



文章编号:1001-9081(2019)06-1696-05

DOI:10.11772/j.issn.1001-9081.2018109193

## 语义驱动的司法文档学习分类方法

马建刚<sup>1,2,3\*</sup>, 马应龙<sup>4</sup>

(1. 中国人民大学 法学院, 北京 100872; 2. 国家检察官学院, 北京 102206;  
3. 河南省人民检察院, 郑州 450004; 4. 华北电力大学 控制与计算机工程学院, 北京 102206)  
(\*通信作者电子邮箱 704362432@qq.com)

**摘要:** 基于海量的司法文书进行的高效司法文档分类有助于目前的司法智能化应用, 如类案推送、文书检索、判决预测和量刑辅助等。面向通用领域的文本分类方法因没有考虑司法领域文本的复杂结构和知识语义, 导致司法文本分类的效能很低。针对该问题提出了一种语义驱动的方法来学习和分类司法文书。首先, 提出并构建了面向司法领域的领域知识模型以清晰表达文档级语义; 然后, 基于该模型对司法文档进行相应的领域知识抽取; 最后, 利用图长短期记忆模型(Graph LSTM)对司法文书进行训练和分类。实验结果表明该方法在准确率和召回率方面明显优于常用的长短期记忆(LSTM)模型、多类别逻辑回归和支持向量机等方法。

**关键词:** 司法大数据; 领域知识模型; 文本分类; 智慧检务; 图长短期记忆模型

**中图分类号:** TP309    **文献标志码:**A

### Semantic-driven learning and classification method of judicial documents

MA Jiangang<sup>1,2,3\*</sup>, MA Yinglong<sup>4</sup>

(1. Law School, Renmin University of China, Beijing 100872, China;  
2. National Prosecutors College of P. R. C., Beijing 102206, China;  
3. The People's Procuratorate of Henan Province, Zhengzhou Henan 450004, China;  
4. School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China)

**Abstract:** Efficient document classification techniques based on large-scale judicial documents are crucial to current judicial intelligent application, such as similar case pushing, legal document retrieval, judgment prediction and sentencing assistance. The general-domain-oriented document classification methods are lack of efficiency because they do not consider the complex structure and knowledge semantics of judicial documents. To solve this problem, a semantic-driven method was proposed to learn and classify judicial documents. Firstly, a domain knowledge model oriented to judicial domain was proposed and constructed to express the document-level semantics clearly. Then, domain knowledge was extracted from the judicial documents based on the model. Finally, the judicial documents were trained and classified by using Graph Long Short-Term Memory (Graph LSTM) model. The experimental results show that, the proposed method is superior to Long Short-Term Memory (LSTM) model, Multinomial Logistic Regression (MLR) and Support Vector Machine (SVM) in accuracy and recall.

**Key words:** judicial big data; domain knowledge model; text categorization; smart procuratorate; Graph Long Short-Term Memory (Graph LSTM) model

### 0 引言

司法机关通过多年的信息化建设应用已经积累了海量的司法文书, 如最高检察院检察信息公开网 2016 年一年就发布起诉书 779478 份, 最高法院的中国裁判文书网已发布判决书 4677 万份(截止 2018 年 6 月), 为开展司法智能化建设应用(如智慧法院、智慧检务<sup>[1]</sup>)提供了数据基础。基于海量的司法文书进行高效的司法文档分类对目前的司法智能化应用极富价值, 如类案推送、文书检索、判决预测和量刑辅助等。

由于司法文档本身的复杂结构司法文档分类是一项具有挑战性的任务<sup>[2]</sup>。文本自动分类在自然语言处理领域是经典的问题。常用的传统文本分类方法有词频-逆文档频率(Term Frequency-Inverse Document Frequency, TF-IDF)<sup>[3]</sup>、词

袋(Bag Of Words, BOW)模型<sup>[4]</sup>、向量空间模型(Vector Space Model, VSM)<sup>[5]</sup>、LDA (Latent Dirichlet Allocation) 主题模型<sup>[6]</sup>等; 然而, 这些方法往往由于其文本表示通常是高维度高稀疏而导致特征表达能力很弱, 针对司法文本的分类结果并不理想。许多研究基于机器学习方法的分类器来分类司法文档<sup>[7-8]</sup>, 如 K 最近邻(K-Nearest Neighbors, KNN)、支持向量机(Support Vector Machine, SVM)<sup>[9-10]</sup>、最大熵<sup>[11]</sup>、决策树<sup>[12]</sup>等。

面向司法领域的文本分类方法需要考虑特定司法领域文本的复杂结构和知识语义以提高司法文本分类的效能<sup>[13]</sup>。司法文书的文本分类应用对分类准确率有着极高的要求, 且司法领域文本数量大、文本结构复杂。马建刚等<sup>[14]</sup>结合司法文档语义背景知识提出了一种基于知识块摘要和词转移距离

收稿日期:2018-11-15;修回日期:2019-01-03;录用日期:2019-01-08。

基金项目:国家重点研发计划项目(2018YFC0831404, 2018YFC0830605);中国博士后科学基金资助项目(2016M591317)。

作者简介:马建刚(1977—),男,河南郑州人,高级工程师,博士,CCF 高级会员,主要研究方向:大数据、智慧检务、智慧司法;马应龙(1976—),男,陕西咸阳人,教授,博士,CCF 高级会员,主要研究方向:大数据、知识工程。



的高效司法文档分类方法,针对词转移距离模型在处理短文本时具有更好效能的特点,抽取司法文档的核心知识块摘要,进而将针对司法文档的分类转换成针对司法文档知识块摘要的分类,提高了分类的效能;然而,文献[14]中对于确定从司法文档所抽取的知识块摘要中哪些属于对分类至关重要的核心知识块摘要还需要领域专家人工干预和确认,在一定程度上降低了司法文档分类的自动化程度、增加了相应的人工成本开销。

针对上述问题,本文提出了一种语义驱动的深度学习方法来进行司法文本分类。首先,针对具体司法领域构建对应的司法领域知识本体以清晰表达文档级语义;然后,基于领域本体检测司法文档中是否存在与领域知识本体中的术语对应或相似的知识信息,为每一个司法文档生成对应的向量模型;接着,利用图长短期记忆(Graph Long Short-Term Memory, Graph LSTM)模型<sup>[15]</sup>对司法文书进行训练和分类;最后,通过实验证明了所提方法的有效性。实验结果表明,该方法要显著优于常用的长短期记忆模型、多类别的逻辑回归模型和支持向量机方法。本文方法与文献[14]方法虽然都利用了领域背景知识,但处理方法上有以下不同:1)本文方法利用领域本体生成司法文档对应的向量表示而不用获取知识块摘要;2)在领域知识本体构建后,本文方法的司法文档分类后续过程皆可以自动化进行,无需领域专家进一步人工干预;最后,本文方法利用Graph LSTM深度学习模型进行司法文档自动化分类。

## 1 司法文书领域知识模型

一个司法文书包含大量信息,但文档中不同部分的信息对分析司法文档的价值是不一样的。因此,构造一个司法文书领域的知识模型对分析司法文书有很大帮助。基于犯罪构成理论构建司法文书领域知识模型,模型包含犯罪构成的四要件,即:主体、客体、主观方面、客观方面。客观方面又包括危害行为和危害结果,同时还包括文书基本信息(如文号)和判决结果信息。本文以交通肇事罪为例建立了司法文书领域知识模型(Legal Document Model, LDM),如图1所示。交通肇事罪的判决书主要包括文档基本信息、主体、客观方面、判决结果等部分。其中文档基本信息包括判决书文号、审判机关、公诉机关、审判员和审判日期等信息。主体和客观方面这两个概念来自刑法中的犯罪构成要件。主体指被告人的信息,包括姓名、职业、年龄、出生日期、是否有前科、是否累犯等信息。交通肇事罪的客观方面会涉及机动车辆类型、危害行为和危害结果等,危害行为包括醉酒驾驶、追逐竞驶等,危害结果则包括人员伤亡、财产损失等。交通肇事罪的判决结果的主刑包括拘役、有期徒刑等。

本文的司法文书知识模型基于半自动化方式进行构建。首先,采用爬虫技术爬取司法权威机构支持建立的在线司法文档数据,根据页面分析爬取其中的司法分类元数据。本文选择从法信在线类案检索系统(<http://www.faxin.cn/index.aspx>)作为页面爬取的数据源。法信是由最高人民法院支持维护的一个司法文档检索系统,其司法分类系统按照罪名进行分类,每个罪名包含一些更细的子分类,因此具有权威性。其次,根据爬取的页面框架抽取出相关的分类术语。最后,通过具有丰富司法领域10年以上工作经验的司法业务专家审核,

最终以半自动化的方式构建司法文书领域知识模型。

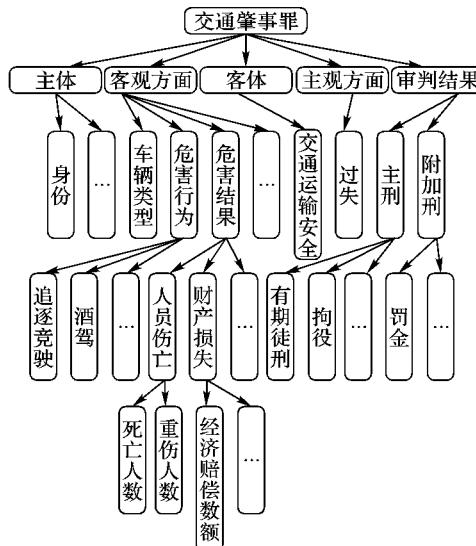


图1 交通肇事罪的司法文书领域知识模型

Fig. 1 Legal document domain knowledge model of traffic accident case

## 2 基于LDM的司法文书自动化知识抽取

自动化知识抽取包括两部分:一是抽取出客观方面部分,这部分内容主要决定了案件的判决结果。二是抽取出司法文书中的判决结果部分,并标准化判决结果,依此为司法文书分类获得可供实验用的带标签的数据集。对每一个司法文书,使用一个可扩展标记语言(extensible Markup Language, XML)文件来保存抽取得到的知识,XML文件的树结构取自于LDM的结构,并与之完全相同。XML文件中的各元素所存储的正是一个司法文书中与LDM的各节点相关的信息,如在图1所示的LDM中,客观方面分支下存有一个酒驾节点,若在一个判决书中检测到犯罪嫌疑人存在酒驾行为,那么在与该判决书对应的XML文件中代表酒驾的元素的值将被设置为1;若未检测到,将被设置为0。

本文采用基于词语相似度匹配和规则的方法来抽取客观方面的知识。需要抽取的知识由LDM确定,不同罪名对应的LDM不同。从图1所示的LDM中可以看到,客观方面中存在两种需要抽取的知识:一是定性的知识,如酒驾、追逐竞驶,只有两种结果,在XML文件中用0代表没有,用1代表有;二是定量的知识,如死亡人数、重伤人数,这种知识需要提取具体的数字。对于定性的知识,首先将判决书分词,然后使用编辑距离判断判决书中的各词与代表待抽取知识的词是否相似,若检测到存在这样一个相似的词,则将XML文件中该元素的值设置为1,否则为0。编辑距离是一种计算词语相似度的算法,计算式如下:

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j), & \min(i,j) = 0 \\ \min\{lev_{a,b}(i-1,j) + 1, lev_{a,b}(i,j-1) + 1, \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)}\}, & \text{其他} \end{cases} \quad (1)$$

例如在抽取酒驾时,如果判决书中存在“喝酒”“酒驾”“醉酒驾驶”等词语时,那么通过编辑距离就能将这些词语判定为酒驾的相似词,就能判定犯罪嫌疑人存在酒驾行为,并在



XML 文件中将对应元素的值设置为 1, 这种做法, 也是基于判决书中可能存在的用词不规范以及自然语言的多样性考虑的。对于定量的知识, 则采用基于规则的方法抽取, 如死亡人数, 会利用“死亡 \* 人”这一规则在判决书中寻找符合的句子, 其中“\*”代表死亡人数, 若能找到, 则将“\*”的值填入 XML 文件的元素中; 若无法找到, 则填入 0, 代表无人死亡。

同样的, 本文采用基于规则的方法抽取审判结果。在司法文档中, 审判结果具有固定的用语和结构, 即被告人 + 姓名 + 犯 + 罪名 + 判处 + 判决结果。利用这个规则, 很容易就能提取出判决结果。本文所抽取的审判结果主要是主刑部分, 这样就能得到形如“有期徒刑五年六个月”的判决结果部分。这里的“五年六个月”中的五和六在文档中是汉字而不是阿拉伯数字, 审判结果的标准化指的是将汉字转化为阿拉伯数字, 同时将月转换为年, 即将“五年六个月”转化为 5.5 年。这样做是为了方便根据刑期对司法文档进行分类。

对于一个保存了抽取所得知识的 XML 文件来说, 可以很容易地使用一个向量来表示整个 XML 的重要信息, 如 XML 中含有  $n$  个元素, 那么可以用一个  $n$  维的向量来代表这个 XML 文件, 向量的每一个分量表示 XML 文件的一个元素值。这个向量可以被认为是保存了一个判决书的关键特征, 基于此向量, 可以作进一步的研究, 如分类、聚类等。这种做法简单明了, 不足的是会丢失 XML 的结构信息。

### 3 基于 Graph LSTM 的司法文书分类

#### 3.1 LSTM 模型

LSTM 是一种循环神经网络 (Recurrent Neural Network, RNN) 的变体, 主要用于序列建模, 其使用门机制处理信息, 解决了 RNN 学习过程中的梯度消失问题, 从而有效地学习到长距离依赖信息。在 LSTM 网络内部, 存在三种门: 输入门、遗忘门和输出门。此外, 相较于普通 RNN 模型, LSTM 内部除了状态  $\mathbf{h}$  之外还有单元状态  $\mathbf{c}$ 。LSTM 用两个门来控制单元状态  $\mathbf{c}$  的内容: 一个是遗忘门, 它决定了上一时刻的单元状态  $\mathbf{c}_{t-1}$  有多少保留到当前时刻  $\mathbf{c}_t$ ; 另一个是输入门, 它决定了当前时刻网络的输入  $\mathbf{x}_t$  有多少保存到单元状态  $\mathbf{c}_t$ 。LSTM 用输出门来控制单元状态  $\mathbf{c}_t$  有多少输出到 LSTM 的当前的输出值  $\mathbf{h}_t$ 。遗忘门公式为:

$$f_t = \sigma(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \quad (2)$$

式中:  $\mathbf{W}_f$  是遗忘门的权重矩阵;  $\mathbf{b}_f$  是偏置项;  $\sigma$  是 sigmoid 函数, 在输入门和输出门的公式中符号意义与之类似。

输入门公式为:

$$i_t = \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \quad (3)$$

输出门公式为:

$$o_t = \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \quad (4)$$

$\tilde{\mathbf{c}}_t$  用于描述当前输入的单元状态, 它是根据上一次的输出和本次输入来计算的:

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c) \quad (5)$$

当前时刻的单元状态  $\mathbf{c}_t$  是由上一次的单元状态按元素乘以遗忘门, 再用当前输入的单元状态按元素乘以输入门, 再将两个积加和产生的:

$$\mathbf{c}_t = f_t \circ \mathbf{c}_{t-1} + i_t \circ \tilde{\mathbf{c}}_t \quad (6)$$

LSTM 最终的输出, 是由输出门和单元状态共同确定的:

$$\mathbf{h}_t = o_t \circ \tanh(\mathbf{c}_t) \quad (7)$$

#### 3.2 基于 Graph LSTM 的司法文书表示和分类

##### 3.2.1 Graph LSTM

Graph LSTM 是一种使用 LSTM 对图类型的数据进行编码的方式, 通常来说这里的图指的是有向无环图, 对于无向图和带环的图, 可以通过拆分的方法将其转换为有向无环图。在 Graph LSTM 中, 一个节点的向量表示是通过其子节点的向量表示学习得来的, 具体而言, 若一个节点  $q$  拥有  $n$  个子节点, 则将这  $n$  个子节点视为一个序列, 然后通过 LSTM 进行序列建模, 即将  $n$  个子节点的向量表示输入到一个 LSTM 中, 最终 LSTM 的输出即为  $q$  的向量表示。对图中所有节点做如此递归的操作, 最终可得到整个图的向量表示。除无子节点的节点之外, 每个节点都有一个与之相对应的 LSTM, 即不同节点的 LSTM 参数不共享。

##### 3.2.2 司法文书表示和分类

对一份判决书进行基于 LDM 的自动化知识抽取后可以得到一个 XML 文件。以图 1 所示的交通肇事罪为例, 得到的 XML 文件包括两部分: 一是客观方面部分; 二是审判结果部分。其中: 客观方面部分经过 Graph LSTM 处理, 得到一个向量表示, 被认为是判决书所描述案情的高级特征; 审判结果部分中主刑的刑期, 则被用来当作分类的标准, 即分类结果。希望本文的模型能对一个判决书中的案情, 也就是案件的客观方面部分进行分类, 得出相应的结果, 即刑期。

具体而言, Graph LSTM 对 XML 中客观方面的处理如图 2 所示。

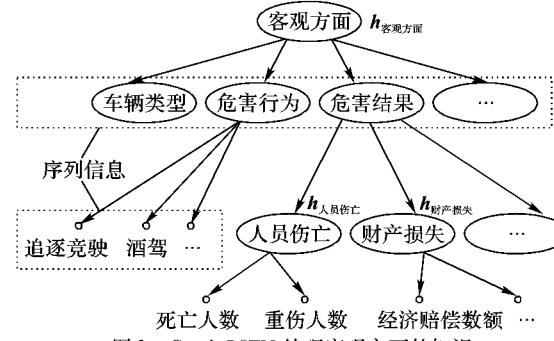


Fig. 2 Processing knowledge of objective aspect with Graph LSTM

图 2 展示了 Graph LSTM 对 XML 信息处理的部分内容, 生成的 XML 文件是树结构, 树是一种简单意义上的图, 所以也可使用 Graph LSTM 对其处理。图中空心小圆代表叶子节点, XML 中属于同一个父节点的叶子节点组成了一组序列信息, 将这组序列信息输入到一个 LSTM 中, 即可得到其父节点的表示, 如图 2 中,  $\mathbf{h}_{\text{人员伤亡}}$  代表“人员伤亡”节点的表示, 是由“死亡人数”“重伤人数”节点的信息经由一个 LSTM 生成的, 而“人员伤亡”“财产损失”等节点的表示又能生成节点“危害结果”的表示, 按这种方式即可递归地得到客观方面的表示, 即  $\mathbf{h}_{\text{客观方面}}$ , 最后即可通过 softmax 层完成分类, 目标函数是负对数似然函数。

对于一个未经审判的案件, 给出其情节, 也就是判决书的客观方面部分, 训练好的模型可以对其进行自动分类, 即给出其刑期, 或者推送类似情节的已判决案件, 以供司法人员参考, 具体分类过程如图 3 所示。其中多层感知机 (Multi-Layer



Perceptron, MLP)是为了增强模型的特征表达能力。

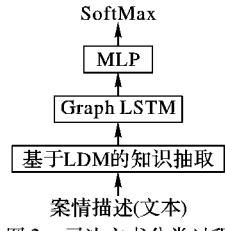


Fig. 3 Judicial document classification process

## 4 实验验证与分析

### 4.1 数据集

实验所用数据集为交通肇事罪判决书,来源于中国裁判文书网(<http://wenshu.court.gov.cn/>),共10 000份,使用其中的80%作为训练集,10%作为验证集,10%作为测试集,数据集的划分是通过随机选择实现的。如前文所说,根据判决结果中主刑的刑期进行分类,具体而言,根据最高人民法院《关于审理交通肇事刑事案件具体应用法律若干问题的解释》中的规定将其刑期划分为4个区间,即:0到6个月,6个月以上到3年,3年以上到7年,以及7年以上。

### 4.2 对比的算法

将本文提出的LDM+Graph LSTM模型与多个算法进行了比较,包括传统的机器学习方法和基于深度学习的算法,传统方法有多类别逻辑回归(Multinomial Logistic Regression, MLR)和SVM,深度学习方法有普通的LSTM。

#### 4.2.1 多类别逻辑回归

多类别的逻辑回归无法处理图数据结构,一种方法是使用一个n维向量(n-vector)作为特征,该向量来自经知识提取之后得到的XML文件,具体可见第2章节所述。在本实验中,根据交通肇事罪的LDM,n取30。另一种方法是使用经典的TF-IDF方法,对于一篇判决书,首先去除审判结果部分,然后将剩余文本的TF-IDF向量作为特征输入到多类别逻辑回归中。

#### 4.2.2 SVM

与多类别逻辑回归相同,基于SVM的方法的输入也是两种,即n维向量和TF-IDF向量。

#### 4.2.3 普通LSTM

普通LSTM对去掉审判结果之后的剩余文本进行序列建模。首先,对文本进行分词等预处理,得到一组词;然后,将所有词按顺序输入到一个LSTM中,得到文本的向量表示,继而通过SoftMax函数进行分类。词由词向量表示,词向量使用的是Word2Vec,在整个数据集上训练得到,维度为200。

### 4.3 模型参数和训练

使用JIEBA<sup>[16]</sup>分词作为分词工具,在实验中,Graph LSTM中各LSTM的隐藏层单元数设置为50,并且使用带动量的随机梯度下降法优化目标函数,批处理的大小为64,学习率设为0.01,动量大小为0.9。

### 4.4 结果分析

实验中使用准确率、召回率和F值作为指标衡量分类效果,其中,F值为准确率和召回率的调和平均值, $F\text{值} = \frac{\text{准确率} * \text{召回率}}{\text{准确率} + \text{召回率}} * 2 / (\text{准确率} + \text{召回率})$ 。在各数据集上不同方法的实验结果如表1所示,表中的Graph LSTM代表本文使用

的基于LDM和Graph LSTM的模型。实验结果表明,相较于其他方法,本文的模型在准确率和召回率上都得到了最好的结果。

对于基于传统机器学习的文本分类方法来说,影响分类效果的因素除了分类方法之外,特征的选择也是很重要的。从表1中可以看到,对于多类别的逻辑回归和SVM这两种方法,使用经过基于LDM的知识提取得到的n维向量作为特征比使用TF-IDF特征能显著地提高分类效果,这证明了经知识提取之后的特征能有效地表达案件情节。

本文模型的分类效果相较于上述使用了n维向量作为特征的两种方法也有很大提升,原因是经过提取所得的知识具有特定结构,而n维向量丢失了这种结构信息,但Graph LSTM能较好地考虑结构信息,因此其分类效果更好。

表1 不同方法的实验结果  
Tab. 1 Experimental results of different methods

分类算法	准确率	召回率	F值
MLR + n-vector	78.7	90.1	84.0
MLR + TF-IDF	74.2	85.3	79.4
SVM + n-vector	80.6	95.2	87.3
SVM + TF-IDF	76.8	86.7	81.5
LSTM	77.3	82.9	80.0
Graph LSTM	94.1	96.2	95.1

本文还通过实验探索了数据集的规模大小对Graph LSTM分类效果的影响,并与传统机器学习方法对比,结果如图4所示。由图4可以看出,在样本数量较少的情况下,Graph LSTM受限于数据集规模,分类效果不如传统的机器学习方法;当逐渐增大数据集规模后,Graph LSTM的分类效果迅速提升,在数据集规模达到6 000份之后,分类效果不再提升,这也是深度学习模型的常见现象。而SVM的分类效果始终变化不大,也就是说,SVM对数据集规模并不敏感。

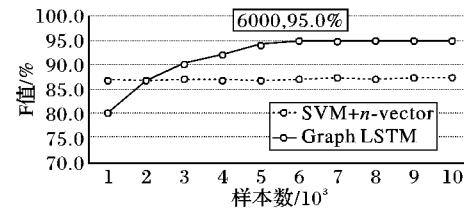


图4 不同数据规模下的分类效果  
Fig. 4 Classification results under different data scales

图5展示了Graph LSTM模型在训练中分类效果随迭代次数的变化趋势。

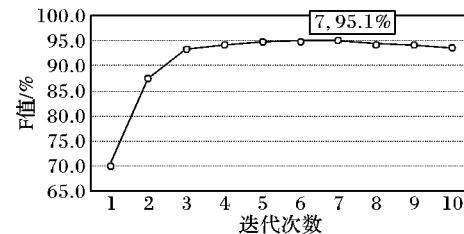


图5 Graph LSTM分类效果随训练迭代次数变化的趋势  
Fig. 5 Trend of Graph LSTM classification effect with number of training iterations

由图5可以看出,在刚开始,该模型效果越来越好,在整个数据集上完成了7次迭代之后,模型的分类效果达到最好,



之后效果逐渐变差。这是因为随着迭代次数的增加,Graph LSTM模型在训练集上出现了过拟合的现象,导致了性能下降。因此,在适当的时候应提前终止模型的训练,避免过拟合。

## 5 结语

本文针对司法文书的相似性分析、实现类案推送为司法人员提供智能辅助办案服务的应用场景,提出了一种语义驱动的司法文档学习分类方法。该方法使用司法领域知识构建了基于领域知识的模型LDM;基于LDM使用结合词语相似度和规则的自动化方法从原始司法文件中提取结构化的知识,并保存到XML文件中;将抽取得到的知识作为原始文本的高级语义特征,并使用Graph LSTM进行分类,相比传统分类方法,显著地提高了分类的效果。

### 参考文献 (References)

- [1] 马建刚. 检察实务中的大数据 [M]. 北京: 中国检察出版社, 2017: 17–23. (MA J G. Procuratorial Big Data [M]. Beijing: China Procuratorial Press, 2017: 17–23.)
- [2] BOELLA G, CARO L D, HUMPHREYS L, et al. Eunomos, a legal document and knowledge management system for the Web to provide relevant, reliable and up-to-date information on the law [J]. Artificial Intelligence and Law, 2016, 24(3): 245–283.
- [3] JING L P, HUANG H K, SHI H B. Improved feature selection approach TF-IDF in text mining [C]// Proceedings of the 2003 International Conference on Machine Learning and Cybernetics. Piscataway, NJ: IEEE, 2003: 944–946.
- [4] GALGANI F, COMPTON P, HOFFMANN A. LEXA: building knowledge bases for automatic legal citation classification [J]. Expert Systems with Applications, 2015, 42(17/18): 6391–6407.
- [5] HAMMOUDA K M, KAMEL M S. Phrase-based document similarity based on an index graph model [C]// Proceedings of the 2002 IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2002: 203–210.
- [6] BLEI D M, NG A Y, JORDAN M I, et al. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(4/5): 993–1022.
- [7] ROITBLAT H L, KERSHAW A, OOT P. Document categorization in legal electronic discovery: computer classification vs. manual review [J]. Journal of the American Society for Information Science and Technology, 2010, 61(1): 70–80.
- [8] NOORTWIJK K V, NOORTWIJK K C. Automatic document classification in integrated legal content collections [C]// ICAIL 2017: Proceedings of the 16th International Conference on Artificial Intelligence and Law. New York: ACM, 2017: 129–134.
- [9] SULEA O, ZAMPIERI M, MALMASI S, et al. Exploring the use of text classification in the legal domain [C]// ASAIL 2017: Proceedings of the Second Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts. New York: ACM, 2017: 419–424.
- [10] SARIC F, DALBELO BASIC B, MOENS M F, et al. Multi-label classification of croatian legal documents using EuroVoc thesaurus [C]// SPLeT 2014: Proceedings of the 2014 Workshop on Semantic Processing of Legal Texts. Reykjavik: European Language Resources Association, 2014: 716–723.
- [11] BAJWA I S, KARIM F, NAEEM M A, et al. A semi supervised approach for catchphrase classification in legal text documents [J]. Journal of Computers, 2017, 12(5): 451–461.
- [12] SILVESTRO L D, SPAMPINATO D, TORRISI A. Automatic classification of legal textual documents using C4.5 [EB/OL]. [2018-10-15]. [http://www.ittig.cnr.it/Ricerca/Testi/Spampinato-Di\\_Silvestro-Torrisi2009.pdf](http://www.ittig.cnr.it/Ricerca/Testi/Spampinato-Di_Silvestro-Torrisi2009.pdf).
- [13] NALLAPATI R, MANNING C D. Legal docket-entry classification: where machine learning stumbles [C]// EMNLP 2008: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2008: 438–446.
- [14] 马建刚, 张鹏, 马应龙. 基于知识块摘要和词转移距离的高效司法文档分类 [J]. 计算机应用, 2019, 39(5): 1293–1298. (MA J G, ZHANG P, MA Y L. Efficient judicial document classification based on knowledge block summarization and word mover's distance [J]. Journal of Computer Applications, 2019, 39(5): 1293–1298.)
- [15] PENG N, POON H, QUIRK C, et al. Cross-sentence n-ary relation extraction with graph LSTMs [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2017: 101–115.
- [16] SUN J J. Jieba Chinese word segmentation tool [EB/OL]. [2018-10-15]. <https://github.com/fxsjy/jieba>.

This work is partially supported by the National Key R&D Program of China (2018YFC0831404, 2018YFC0830605), the Postdoctoral Science Foundation of China (2016M591317).

**MA Jiangang**, born in 1977, Ph. D., senior engineer. His research interests include big data, smart procuratorate, smart judiciary.

**MA Yinglong**, born in 1976, Ph. D., professor. His research interests include big data, knowledge engineering.