Model Compression[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 4314-4323.
- [25] Lin T Y, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, PP(99):2999-3007.
- [26] Xiao C, Zhong H, Guo Z, et al. Cai2018: A large-scale legal dataset for judgment prediction [J/OL]. (2018-07-04). <https://arxiv.org/abs/1807.02478>.
- [27] Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for chinese bert [J/OL]. (2019-10-29). <https://arxiv.org/abs/1906.08101>
- [28] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. *The journal of machine learning research*, 2014, 15(1): 1929-1958.
- [29] Kingma D P, Ba J. Adam: A method for stochastic optimization[C]// Proceedings of International Conference on Learning Representations (ICLR) 2015, 2015:1-15.
- [30] Kim Y. Convolutional Neural Networks for Sentence Classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1746-1751.
- [31] Song Y, Shi S, Li J, et al. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 175-18

## 作者简介

潘瑞东(1996—), 男, 硕士生, 从事自然语言处理等研究, E-mail: 1208668157@qq.com;

孔维健(1983—), 男, 讲师, 博士, 从事机器学习算法及应用研究, E-mail: kongweijian@dhu.edu.cn.

齐洁(1978—), 女, 教授, 博士, 从事多智能体系统、复杂系统建模与控制的研究, E-mail: jieqi@dhu.edu.cn.