# 基于深度学习的未成年人案件文书识别方法研究

杨圣豪 吴玥悦 毛佳昕 刘奕群 张敏 马少平
清华大学计算机系北京信息科学与技术国家研究中心

## 论文摘要

在当今这个互联网时代，网络上各式各样的信息会涉及到多个领域，而在法律结构的工作过程中，案件文书作为司法信息的重要内容需要在审判之后向公众公开，在这其中某些涉及未成年人的案件文书则极有可能会造成未成年人的个人隐私信息泄露。我们的目标是从大量案件文书中准确识别并分离出涉及未成年人信息的文书，从而有针对性地对其进行一些隐私保护处理。对于这样一个文本分类问题，我们基于现实数据集从特征工程和分类模型两个方面进行研究。由于现实数据集中有标签样本缺乏，难以进行有效的监督学习，为解决此问题，创新性地引入了半监督学习方法。在数据集样本数量有限的情况下，使用了**PU learning**方法，尽可能使模型的分类性能最大化。另外为提高分类模型的分类效果，使用关键词筛选的方法对分类结果进行后处理。为了解决关键词归纳过程需耗费大量时间成本的问题，我们应用**Active learning**方法分析数据集中样本语料的关键词。经过分类模型预测和关键词后处理，在面向现实场景比例构建的测试集上取得了**97.33%**的召回率和**80.22%**的精确率。

## 系统模型

**PU learning + Active learning**

## 论文简介

本文针对未成年人案件文书识别任务中文书语料较长，正负例样本不均衡等问题引入了两种半监督学习方法，其中**PU learning**用于分类模型的训练过程以获得较好的性能，**Active learning**用于分析文书语料的关键词，从而对分类结果进行后处理进一步提高分类性能。

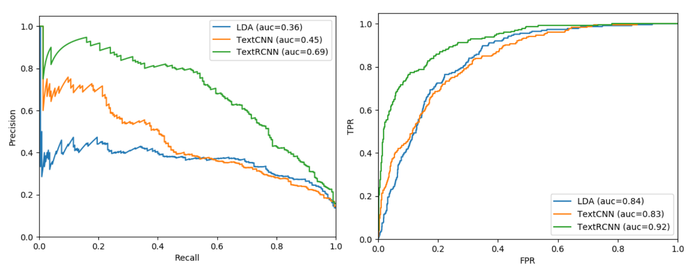
## 算法原理

**PU learning**：是一种使用**Positive**样本和**Unlabeled**样本进行的半监督学习方法，正当有效地利用有限的正例样本和大量未标注样本进行训练得到的分类模型，其对于样本的预测概率应该和该样本属于正例的概率成正相关。

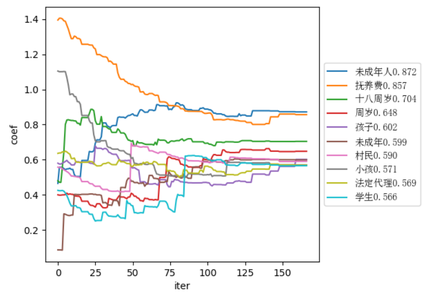
**Active learning**：作为一种半监督方法，其常常被应用到样本不均衡的场景中，其算法本身可以主动地提出对样本的标注，而不是由人来随机抽取样本，这使得我们可以用尽可能低的人工标注成本来获得更好的模型性能。

## 实验仿真

•选择三种分类模型使用**PU learning**方法进行训练，训练后的分类模型对于测试集样本的预测效果如下图中的**P-R**曲线和**ROC**曲线所示：




•使用**Active learning**进行关键词分析，在迭代过程中关键词权值的变化情况如下图所示：




•使用获取到的关键词对于**P-R**曲线和**ROC**曲线表现最好的分类模型**TextRCNN**的预测结果进行后处理，具体方法为：对被分类模型预测为负例的样本使用关键词筛选方法，对于出现了关键词的样本，将其预测结果从负例变为正例，没有出现则保持为负例，从而得到最终的分类结果。

## 论文结论

本文介绍了基于深度学习的未成年人案件文书识别方法，首先对文本特征提取和文本分类的相关工作进行了介绍，并针对现实场景中有标签数据集缺少的问题介绍了**PU learning**和**Active learning**两个半监督学习方法，其中，**PU learning**应用于分类模型的训练中，输入特征使用了**Word2Vec**方法对原始语料进行特征提取后得到的词向量。为了提高分类模型的分类效果，我们在分类模型的预测结果的基础上进行关键词后处理。为了省时有效地归纳关键词，我们基于**Active learning**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过程，我们设计了一个可视化的人机交互系统。通过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，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，我们在面向现实场景构建的数据集上可以取得**0.9733**的**Recall**值和**0.8022**的**Precision**值。

西安电子科技大学 XIDIAN UNIVERSITY