第49卷第1期
2021年1月

华南理工大学学报(自然科学版)
Journal of South China University of Technology
(Natural Science Edition)

Vol. 49 No. 1
January 2021

doi: 10.12141/j.issn.1000-565X.200513

基于半监督学习的涉及未成年人案件文书识别方法

杨圣豪 吴玥悦 毛佳昕 刘奕群† 张敏 马少平

(清华大学 计算机科学与技术系//北京信息科学与技术国家研究中心，北京 100084)

摘 要: 案件文书作为司法信息公开的重要内容，需要在审判之后向公众公开，某些涉及未成年人的案件文书极有可能会造成未成年人的个人隐私信息泄露。为了能从大量案件文书中准确地识别出涉及未成年人信息的文书，进而有针对性地对其进行隐私保护处理。同时，为解决现实数据集因有标注样本缺乏而难以进行有效的有监督学习的问题，文中提出了基于半监督学习的涉及未成年人案件文书识别方法。首先，对案件文书语料文本进行预处理后分别使用 Word2Vec 和 BERT-wwm-ext 对文本进行特征提取，将长语料文本转换为可作为分类模型输入的数据格式；接着，采用 PU 学习方法训练分类模型，在正例样本极少的情况下借助大量未标注样本构建有效的分类器；然后，在分类模型预测结果的基础上，使用主动学习方法获取关键词并对模型预测结果进行筛选处理，以进一步提升预测效果。在基于现实场景比例构建的测试集上，文中提出的案件文书识别方法取得了98.67%的召回率和81.02%的准确率。

关键词: 文本分类；文本特征提取；深度学习；半监督学习

中图分类号: TP391 **文章编号:** 1000-565X(2020)01-0029-10

近年来，随着司法工作的逐步信息化和透明化，作为司法信息公开的一部分，案件文书会在中国裁判文书网等一些政府网站上进行公开，伴随而来的是案件当事人的个人隐私信息泄露的问题，特别是未成年人隐私信息的泄露[1]，会对未成年人的生活带来不良影响。因此，有必要从大量案件文书语料中准确识别出涉及未成年人信息的语料，帮助相关工作人员在公开案件文书前进行一些隐私保护处理，从而达到保护未成年人个人隐私信息的目的。

从大量案件文书语料中准确识别出未成年人信息的语料，属于文本分类问题。文本分类是依据文本内容为其分配标签的过程。例如，新产生的文章可以通过文本分类来分配主题以便进行归档，聊天对话可以按语言风格进行组织，商品评论可以按情感进行组织等。文本分类任务在各个领域中有着普遍的运用，例如情感分析、主题标签、垃圾邮件检测和意图检测等信息检索方面的任务。随着科学技术的不断进步，文本分类在医学、心理学、法学、工程学等许多方面也有着广泛的应用[2]。

与传统的文本分类研究中文本的序列长度相对较短不同，案件文书语料的内容往往达到上千字级别，而且语言描述比较冗杂，使用传统的特征提取和文本分类方法无法取得令人满意的效果。另外，由于实际案件文书缺乏足够的有标注数据，在现有

收稿日期: 2020-08-25

基金项目: 国家重点研发计划项目(2018YFC0831700)；国家自然科学基金资助项目(61732008, 61532011)

Foundation items: Supported by the National Key R&D Program of China (2018YFC0831700) and the National Natural Science Foundation of China (61732008, 61532011)

作者简介: 杨圣豪(1998-)，男，主要从事信息检索研究。E-mail: yangsh824@gmail.com

† **通信作者:** 刘奕群(1981-)，男，博士，教授，主要从事网络信息检索、网络用户行为分析研究。E-mail: yiqunliu@tsinghua.edu.cn

的数据集中只有少量正例样本而没有负例样本。与未成年人相关的案件文书在现实场景下存在的比例极小，相对而言，所有与未成年人不相关的案件文书都可以看作负例，但这些负例文书太多样化并且难以收集，这就导致了数据集中正负例样本极度不均衡，使得分类模型难以进行有效的监督训练。目前大多数文本分类方法基于大量的已标注数据集，为此本文提出了基于正例、无标记样本(PU)学习和主动学习两种半监督学习方法的未成年人案件文书识别方法。首先将所有案件文书语料按照一定比例划分为训练集和测试集，对语料文本进行预处理后分别使用 Word2Vec 和 BERT-wwm-ext 对文本进行特征提取，将长语料文本转换为可作为分类模型输入的数据格式；接着采用 PU 学习方法对分类模型进行训练，在正例样本极少的情况下借助大量未标注样本构建有效的分类器；然后在分类模型预测结果的基础上，使用主动学习方法获取关键词并对模型预测结果进行筛选处理，以进一步提升预测效果；最后在面向现实场景构建的测试数据集上进行实验，分析本文识别方法的性能。

1 相关工作

1.1 文本特征提取

文本作为一种半结构化或非结构化的信息，为适应计算机处理，必须转换为可识别的格式。文本特征提取可以实现将文本转换为计算机可以处理的数值模式[3]。向量空间模型[4]作为一种经典的文本特征提取方法，自1975年被提出后一直有着广泛的应用。作为一种向量空间模型的词袋模型，其主要思想是将一段文本用其所包含的词作为特征来表示，这种表示方式只考虑词频，并不能很好地反映词的重要性[5]。TF-IDF 模型提供了一种反映词重要性的方法，对于某个词，当它在当前语料中出现的次数较多，而在其他语料中出现的次数较少时，其 TF-IDF 值会越大，也就表示其重要性越高[6]。Dumais[7]提出的潜在语义分析(LSA)，通过分析自然文本的代表性语料库来对单词和段落的含义进行计算机建模和仿真，其通过奇异值矩阵分解的方式解决了一义多词的问题。Blei 等[8]提出的潜在的狄利克雷分配(LDA)是一个生成概率模型，它收集文本语料库的离散数据并分析文本，以获得它属于某个主题的概率，并使用这些主题概率进行聚类和分类。

向量空间模型本质上是将文本用一组特征值表示，其不足之处是忽略了文本中每个词的内在联系，缺乏前后文语义。词向量模型则是将文本中的每个词表示为向量，one-hot 表示是一种易理解的词向量模型，其基本思想是将某个词表示成这样一种向量：向量的维度为词的总数，其中只有对应该词在所有词中位置的元素为1，其他都为0。这种表示方法会使词向量的维度很大。与 one-hot 表示相对的分布式表示方法，会将每个词表示成特定维度的稠密连续向量，其思想是假设两个词的上下文相同，那么这两个词的语义应该是一样的，因此使用分布式表示方法获得的词向量存在着语义联系，如具有相近含义的词会具有类似的形式，这是 one-hot 表示所做不到的。Mikolov 等[9-10]提出的 Word2Vec 基于两种语言模型(CBOW 和 Skip-Gram)，通过将词或者词的上下文作为输入，去预测词的上下文或词，使用训练过程中模型的参数得到词的向量表示。Pennington 等[11]提出的 Glove 模型是一种基于全局向量的词向量模型，融合了全局矩阵分解和局部上下文窗口两种方法的优点。Word2Vec 和 Glove 都是基于固定表征的词向量方法，难以解决一词多义问题，而一些基于语言模型的动态表征方法(如 ELMo[12]、GPT[13]、BERT[14]等)生成的词向量可以学习到词在不同语境中的含义变化。

1.2 文本分类

早期的文本分类模型一般会使用 SVM[15]、LDA[16]等一些传统的机器学习算法。近年来随着深度学习算法的发展，出现了大量以卷积神经网络(CNN)或循环神经网络(RNN)为基础的文本分类方法。FastText[17]的输入为经过词向量模型得到的词向量，并将所有词向量进行平均，最后经由 Softmax 层得到结果，整体架构较为简单。Text-CNN[18]同样使用词向量作为输入，经由卷积层和池化层后得到特征向量，最后经由一个全连接层和 Softmax 层输出预测结果。RNN 因其递归结构非常适合处理不同长度的文本而在文本分类中被广泛应用。长短期记忆网络(LSTM)[19]相对于传统 RNN，增加了一个新的传递状态，并且加入了一些选择门控对传递过程中的信息进行取舍。TextRCNN[20]综合考虑了 CNN 和 RNN 结构的优缺点，先使用双向 RNN 结构作为 CNN 中的卷积层来获取上下文信息，并与词本身结合作为词的表示特征，然后将获取到的特征输入到池化层，之后的结构与 TextCNN 相似。

近年来，BERT 和 XLNet[21]等预训练语言模型陆

续出现。BERT 是一种双向 Transformer 结构的编码器，基于 Masked 语言模型和下一个句子预测（NSP）进行预训练。XLNet 在预训练中引入了排序语言模型（PLM），并采用 Transformer XL 使其对长文本序列的处理效果更好。BERT 和 XLNet 等预训练模型在机器阅读理解、句子分类等任务中都取得了出色的表现，也是时下研究的热点，但预训练模型在文本长度上的限制以及在较长文本上的训练耗时较长，使得它们在长文本分类任务中并没有受到青睐。

1.3　半监督学习方法

大多数文本分类任务会基于大量的有标注数据进行有监督的学习，但在现实情况下，往往很难收集到标准化的成对数据集，在无法进行有效的有监督学习方法的情况下，可以使用半监督学习方法。

1.3.1　PU 学习

PU 学习是一种使用正例样本和未标记样本进行训练的半监督学习方法[22]。PU 学习有 3 种常见的实现方法：标准法[23]、PU Bagging[24] 和两步法[25]，其中 PU Bagging 方法使用一种来自于统计学中 Bootstrap 的思想，具体步骤如下：

（1）随机抽取等同于正例样本个数的未标注样本，并将它们看做负例样本，构建一个 Bootstrap 训练集；

（2）使用构建好的 Bootstrap 训练集训练分类器；

（3）使用训练好的分类器预测那些没有被抽取出来的未标注样本，并记录其被预测成正例样本的概率；

（4）重复 T 次以上步骤，最后求所有未标注样本被预测成正例的概率的平均值。

文献［24］使用以下伪代码来描述整个 PU Bagging 的过程：

{ Input: P, U, K = size of bootstrap samples, T = number of bootstraps
Output: a scores s: $U \rightarrow \mathbf{R}$
Initialize $\forall x \in U, n(x) \leftarrow 0, f(x) \leftarrow 0$
for $t = 1$ to T do
Draw a bootstrap sample U_t of size K in U.
Train a classifier f_t to discriminate P gainst U_t.
for any $x \in U \setminus U_t$ update
$f(x) \leftarrow f(x) + f_t(x)$
$n(x) \leftarrow n(x) + 1$
end for
Return $s(x) = f(x) / n(x)$ for $x \in U$}

1.3.2　主动学习

主动学习作为一种半监督学习方法，常被应用到样本不平衡的场景中，在这种情况下往往需要进行一些样本标注。主动学习最大的特点是其算法本身可以主动地提出对样本进行标注，而不是由人工随机抽取样本，从而可以用尽可能低的人工标注成本来获得更好的模型性能[26]。

主动学习模型可以表示为

$$A = (C, Q, S, L, U) \tag{7}$$

式中：C 为分类模型；Q 为从未标注样本集 U 中查询信息的查询者；S 为标注者，可以标注 U 中的样本；L 为有标注的样本集。使用图 3 描述各个角色之间的联系。

如图 1 示，主动学习的基本思路可以描述如下：

（1）从未标注样本集 U 中随机抽样，S 对其进行标注，加入到有标注样本集 L 中；

（2）使用有标注样本集 L 训练分类器 C，得到预测值；

（3）Q 从分类结果中选择（查询）一些样本，S 对其进行标注；

（4）重复步骤（2）、（3），直到没有更多样本或者当前分类器对样本的预测正确或者选择（查询）的样本也无法人工进行分类时结束。

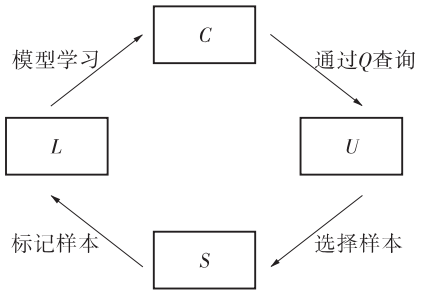


图 1　主动学习模型

Fig. 1　Active learning model

在主动学习算法中，Q 如何选择有价值的样本是一个值得探究的问题，也是主动学习的一个核心问题。文献［27］中介绍了 3 种样本选择策略：①RS 法，即从 U 中随机选择样本，是一种被动选择策略，是为了与主动选择策略作对比而提出的；②LC 法，即选择那些在分类结果中后验概率不太可信的样本——难以判别类别的样本；③BT 法，即选择那些在分类结果中属于某两个类的后验概率差异最小的样本——难以区分是类 1 还是类 2 的样本，其着重于某两类之间的边界区域，目的是丰富训练集的多样性。

2　基于 PU 学习的案件文书识别

本文提出的基于 PU 学习的案件文书识别框架如图 2 所示。

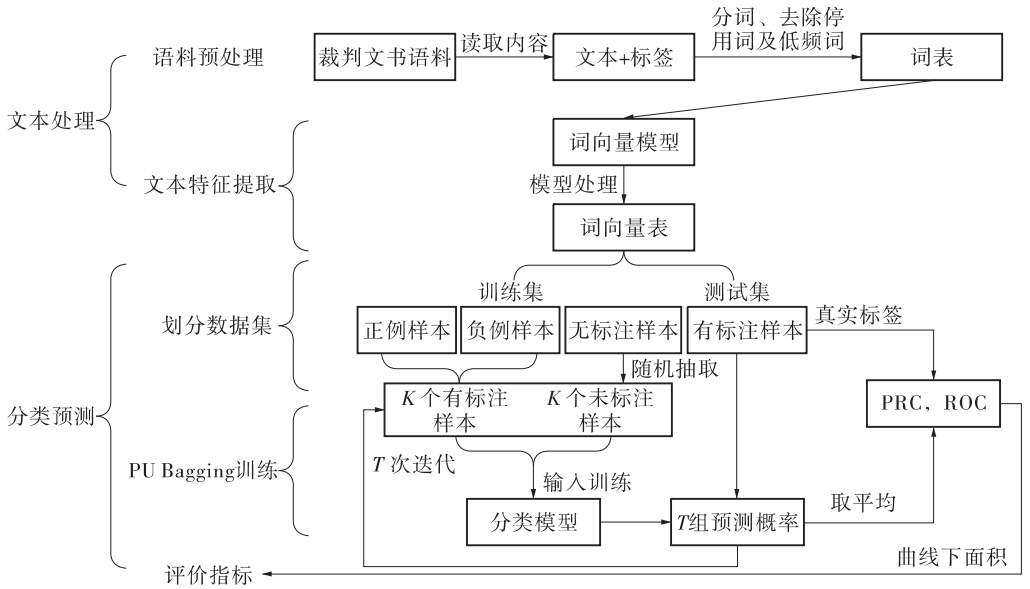


图2 基于PU学习的案件文书识别框架

Fig. 2 Framework of recognition of case documents based on PU learning

2.1 数据集

将所有案件文书语料按照一定比例划分为训练集和测试集，训练集中包括3种样本（正例样本、负例样本和未标注样本），测试集中则全部都是有标注样本。

本文期望构建一个符合真实应用场景的测试集，即构建的测试集中正负例样本的比例应大致符合真实场景下正负例样本的比例。由于现阶段并没有涉及未成年人的案件文书在海量文书中所占比例的研究，只能根据以往的标注经验对正负例样本占比进行估计。在之前的一次标注经历中，从1 000个未标注样本中标注得到81个正例样本。另外，根据经验，在现实情况中，涉及未成年人的案件文书确实较少。综合考虑后，本文最终使用的测试集构成为225个正例样本和1 535个负例样本，共1 760个样本；训练集共有7 867个样本，其中正例样本313个，负例样本54个，未标注样本7 500个。

2.2 文本预处理和特征提取

2.2.1 使用Word2Vec进行文本特征提取

对于原始案件文书语料，首先读取其中的文本内容，并进行去除空格、换行符等清理操作；接着使用分词工具对文本内容进行分词、去除停用词和低频词操作，这样每篇语料都被表示成一组词表；然后使用所有语料的词表训练Word2Vec模型，再使用训练得到的模型将所有语料的词表中的每个词都转换成词向量表示，这样就将所有案件文书语料表示成一个个词向量表。

2.2.2 使用BERT进行文本特征提取

将语料文本直接作为BERT预训练模型的输入，最终将每个语料文本都转化为词向量的集合，本文使用的BERT预训练模型为BERT-wwm-ext，其得到的词向量维度为768。

2.3 分类模型的训练和预测

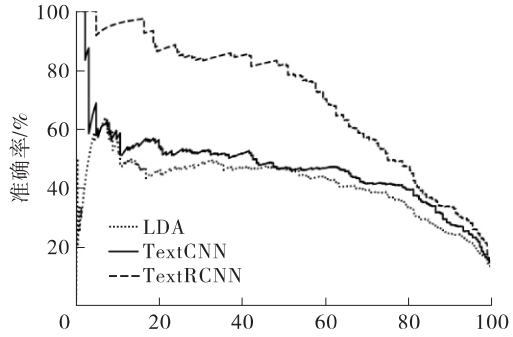
本文使用两个深度学习模型（TextCNN、Text-RCNN）和一个机器学习模型LDA作为分类模型，并采用PU学习方法对分类模型进行训练和预测，具体过程如下：

（1）将训练集中的367个已标注样本和从7 500个未标注样本随机抽取的等量未标注样本输入到分类模型，进行分类模型的训练；

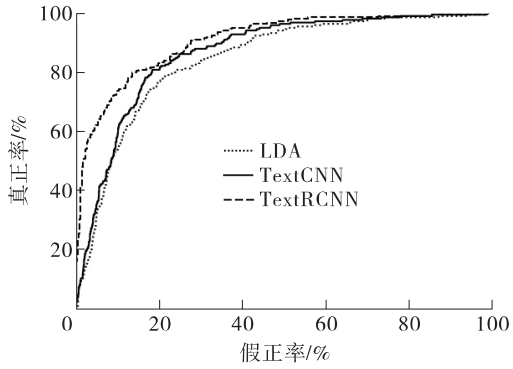
（2）使用训练得到的分类模型对测试集中各个样本进行预测，记录测试集中每个样本被预测为正例样本的概率；

（3）返回步骤（2）反复进行T次，最终获得T组预测概率，计算这T组预测概率的平均值。

将测试集中各个样本的平均预测概率与其真实标签作对比，计算并绘制出准确率(P)－召回率(R)曲线（PRC）和受试者工作特征（ROC）曲线，结果如图3所示，P-R和ROC曲线的线下面积如表1所示。

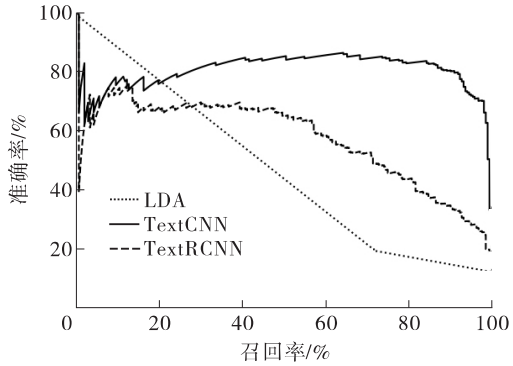


P-R 曲线

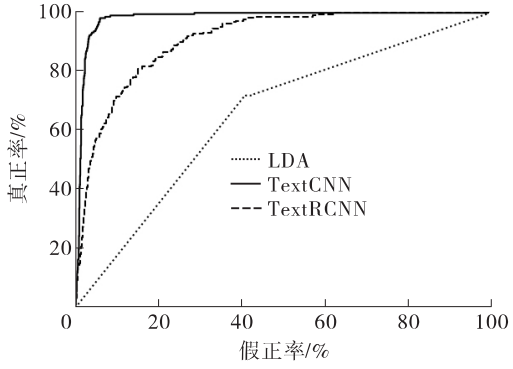


ROC 曲线

(a)　使用 Word2Vec 进行文本特征提取



P-R 曲线



ROC 曲线

(b)　使用 BERT 进行文本特征提取

图 3　3 种分类模型的预测结果

Fig. 3　Prediction results of 3 classification models

表 1　3 种分类模型预测结果的曲线下面积

Table 1　Area under the curve of prediction results of 3 classification models

分类模型	文本特征提取方法	曲线下面积	
		P-R 曲线	ROC 曲线
LDA	Word2Vec	0. 41	0. 85
	BERT	0. 48	0. 65
TextCNN	Word2Vec	0. 47	0. 87
	BERT	0. 81	0. 98
TextRCNN	Word2Vec	0. 70	0. 91
	BERT	0. 58	0. 91

由图 3 和表 1 可以看出，TextCNN 在使用 BERT 词向量作为输入的情况下，分类效果最佳，可以达到 0. 81 的 *P-R* 曲线下面积和 0. 98 的 ROC 曲线下面积。

3　基于主动学习的关键词后处理方法

为在 TextCNN 得到的分类结果的基础上进一步提升分类效果，本文考虑先归纳出与未成年人案件文书相关的关键词，并使用关键词对分类结果中被预测为负例的样本进行关键词筛选，以进一步召回更多的正例，从而提升分类效果。

一般来说，为归纳关键词需要进行大量的标注并根据经验进行总结，这种方式一方面会耗费巨大的人力和时间，另一方面根据经验总结得到的关键词是否准确、有效，也有待商榷。为此，本文提出使用主动学习方法，借助其算法本身会不断提出新的标注需求进而逐步提升模型性能的特点，通过观察主动学习过程中分类模型对各个词的权值系数变化情况，分析出最有助于提升模型性能的词，而这些词就可以用来做关键词筛选的关键词。

整个关键词后处理方法基于 Libact[28] 库实现，同时为便于观察词的权值系数变化情况，本文使用基于词袋模型的特征提取方法，并使用线性分类器进行分类，具体步骤如下：

(1)　在有标注训练集上对线性分类器进行训练；

(2)　使用训练后的线性分类器对测试集进行预测，记录分类结果的 *P-R* 曲线和 ROC 曲线下面积，并记录分类器对于各个词的权值系数；

(3)　使用训练后的线性分类器对未标注数据集进行预测，根据样本选择策略选取待标注样本；

(4)　对待标注样本进行标注后加入到有标注训练集中，重新对线性分类器进行训练；

(5) 重复步骤(2)、(3)、(4),直到达到可以结束的条件。

3.1 关键词分析过程

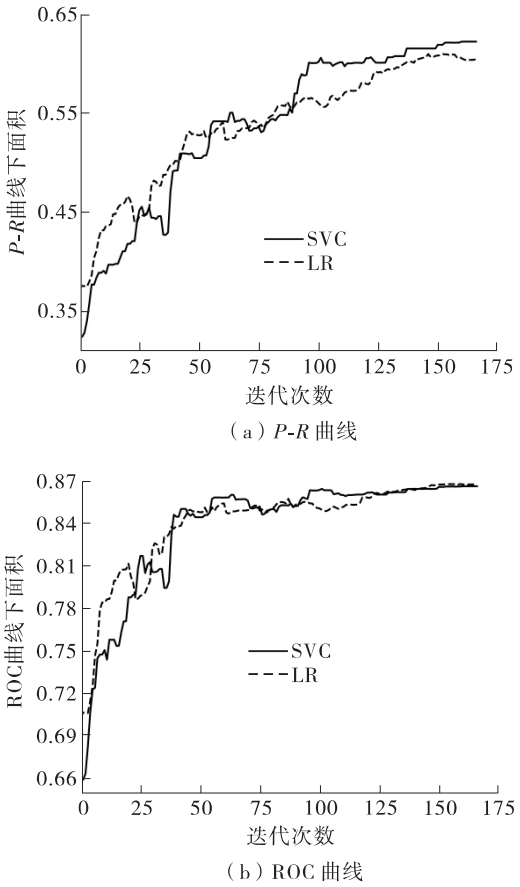
首先基于词袋模型进行特征提取,步骤如下:

(1) 将所有语料转换成使用其中所有词的词频进行编码后得到的向量,此时向量维度等于词的个数,超过1万;

(2) 使用TF-IDF模型将所有语料向量中的词频值转换成tf-idf权值,并通过只保留词频排在前1500的词来去除大量低频词,从而将所有语料向量的维度降低到1500。

随后选用两种线性分类模型:逻辑回归(LR)和支持向量分类器(SVC,线性核),将基于词袋模型获得的特征作为输入,使用Libact库进行主动学习训练,使用Libact中的ALBL样本选取策略。初始的有标注训练集中样本的个数为200,在迭代过程中逐一标注未标注样本,每次标注过后重新训练分类模型并对测试集进行预测,记录预测结果的*P-R*曲线和ROC曲线的曲线下面积。经过167次标注后,*P-R*曲线和ROC曲线的曲线下面积的变化趋势如图4所示。由图中可知,随着迭代的进行,两个模型的*P-R*曲线下面积和ROC曲线下面积都有着明显的增长趋势,并且在曲线最大值处,两个模型均能取得0.6左右的*P-R*曲线下面积和0.85左右的ROC曲线下面积,总体而言,SVC的分类效果更好。

在词袋模型中共有1500个词,在主动学习过程中,每次模型训练过后会得到一组一一对应这些词的权值系数,对整个迭代过程中的所有权值系数进行记录,取最后一轮迭代后权值系数最大的10个词,将他们的权值系数随迭代的变化情况绘制成折线图。图5是使用SVC模型进行主动学习后得到的折线图(图例中的数字代表该词最终的权值系数)。如图5所示,在权值系数排名前10的词中,“未成年人”“周岁”“未成年”“学生”的权值系数随着迭代的进行一直呈上升趋势;“十八周岁”“孩子”既有上升趋势也有下降趋势,但最终的权值系数排名较高;“抚养费”“小孩”作为和未成年人有一定关系的词,其最初的权值系数很高,但随着迭代的进行,其权值系数有明显的下降;“法定代理”“村民”作为和未成年人没有太大关系的词,其权值系数在迭代过程中并没有很明显的变化趋势,整体都在0.6左右。最终选择“未成年人”“周岁”“未成年”“学生”“十八周岁”“孩子”“抚养费”“小孩”这8个词来进行下一步的关键词筛选后处理。



(a) *P-R*曲线

(b) ROC曲线

图4 两种线性分类模型的曲线下面积随迭代次数的变化

Fig. 4 Variation of the area under the curve of two linear classification models with the number of iterations

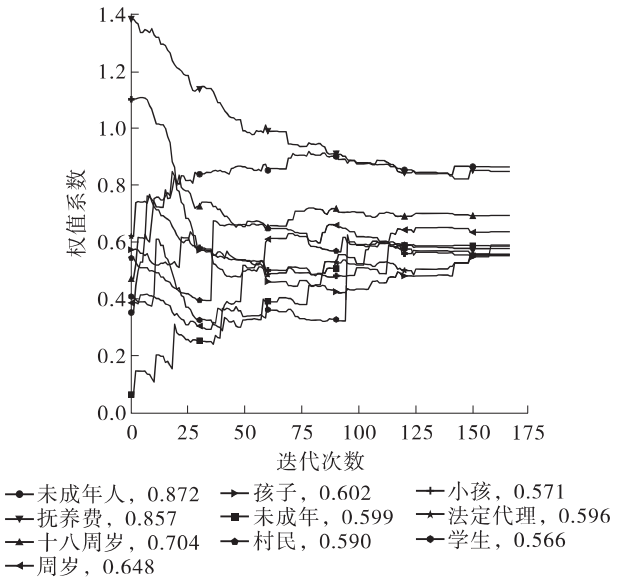


图5 主动学习过程中词的权值系数的变化情况

Fig. 5 Variation of weight coefficients of words during active learning

3.2　关键词后处理

首先使用关键词筛选方法对测试集进行测试，以检验本文通过主动学习分析得到的关键词在关键词筛选中的效果，具体方法为：遍历所有测试集中的语料内容，将出现 3.1 中 8 个关键词的样本判定为正例，否则判定为负例，最后将该分类结果与真实标签对比，得到的召回率和准确率分别为 93.33% 和 87.40%，说明分类效果较好。

随后在 2.3 节 TextCNN 分类结果的基础上(此时测试集中的每个样本都有一个分数，这个分数是该样本可能为正例的概率)，使用关键词筛选进行后处理，以提升分类效果。根据 TextCNN 分类结果的 *P*-*R* 曲线（如图 3（b）所示)，曲线上的每一个点均对应一组准确率、召回率和阈值，选取准确率和召回率之和最大时的阈值（其值为 99.53%)，根据该阈值将测试集中所有样本的分数排序，分数大于该阈值的样本归为正例，小于该阈值的样本归为负例，从而得到一组预测结果，将该预测结果与真实标签对比，得到的召回率和准确率分别为 92.00% 和 80.23%。

随后对上述被预测为负例的样本使用关键词筛选方法，将 3.1 节中分析得到的 8 个词作为关键词，对于出现了关键词的样本，将原本因分数小于阈值而被归为负例的样本改判为正例，没有出现关键词的样本则保持为负例，从而得到一组新的预测结果，将新预测结果与真实标签对比，得到的召回率和准确率分别为 98.67% 和 81.02%。

由表 2 可以看出：使用 PU 学习、主动学习和关键词筛选后，TextCNN 的分类结果的召回率(98.67%）和准确率（81.02%)，相较于只使用 PU 学习进行 TextCNN 训练时分别提升了 6.67% 和 0.79%，相较于只使用主动学习分析得到的关键词进行关键词筛选时，召回率提高了 5.34% 但准确率降低了 6.38%；只进行传统的有监督学习方法时，TextCNN 的召回率和准确率都是最差的。

表 2　是否使用 PU 学习和主动学习时的实验结果对比

Table 2　Comparison of experimental results whether using PU learning and active learning or not　%

方法	召回率	准确率
TextCNN	41.33	11.40
PU 学习 + TextCNN	92.00	80.23
主动学习 + 关键词筛选	93.33	87.40
PU 学习 + TextCNN + 主动学习 + 关键词筛选	98.67	81.02

值得注意的是，虽然单纯使用关键词筛选获得的召回率和准确率综合来看是最高的，但本文的研究目标是尽可能多地识别到正例样本，尽可能提高召回率，同时确保准确率不会太低。在 TextCNN 模型分类结果的基础上加入关键词后处理的方法，可以有效地提升召回率，同时也有不错的准确率，因此就本文研究任务而言，可以认为这是最优的结果。

3.3　人机交互系统

为了使主动学习中的标注过程更加友好，本文

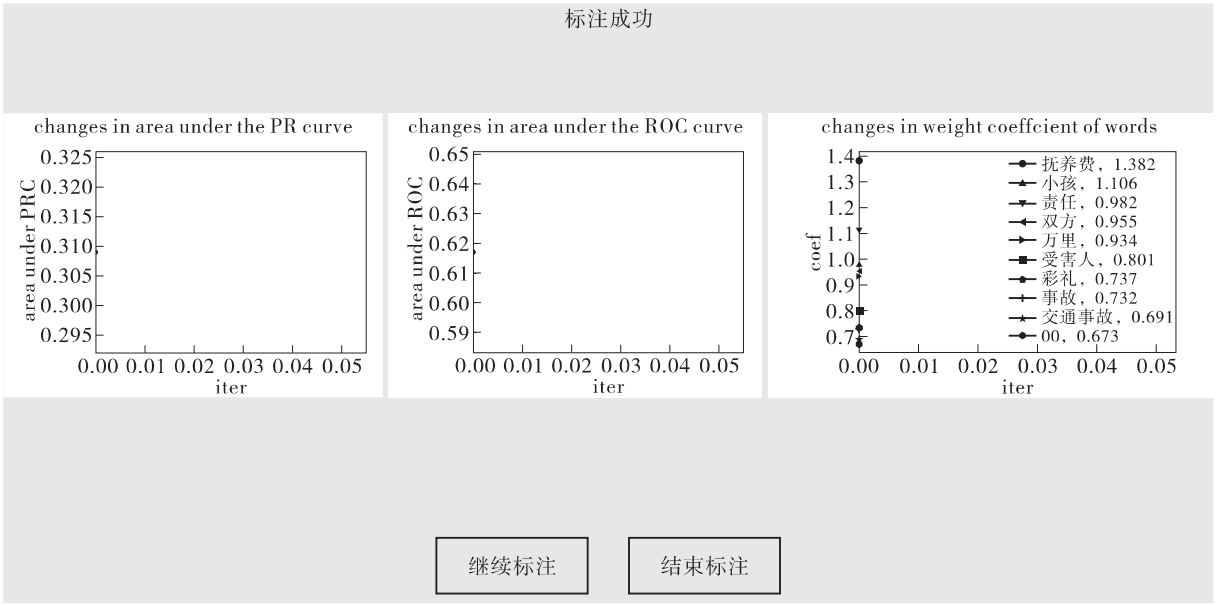


图 6　第一次标注后的测试结果

Fig. 6　Test results of the first labeling

基于 Flask 框架设计了一个人机交互系统。该系统的主要功能为：对主动学习算法选取的待标注样本进行标注，并将新标注的样本加入到训练集中，重新训练模型并重新在测试集上进行测试，将标注前后模型在测试集上的分类效果进行对比展示。

首先根据主动学习算法中的样本选择策略给出待标注的样本，然后在系统中进行标注，标注成功后会看到使用新标注训练集训练后的分类器在测试集上的分类结果以及分类器对于各个词的权值系数情况，结果如图 6 所示。进行第二次标注后，结果会更新，如图 7 所示。进行多次标注后，可以看到折线图上有明显的变化趋势，如图 8 所示。再次打开系统时，就会显示目前最新的测试结果，如图 9 所示。

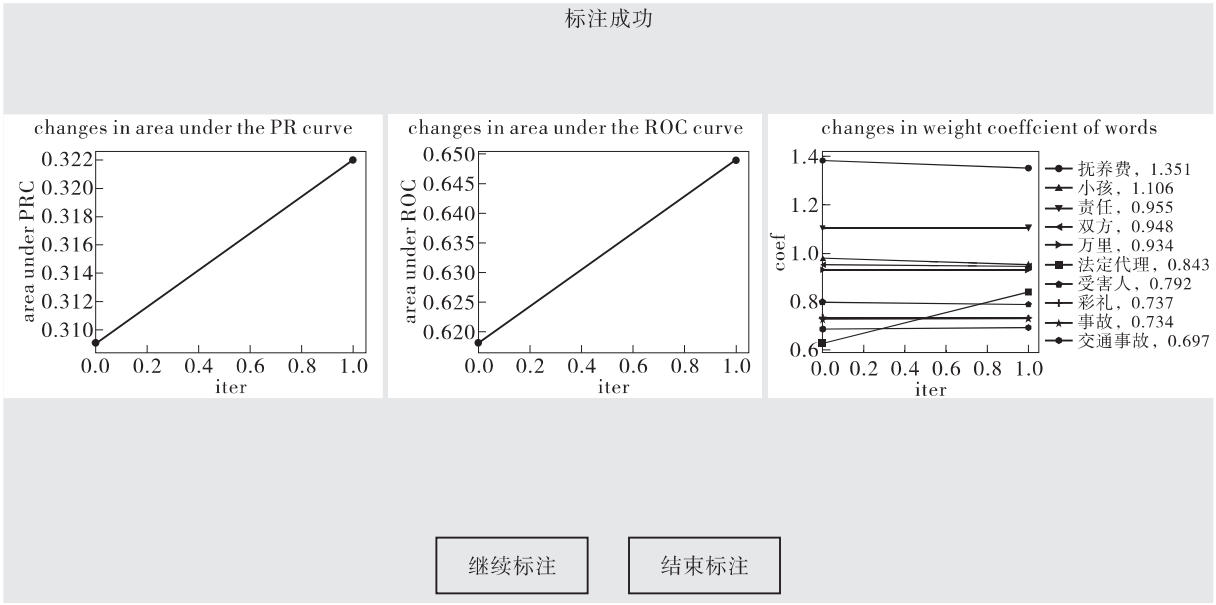


图 7 第二次标注后的测试结果

Fig. 7 Test results of the second labeling

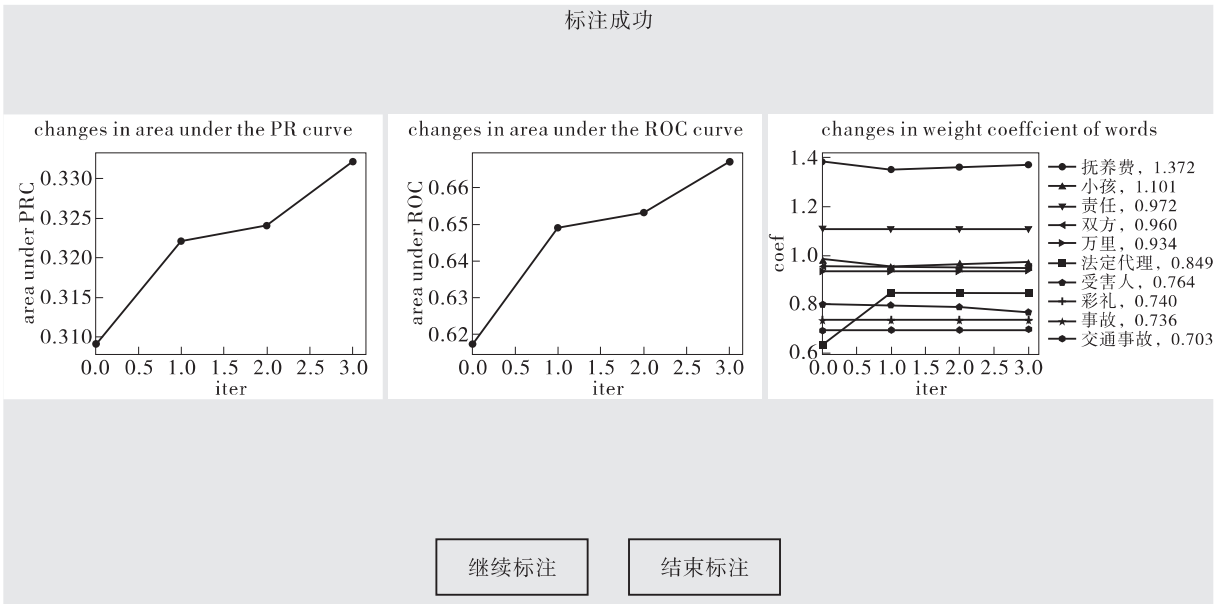


图 8 多次标注后的测试结果

Fig. 8 Test results of several labeling

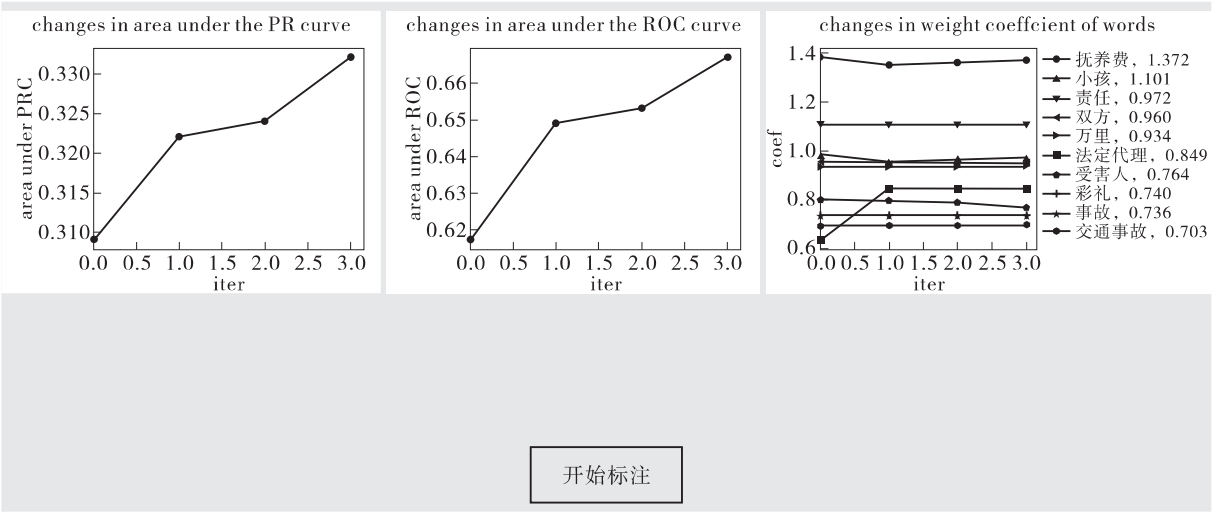


图 9　在首页展示目前最新的测试结果

Fig. 9　Display the latest test results on the homepage

4　结论与展望

针对现实场景中有标注数据集缺少的问题，本文提出了基于半监督学习的未成年人案件文书识别方法。首先将所有案件文书语料按照一定比例划分为训练集和测试集，对语料文本进行预处理之后分别使用 Word2Vec 和 BERT-wwm-ext 对文本进行特征提取，将长语料文本转换为可作为分类模型输入的数据格式；接着采用 PU 学习方法对分类模型进行训练，在正例样本极少的情况下借助大量未标注样本构建有效的分类器；然后在分类模型预测结果的基础上，使用主动学习方法获取关键词并对模型预测结果进行筛选处理，以进一步提升预测效果。为了省时有效地归纳关键词，本文基于主动学习方法，使用线性分类器并以基于词袋模型的特征作为输入进行训练。为更方便地进行训练，本文设计了一个可视化的人机交互系统，通过该系统可观察训练过程中分类器对于各个词的权值系数的变化情况，进而总结出可用于后处理的关键词。经过分类模型预测和关键词后处理，本文识别方法在面向现实场景构建的数据集上可以取得 98.67% 的召回率和 81.02% 的准确率。

本文在进行主动学习训练的过程中设计了一个人机交互系统，目的是方便实验中的标注过程，未来可以考虑将这个系统部署到公共网络中，由更多的标注者来完成更多的标注，进一步提高关键词分析的有效性和精度。本文提出的未成年人案件文书识别方法，克服了样本不均衡的问题，为处理长文本分类任务提出了一个思路，未来还需要在现实应用中检验该识别方法的稳定性。

参考文献：

[1] 郭英崽. 裁判文书上网的隐私权保护问题研究[D]. 南昌：江西师范大学，2015.

[2] KOWSARI K，MEIMANDI K J，HEIDARYSAFA M，et al. Text classification algorithms：a survey[J]. Information，2019，10(4)：150/1-68.

[3] 台德艺，谢飞，胡学钢. 文本分类技术研究[J]. 合肥学院学报（自然科学版），2007，17(3)：61-64.
TAI De-yi，XIE Fei，HU Xue-gang. A research of text categorization[J]. Journal of Hefei University (Natural Science)，2007，17(3)：61-64.

[4] SALTON G，WONG A，YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM，1975，18(11)：613-620.

[5] ZHANG Y，JIN R，ZHOU Z H. Understanding bag-of-words model：a statistical framework[J]. International Journal of Machine Learning and Cybernetics，2010，1：43-52.

[6] RAMOS J. Using TF-IDF to determine word relevance in document queries[C]//Proceedings of the First Instructional Conference on Machine Learning. Piscataway：[s. n.]，2003：133-142.

[7] DUMAIS S T. Latent semantic analysis[J]. Annual Review of Information Science and Technology，2004，38(1)：188-230.

[8] BLEI D M，NG A Y，JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research，2003，3：993-1022.

[9] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space[EB/OL].(2013-09-07)[2020-08-20]. https://arx-

iv.org/abs/1301.3781.

[10] LE Q, MIKOLOV T. Distributed representations of sentences and documents [C] // Proceedings of the 31st International Conference on Machine Learning. Beijing: JMLR, 2014: 1188-1196.

[11] PENNINGTON J, SOCHER R, MANNING C D. Glove: global vectors for word representation [C] // Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014: 1532-1543.

[12] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations [C] // Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics, 2018: 2227-2237.

[13] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. (2020-07-14) [2020-08-20]. https://www.bibsonomy.org/bibtex/273ced32c0d4588eb95b6986dc2c8147c/jonaskaiser.

[14] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C] // Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.

[15] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.

[16] FUKUNAGA K. Introduction to statistical pattern recognition [M]. 2nd ed. San Diego: Academic Press, 2013.

[17] JOULIN A, GRAVE E, 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. Valencia: Association for Computational Linguistics, 2017: 427-431.

[18] ZHANG Y, WALLACE B C. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification [C] // Proceedings of the Eighth International Joint Conference on Natural Language Processing. Taipei: Asian Federation of Natural Language Processing, 2017: 253-263.

[19] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[20] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification [C] // Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin: AAAI Press, 2015: 2267-2273.

[21] YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding [C] // Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019. Vancouver: NeurIPS, 2019: 5754-5764.

[22] LIU B. Web data mining: exploring hyperlinks, contents, and usage data [M]. Berlin/Heidelberg: Springer, 2007.

[23] ELKAN C, NOTO K. Learning classifiers from only positive and unlabeled data [C] // Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas: ACM, 2008: 213-220.

[24] MORDELET F, VERT J P. A bagging SVM to learn from positive and unlabeled examples [J]. Pattern Recognition Letters, 2014, 37: 201-209.

[25] KABOUTARI A, BAGHERZADEH J, KHERADMAND F. An evaluation of two-step techniques for positive-unlabeled learning in text classification [J]. International Journal of Computer Applications Technology and Research, 2014, 3(9): 592-594.

[26] SETTLES B. Active learning literature survey [R]. Madison: University of Wisconsin-Madison, 2009.

[27] SUN S, ZHONG P, XIAO H, et al. An MRF model-based active learning framework for the spectral-spatial classification of hyper spectral imagery [J]. IEEE Journal of Selected Topics in Signal Processing, 2015, 9(6): 1074-1088.

[28] YANG Y Y, LEE S C, CHUNG Y A, et al. Libact: pool-based active learning in python [EB/OL]. (2017-10-01) [2020-08-20]. https://arxiv.org/abs/1710.00379.

(下转第46页)

3D Object Detection Based on Point Cloud Bird's Eye View Remapping

WU Qiuxia LI Lingmin
(School of Software Engineering, South China University of Technology, Guangzhou 510006, Guangdong, China)

Abstract: Image and point cloud are the common data formats for 3D object detection, for images have a superior object recognition capability and point clouds contain accurate spatial information. In order to utilize the above mentioned advantages of both images and point clouds, a 3D object detection method named Bird-PointNet based on bird's eye view of point cloud remapping approach was proposed. First, point cloud was encoded into bird's eye view format for object recognition and rough positioning. Then the results from bird's eye view detection was remapped into the point cloud's space for precise detection. Experiments on the BEV detection benchmark and the 3D detection benchmark of KITTI dataset have demonstrated that the proposed Bird-PointNet method has a higher accuracy of 3D detection, compared with the baseline method that only with bird's eye view coding of point cloud.
Key words: object detection; autopilot driving; point cloud; bird's eye view

(上接第38页)

Juvenile Case Documents Recognition Method Based on Semi-Supervised Learning

YANG Shenghao WU Yueyue MAO Jiaxin LIU Yiqun ZHANG Min MA Shaoping
(Department of Computer Science and Technology//Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: As an important content of judicial information disclosure, case documents should be disclosed to the public after the trial. Some case documents involving juvenile are likely to cause the disclosure of juvenile personal privacy information. In order to conduct targeted privacy protection processing, the first step is to accurately identify documents involving juvenile information from a large number of case documents. At the same time, in order to solve the problem of difficulty in effective supervised learning due to the lack of labeled samples in the real data set, this paper proposed a juvenile case documents recognition method based on semi-supervised learning. Firstly, the corpus text of the case document was pre-processed, and then the features of the text were extracted with Word2Vec and BERT-wwm-ext. After the above processing, the long corpus text was converted into the data format that can be used as the input for the classification model. Then the classification model was trained with the PU learning method, and an effective classifier was constructed with a large number of unlabeled samples under the condition of few positive examples. Then, based on the prediction results of the classification model, active learning method was employed to obtain keywords and screen the prediction results, so as to further improve the prediction effect. Finally, the case documents recognition method proposed in this article achieves a recall of 98.67% and a precision of 81.02% on the test set constructed based on the proportion of real scenes.
Key words: text classification; text feature extraction; deep learning; semi-supervised learning