

基于因果图分析的可解释司法判决预测方法研究

张虎, 张振, 范越, 郭佳钰

山西大学计算机与信息技术学院, 山西 太原 030006

摘要

随着人工智能技术的发展和海量司法数据的公开, 面向“智慧司法”服务的司法判决预测 (legal judgment prediction, LJP) 任务受到了学术界和工业界的广泛关注, 该任务旨在根据有限的案件事实描述文本来预测案件的罪名、法条和刑期。然而, 现有工作缺乏对易混淆司法案件的智能决策的研究, 且相关模型通常缺乏可解释性, 这会导致模型预测严重依赖领域专家, 阻碍LJP在不同法律体系中的应用。为此, 提出了一种基于因果图分析的司法判决预测 (prediction of legal judgment based on causal graph analysis, CGLJ) 方法, 首先从非结构化的法律事实描述文本中挖掘要素之间的因果关系, 然后采用易混淆罪名聚类的构图方法构建因果图, 既考虑了相似事实描述之间的差异, 又增强了事实描述和法律法规之间的相互作用, 最后将构建好的因果图融入深度神经网络进行联合推理, 得到判决预测结果。此外, 还对模型预测过程中的因果图推理过程进行了可视化, 为判决结果提供了更好的可解释性。在2018中国“法研杯”司法人工智能挑战赛 (CAIL2018) 司法判决预测数据集上的实验结果表明, 该方法相比基线模型取得了更好的效果。

关键词

司法判决预测; 因果图; 可解释性

中图分类号: TP18

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2024023

Research on interpretable legal judgment prediction method based on causal graph analysis

ZHANG Hu, ZHANG Zhen, FAN Yue, GUO Jiayu

School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China

Abstract

With the development of artificial intelligence technology and the disclosure of massive judicial data, the LJP task for "smart justice" services has received widespread attention from academia and industry. The task aims to predict the charges, laws, and sentences of a case based on limited factual descriptions of the text. However, existing work lacks research on intelligent decision-making in easily confusing judicial cases, and related models often lack interpretability, which leads to heavy reliance on domain experts for model predictions and hinders the application of LJP in different legal systems.

To this end, this article proposes a judicial judgment prediction method CGLJ based on causal graph analysis. Firstly, the causal relationships among elements are mined from unstructured legal fact description texts. Then a causal graph is constructed using a composition method of easily confused accusation clustering. It not only considers the difference among similar fact descriptions, but also enhances the interaction between fact descriptions and laws and regulations. Finally, the constructed causality diagram is integrated into a deep neural network for joint inference to obtain the decision prediction result. In addition, this paper also visualizes the causal diagram inference process in the model prediction, providing better interpretability for the judgment result. The experimental result on the CAIL2018 judicial judgment prediction dataset shows that the proposed method achieves better result than the baseline models.

Key words
legal judgment prediction, causal diagram, interpretability

0 引言

近年来, 随着以裁判文书为代表的司法大数据的不断公开和人工智能技术的不断发展, 人工智能技术和司法服务的融合应用已成为“智慧司法”的研究热点^[1]。司法领域中的大部分数据以文本形式呈现, 如判决书、合同和法律意见书等, 而司法判决预测任务是从给定案件的事实描述文本中自动预测判决结果(包括罪行、相关条款和刑期), 因此, LJP模型在“智慧司法”研究中起着至关重要的作用。

一个具备优越性能的LJP模型可以通过预测可靠的判决结果为不熟悉法律的人提供法律援助, 也可以辅助专业人员判案以节省时间。然而, 现有的深度学习模型大多呈现黑盒性质^[2], 无法得到用户的完全信任。此外, 现有工作中针对易混淆司法案件的判决预测研究较少, 模型对这类案件的预测结果常常出现错误, 影响LJP在实际应用中的性能。

目前, 一些研究者尝试通过在文本中挖掘具有因果结构的领域知识来实现一个可解释的推理预测过程。因为文本中的因果知识对深度学习模型的特征选取具有重

要意义, 将存在“因”关系的特征加入深度学习模型, 不仅能提升模型性能, 还能够给出可信的解释^[3-4]。与传统的描述性或预测性任务相比, 因果推理基于反事实的角度出发讨论预测结果, 其目的在于让模型理解对一个变量的干预是如何影响另一个变量的, 具体来说是通过估计一个特定因果强度的大小来描述单个原因变量对一个结果变量的影响大小。然而, 在实际应用中, 如何处理文本中的潜在混杂因素(即要素之间可能存在的潜在因果关系)是因果推理面临的一个重要挑战。混杂因素难以发现, 常常会导致在具体推理时预测出错误的结果, 因此研究易混淆司法案件的判决预测极为重要。

在具体的法律案件中, 不同的法律因素均会影响法律案件的判决结果。然而, 这些因素在不同的案件类型中有很大的不同。因此, 在判决推理时考虑到所有因素是不可行的, 应该考虑与问题最相关和最有用的因素。从因果分析的角度来看, 当法官面对一个涉及多方当事人以及复杂事件的案件时, 因果推理可以帮助法官找出确切可区分的因素。如图1所示, 模型从以下3个角度进行判决推理: 从非结构化的事实描述中获取行为人的直接行为因素, 如打伤、砍伤等; 从非结构化的事实描述中

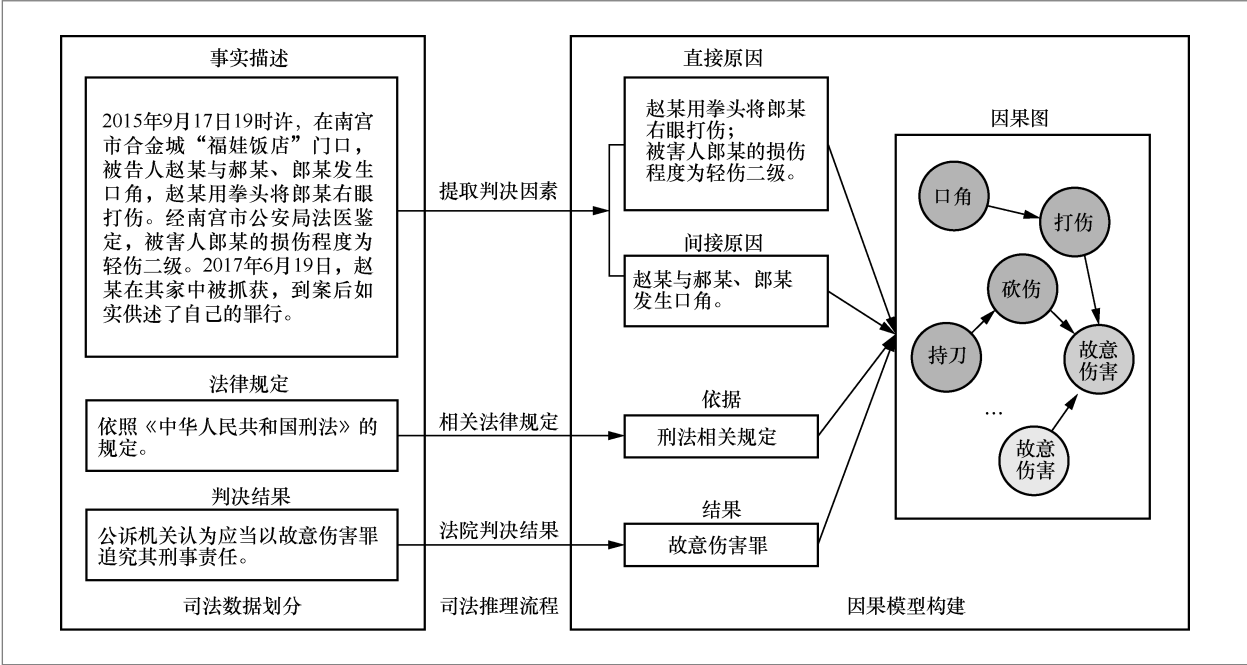


图 1 易混淆司法案件判决推理因果图构建示例

获取行为人的间接行为因素,如口角、持刀等;法律规定中的相关依据。如果能够基于这些不同因素对易混淆的司法案件进行聚类,然后对其进行单独判断,那么模型预测的结果会更有准确性和说服力。此外,可见因素对司法判决结果的支撑使得司法判决过程更加具备可解释性。在此基础上,本文提出了一种基于因果图分析的司法判决预测模型CGLJ,该模型首先采用基于神经网络的方法对易混淆判决结果聚类,然后从不同簇群的法律事实描述中挖掘出要素词之间的因果关系,最后按照不同簇群分别构建因果图,进一步寻找事实描述中的“因”关系进行推理,得到更准确的判决预测结果。

本文主要贡献如下:

①基于裁判文书中非结构化的事实描述和因果逻辑理论,提出了基于易混淆类别标签聚类构建司法因果图的方法,建立了基于因果图和深度神经网络的司法判决

预测模型;

②基于构建的司法因果图对判决预测的推理过程进行了可视化,验证了所提模型对判决预测结果的可解释性;

③在CAIL2018司法判决预测数据集上进行多项对比实验,验证了所提模型的有效性。

1 相关工作

1.1 司法判决预测

司法判决预测是智慧司法领域中重要的研究任务之一,在学术界和工业界受到了广泛关注。由于机器学习方法的局限性,早期的研究工作主要集中在使用数学或统计方法对现有法律案例进行分析。近年来,随着深度学习技术的发展,许多研究人员将LJP任务转化为自然语言处理

(natural language processing, NLP)中的一项分类任务,并尝试通过提取特定任务的表征来提高LJP的性能^[5-6]。此外,一些研究人员探索了如何将NLP技术与LJP的先验知识相结合,进而提升判决预测的准确性。Luo等人^[7]提出了一个基于注意力机制的模型,利用法律条款中的知识来增强LJP的性能;Hu等人^[8]探讨了如何使用基于属性的模型来处理不频繁和混乱的指控;Zhong等人^[9]分析了LJP中不同任务之间的拓扑依赖关系;Ye等人^[10]根据事实和预期指控产生法院意见。然而,虽然现有工作在判决预测准确度上不断提升,但是大多数模型仍然缺乏可解释性,这阻碍了其在现实场景中的应用。

1.2 可解释性

随着深度学习的不断发展,深度学习的“黑盒”特性逐渐引发用户信任问题,因此可解释性引起了研究人员的广泛关注。从可解释性方法的角度来看,现有工作大致分为两个方向。一种是模型内在可解释方法,这类方法将可解释性模块加入模型中一起训练,同最终任务进行联合优化。如Croce等人^[11]通过模型反向计算每个神经元的相关性分数,确定每个样本对测试数据的影响。此外,研究人员发现引入注意力机制为模型的可解释性提供了帮助,如Sun等人^[12]通过注意力权重来分析要素词对预测结果的重要程度。另一种是事后可解释方法,这类方法需要模型在给出预测结果后执行额外的操作以得到解释。Han等人^[13]在训练完模型后,对每个样本使用Mask方法分离出不受样本影响的神经元,以衡量训练样本对测试样本的影响。Wallace等人^[14]利用测试样本的隐层表示找到相近的训练样本来帮助预测。Lei等人^[15]、Narang等人^[16]通过抽取或者生成的方式输出测试样本中对预测结果影响大的

文本内容。近年来,有部分研究者对智慧司法领域的可解释性进行研究。Jiang等人^[17-18]使用深度强化学习的方法从案件的事实描述中提取简短且具有决定性的文字片段作为解释。Zhong等人^[19]提出了一种基于强化学习的迭代问答模型,通过对预测过程进行可视化给出可解释的预测结果。然而在复杂场景下,现有方法得到的解释只能证明与预测结果之间存在有限的相关性,缺乏一定的因果联系,这对于用户理解模型预测结果来说,仍然存在不足。

1.3 因果推理

因果推理是人类的一项核心认知能力,借助因果推理能力,人类可以理解已观测到的各种现象,并预测将来可能发生的事件。因此,将因果推理引入深度学习模型中,可以帮助模型充分考虑数据之间的因果关联,进而提升模型预测结果,还可以得到对模型预测结果的解释证据。近年来,相关研究尝试将文本数据与因果推理方法相结合以提升目标任务的性能。Landeiro等人^[20]在文本分类任务中将文本数据和类别标签之间的因果关系进行建模。Ponti等人^[21]通过发现前馈神经网络中的事件相关特征来识别话语中的隐含因果关系,从而辅助推理。Paul等人^[22]使用倾向性得分匹配包含相似单词倾向的文档,追溯特征对情感的影响是否具有重要的因果关系,从而进行因果推理。这些研究仅仅将文本视为因果图中的单个节点,没有考虑文本内部的混杂因素。研究人员发现将因果关系构建成因果图可以更好地进行推理。例如,Liu^[23]等人提出以自监督方式处理来自法律事实描述中的混杂因素并构建因果图,进而使用因果图消除一些类似指控的歧义。相比传统的因果推理方法,基于因果图的推理方法

尝试将文本作为混杂因素,从多个文本描述中捕捉细微差别,进而获取更详细的因果知识,帮助模型学习到文本数据的因果关系。

2 CGLJ模型

本文提出的CGLJ模型如图2所示,该模型主要包括3个部分,即易混淆类别聚类、因果图构建、判决生成和解释生成。首先,将法律事实描述文本作为模型的输

入,通过预训练语言模型BERT对其进行编码,获得事实描述的文本语义表示;其次,利用具有多层卷积核的卷积神经网络(convolutional neural networks, CNN)获得事实描述的局部特征信息进行预测,根据预测结果对存在类似的易混淆实体进行聚类;然后,针对聚类后的实体分别构建因果图,利用因果图对易混淆实体进行再分类,得到更准确的罪名或法条实体;最后,对该事实描述及罪名法条构建可解释的因果图,从而生成证据进一步提升判决预测的可信度。

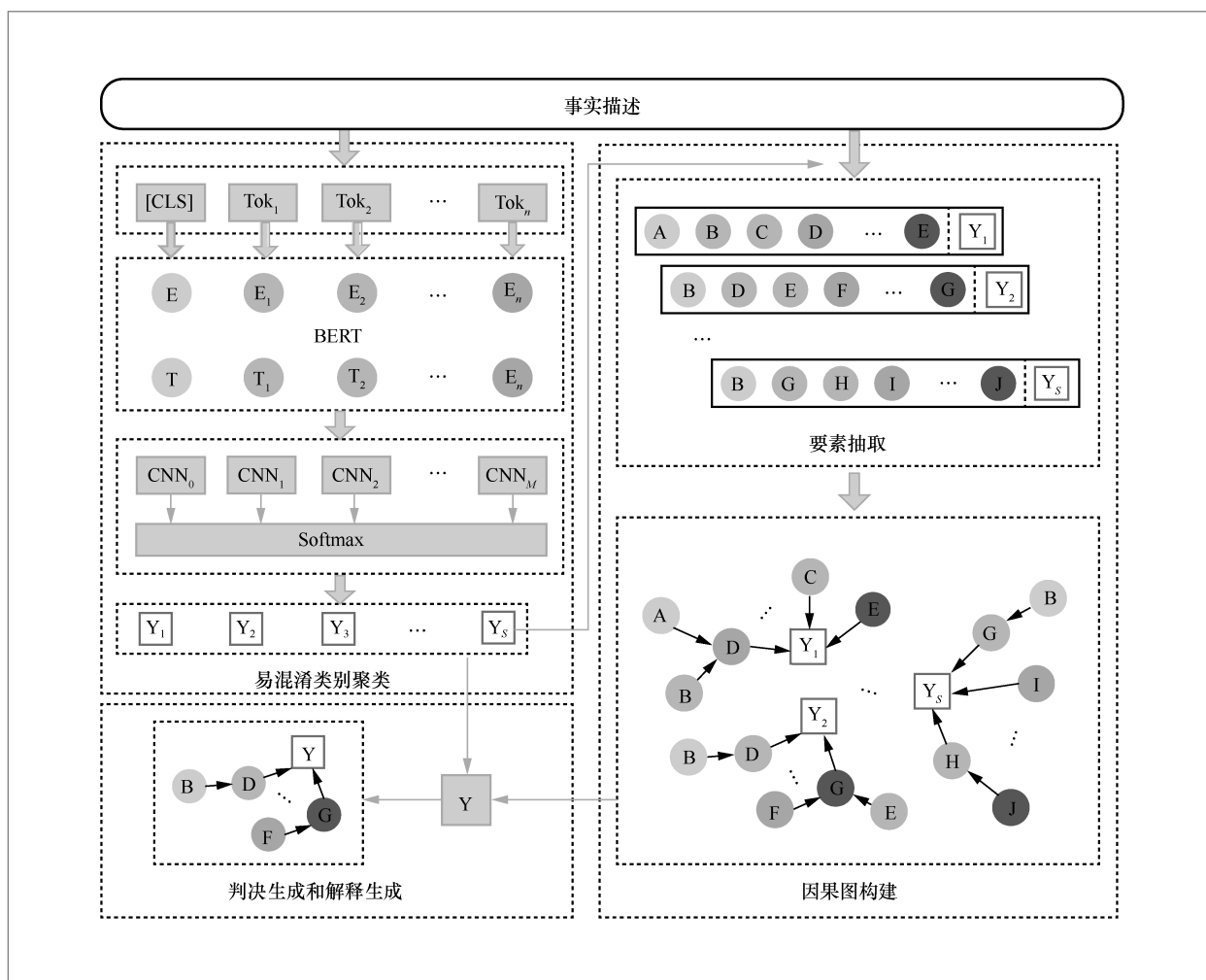


图2 CGLJ模型整体架构

2.1 易混淆类别聚类

易混淆实体之间具有相似的因果结构,针对这样的实体生成各自的因果图可能会导致错误的预测。因此,本文采用神经网络来实现对易混淆实体类别的聚类,具体方法如下。

首先,将刑事案件的事实描述定义为单词序列 $D=\{d_1, d_2, \dots, d_N\}$ 、相关罪名或法条 $C=\{c_1, c_2, \dots, c_N\}$ 分别输入预训练语言模型BERT中,获得字或词的向量表示,如式(1)~式(3)所示。

$$\text{input}=[\text{CLS}]+D+[\text{SEP}]+C+[\text{SEP}] \quad (1)$$

$$\mathbf{S}_i=\text{BERT}(\text{input}) \quad (2)$$

$$\mathbf{S}=\text{concat}(\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_n) \quad (3)$$

其中,特殊字符[CLS]为序列的第一个字符,[SEP]为分隔符,input为BERT的输入。由于案件事实描述多为复杂长文本场景,而BERT模型输入长度为512,因此,将事实描述中的每个句子分别输入BERT进行编码,得到每个句子的向量表示 $\mathbf{S}_i$,将其拼接得到事实描述文本的向量表示 $\mathbf{S}$ 。本文使用BERT的base版本作为编码器,得到的每个字或词的向量维度为768。

其次,经过BERT编码后的事实描述文本输入CNN模型进一步提取其局部特征信息。CNN模型利用不同的卷积核提取不同维度的特征,然后融合这些特征得到更好的语义信息表达。具体来说,使用不同的滤波器 W 在特征矩阵 $\mathbf{S}$ 上获取各自的局部特征 $\mathbf{C}_i$,并将其拼接组成新的特征向量 $\mathbf{k}$,如式(4)~式(5)所示。

$$\mathbf{c}_i=\tanh(W\mathbf{S}_{(i, i+h-1)}+b) \quad (4)$$

$$\mathbf{k}=[\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{n-h+1}] \quad (5)$$

其中, b 是偏置项, $\mathbf{S}_{(i, i+h-1)}$ 表示特征向量 $\mathbf{S}$ 中第 i 到 $i+h-1$ 项的局部特征, i 代表了滑动窗口的大小, h 表示局部特征的长度。拼接不同的滤波器得到案情事实描述特征 $\mathbf{C}$,然后选取最大值作为案情事实描述上的最终特征表示 $\hat{\mathbf{C}}$,再利用Softmax函数将特征 $\hat{\mathbf{C}}$ 转化每个罪名或法条所对应的标签概率,如式(6)~式(8)所示。

$$\mathbf{C}=\text{concat}(\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_s) \quad (6)$$

$$\hat{\mathbf{C}}=\max(\mathbf{C}) \quad (7)$$

$$q_M=\text{softmax}(W'\hat{\mathbf{C}}+b) \quad (8)$$

其中, W' 为模型参数, b 为偏置, M 为标签数。选取最大值对应的标签作为预测实体标签,如式(9)所示。

$$L_s=\arg\max(q_M) \quad (9)$$

对比预测标签与真实标签,选取多次预测错误的标签与真实标签进行易混淆聚类,如式(10)所示。

$$\alpha=\left(\sum_{i=1}^m f(L_s \neq L)\right)-\varepsilon \quad (10)$$

其中, L 为真实标签, ε 为超参数。当 $\alpha \geq 0$ 时,表明预测错误的标签与真实标签存在明显的易混淆关系,将其聚类在一个簇群内;当 $\alpha < 0$ 时,表明预测错误的标签与真实标签不存在明显的易混淆关系,将真实标签单独构建一个新的簇群。在聚类时,对于已经出现簇内的实体标签不再重复添加。此时,对于有 M 个罪名或法条实体的集合,将其划分为 h 个簇,每个簇内的实体标签都可表示为 $Y=\{y_1, y_2, \dots, y_s\}$,其中 s 为每个簇内各自拥有的实体标签数。这样属于两个不同簇的一对实体被分隔开来,同时构成的因果图也保留了各自因果关系的独立性。

2.2 因果图构建

基于易混淆类别聚类方法,本文使用最大祖先图(maximal ancestral graph, MAG)对簇内所有实体类别进行建模。MAG是一种对有向无环图(directed acyclic graph, DAG)进行推广的混合图,包含DAG的所有特征^[24]。MAG包含两种边,即单向边($\rightarrow$)和双向边($\leftrightarrow$)。对于因果图来说,不存在互为因果的情况,因此,本文去除双向边的情况,只保留单向边。基于此构建的因果图,每个实体类别都与一个MAG图相关联,属于同一簇群的实体都被包含在统一的MAG图中。接下来对如何构建因果图展开描述。

首先,提取事实描述文本中的要素词。本文使用清华大学中文词法分析工具包THULAC^[25]对案情事实描述进行分词,获得因果图的候选节点。为了对每个实体标签提取更多可区分的要素词,采用YAKE^[26]方法计算要素词对实体标签的重要性。YAKE方法主要采用词的位置、全文词频、上下文之间的关系、句子间的词频4个指标来计算候选要素词得分。然而,很多候选要素词具有类似的含义。因此,本文引入k-means聚类算法,从候选要素词中选出重要性得分较高的*i*个单词,将其聚类为*j*个类,将*j*个要素词类与*M*个实体标签一起构成因果图的节点。对于每一个案情事实描述来说,这些要素词和实体标签都是二元的,即存在记为0,不存在记为1,将其记录起来作为新的事实描述表示。

其次,挖掘不同要素词以及实体标签之间的因果关系,进而为司法因果图构边。本文引入贪婪快速因果推理(greedy fast causal inference, GFCI)^[27]方法作为构图方法来生成MAG图。GFCI结合了基于分数和基于约束的算法,从基于分数

的角度来说,主要在上一步生成的事实描述表示中采用快速贪婪搜索的方法搜索到得分最高的图。这可能会导致因果关系的缺失或者错误,因此,GFCI采取基于约束的方法对其中的一些因果关系进行限制或重构。本文从司法判决的角度针对以下3种情况进行约束。第一,限制实体标签到要素词的边,例如由于“砍杀”导致判处“故意杀人罪”可以成立,反过来则不成立。第二,案情事实描述通常按照时间线的先后顺序来书写,因此,不允许时间靠后的要素词导致时间靠前要素词的情况出现。第三,法律法规中出现的行为往往是判处相应法条或罪名的直接原因,因此,在因果图连边时,增加这类要素词和实体标签之间的边。基于此,通过GFCI可以初步生成一个MAG图。

然后,构建事实描述因果图,即通过生成的MAG图抽取出因果图。此时,通过GFCI构成的MAG图中有3种可能边: $\rightarrow$ 、 $\leftarrow$ 和 $\leftrightarrow$ 。对于上述情况,随机抽取*Q*种组合来构建因果图。针对不存在互为因果的情况,通过GFCI方法只保留单向边而删除双向边。由此构建了*Q*种可能的抽样因果图。

最后,对所有边的因果强度进行估计。在讨论某个因果事件时,导致“果”的“因”可能会有多个,且每个“因”起到的作用都有所不同。因此,本文通过因果关系的强度来细化抽样的因果图,对于具有强因果关系的边赋予高的因果强度,没有因果关系或具有弱因果关系的边赋予近零的强度。为此引入do算子,将平均治疗效果(average treatment effects, ATE)作为因果强度的衡量标准。具体来说,给定一个边 $X \rightarrow Y$,如果将原因*X*从1改为0,判断结果*Y*如何变化,如式(11)所示。

$$\psi_{X,Y} = E[Y | \text{do}(X = 1)] - E[Y | \text{do}(X = 0)] \quad (11)$$

其中, E 表示期望, $do(X=1)$ 表示对 X 的干预。此外, 利用倾向评分匹配(propensity score matching, PSM)来估计ATE。PSM是一种使用反事实推断作为理论框架的方法, 通过干预效应分析来计算有该因果关系的概率。此时图 G 中 $X \rightarrow Y$ 边的强度表示如式(12)所示。

$$\hat{\psi}_{X,Y}^G = \left[\sum_{i:x_i=1} (y_i - y_j) + \sum_{i:x_i=0} (y_j - y_i) \right] / N \quad (12)$$

其中, X_i, y_i 分别是实例 i 的要素词、实体标签, j 为 i 的负样本中最相似的实例。由此构建的 Q 种抽样因果图都有对应不同的 $X \rightarrow Y$ 边的因果强度 $\hat{\psi}_{X,Y}^G$ 。此时, 初步构成了易混淆簇群的因果图。

2.3 判决生成和解释生成

基于上述得到的聚类因果图, 本文融合深度神经网络与因果图进行联合推理。通过上述神经网络方法, 初步确定可能存在的司法判决结果 L_s 。对于 L_s 所在因果图的簇内只有一个实体标签的情况, 仍使用 L_s 作为最后的司法判决结果。而当其所在因果图的簇内有多个易混淆实体标签时, 只通过神经网络仍难以准确预测判决结果。因此, 针对易混淆实体标签, 本文利用 L_s 所在簇的抽样因果图再次进行推理, 此时只需要从案情事实描述中提取要素词并将其与因果图对应, 然后确定哪个实体标签更适合这种情况。具体来说, 计算 Q 个抽样因果图中每个原因要素词 X_j 对于实体 Y_i 的总体因果强度, 如果为0, 则表示不存在该因果关系, 如式(13)所示。

$$\tilde{\psi}_{X_j,Y_i} = \sum_{q=1}^Q \hat{\psi}_{X_j,Y_i}^{G_q} \quad (13)$$

对于需要判决的案情描述, 计算每个

实体标签的分数, 如式(14)所示。

$$S(Y_i) = \sum_{X_j \in \text{Tr}(Y_i)} \tilde{\psi}_{X_j,Y_i} \times \gamma(X_j) \quad (14)$$

其中, $\gamma(X_j)$ 为虚拟变量, 如果 X_j 存在于案情描述中, 该值为1, 反之则为0。 $\text{Tr}(Y_i)$ 为因果图中关于实体标签 Y_i 的所有原因要素词集合。此时, 取最大值对应的实体标签作为该案情描述的判决结果, 并且将该实体标签所在的因果图簇作为解释进行输出。

3 实验结果与分析

3.1 实验数据

本文在2018中国“法研杯”司法人工智能挑战赛(CAIL2018)的司法判决预测数据集CAIL-big上进行实验。为了进一步探索CGLJ模型在单分类问题上的实用性, 筛选出适用多个法律条款和多个罪名指控的案例样本, 将其分解为多个单分类问题的案例样本, 处理后的数据集为CAIL-deal, 两个数据集的统计结果见表1。此外, 刑期预测通常会受时间、地点因素的干扰, 因此, 本文主要针对法条预测和罪名预测两个子任务开展研究。

3.2 评价指标

司法判决的结果预测本质上可以看作多个二元分类问题, 而且CAIL-big数据集存在数据极度不平衡的情况。因此, 主要考虑宏F1(macro-F1)指标来评估预测结果的正确性, 具体计算式如下。

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (15)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

$$\text{F1-score}_i = 2 \frac{\text{recall}_i \times \text{precision}_i}{\text{recall}_i + \text{precision}_i} \tag{17}$$

$$\text{macro-F1} = \frac{\text{F1-score}_1 + \text{F1-score}_2 + \dots + \text{F1-score}_n}{n} \tag{18}$$

其中，TP为实际为此类且预测也为此类的案例个数，FP为预测为此类但实际不为此类的案例个数，FN为实际为此类但预测不为此类的案例个数。macro-F1实际上是所有类别中F1-score的平均值。

3.3 实验设置

实体标签聚类模块使用官方的中文BERT模型作为事实编码器，事实描述被处理为字符序列，最大输入序列长度设置为512。实验中采用Adam优化器，将学习速率设置为 2×10^{-5} ，epoch设置为3，batch_size设置为16，并通过司法判决预测数据集的特点对BERT模型进行微调，从而提升下游任务的效果。

在构图模块中，本文选取不同的训练比例（1%、10%、30%、50%、100%）来研究不同训练比例对性能差距的影响。不同训练比例会得到不同数量的关键词。因此，当训练集比例为1%时，选择20个关键词，并将这些关键词聚类为15个原因要素；当训练集比例为10%和30%时，选择30个关键词，并将这些关键词聚类为25个原因要素；当训练集比例为50%和100%时，选择60个关键词，并将这些关键词聚类为40个原因要素。考虑到GFCI方法构成MAG图的特点，本文针对不同训练比例选取相同数量的其他实体标签对应的案情描述作为负样本。此外，其他超参数都是根据经验选择的，对于同一数据

表1 数据集统计

数据集	训练集样本数	测试集样本数	法条数	罪名数
CAIL-big	1 587 979	185 120	118	130
CAIL-deal	1 395 496	164 173	67	85

集中的不同模型和相同的训练集，比例保持不变。

3.4 实验结果与分析

本文在法条预测与罪名预测两个司法判决预测子任务上进行了实验，实验结果见表2、表3。与CNN、LSTM的对比，本文模型CGFJ在两个子任务上的性能普遍优于单纯使用神经网络进行推理的性能。其中，使用CNN模型对100%的数据

表2 不同模型在法条预测任务上的宏F1值

模型	法条预测				
	1%	10%	30%	50%	100%
LSTM	0.34	0.41	0.51	0.66	0.79
CNN	0.32	0.43	0.57	0.74	0.83
BERT+CNN	0.35	0.48	0.62	0.80	0.89
LSTM+Casual	0.51	0.57	0.64	0.75	0.87
CNN+Casual	0.54	0.61	0.73	0.87	0.91
CGFJ(BERT+CNN+Casual)	0.55	0.62	0.75	0.89	0.937

表3 不同模型在罪名预测任务上的宏F1值

模型	罪名指控预测				
	1%	10%	30%	50%	100%
LSTM	0.29	0.34	0.43	0.54	0.69
CNN	0.31	0.37	0.48	0.62	0.79
BERT+CNN	0.34	0.41	0.56	0.69	0.84
LSTM+Casual	0.46	0.51	0.57	0.64	0.76
CNN+Casual	0.47	0.54	0.61	0.71	0.83
CGFJ(BERT+CNN+Casual)	0.51	0.56	0.65	0.77	0.87

进行训练时,在表2的法条预测任务上性能提升了0.107,在表3的罪名预测任务上性能提升了0.08。在CNN、LSTM上分别加入因果思想后,模型的效果都有所提升。表3中,对100%的数据进行训练时,LSTM+Casual的判决推理准确率比单纯使用LSTM高0.07。神经网络虽然具有从大量非结构化数据中学习要素词的优势,但是难以直接识别易混淆实体类别中的可区分要素词。而本文模型构建的因果图可以准确提取出可区分要素词聚类构图,在极大保留因果关系独特性的同时获得更好的推理能力。此外,与第2.1节的易混淆类别聚类方法使用的模型相比,BERT+CNN比LSTM、CNN表现得更好,这表明BERT+CNN能够更好地进行聚类类别标签。

同时,本文模型有助于判断浓缩的关键信息,尤其是在小样本数据上表现突出。从实验结果来看,一方面,在小样本数据中,基于因果的方法比神经网络方法表现得更稳定;另一方面,神经网络在少量数

据的训练下往往会出现司法判决结果出错的情况。然而,随着训练数据的增加,性能差距变得越来越小。例如,在表2法条预测任务中,使用1%的数据作为训练数据时,CGFJ比CNN的准确率高0.23,这远远高于100%训练数据时两者的差距。

3.5 案例分析

为了直观验证CGFJ提取可区分特征的效果,图3对非国家工作人员受贿罪和受贿罪这两个易混淆罪名进行举例说明,左侧加粗部分为事实描述聚类编码后注意力权重较大的词语。这两个案例中的重要信息非常相似,都包含有“职务便利”“谋取利益”“人民币”等词,仅由此进行判决预测,都会被判为受贿罪。因此,当仅简单使用神经网络来预测法律条款、罪名指控时,这类案例往往会被误判。故将这两个案例的事实描述与对应的因果图簇(如图3右侧部分)进行对照,专注于提取可区分的要素词,如被告的身份信息(如“海某城建燃气

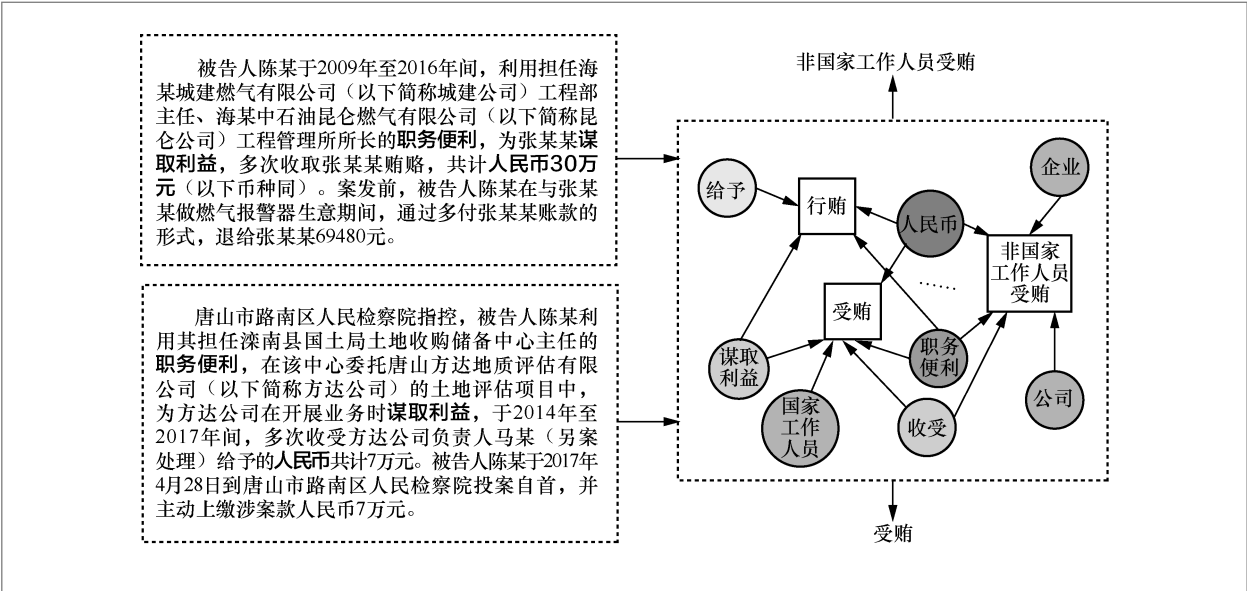


图3 案例分析图

有限公司工程部主任”“滦南县国土局土地收购储备中心主任”），从而有效区分这两个案件适用的罪名指控。

4 结束语

本文针对司法判决预测任务，提出了一个基于因果推理的模型CGFJ，以解决司法领域中易混淆判决的问题。在CGFJ中，提出了一种聚类构建因果图的方法，用于区分易混淆的法律条款、罪名指控的关键特征。该方法不仅考虑了相似事实描述之间的差异，还考虑了事实描述和法律法规之间的相互作用。此外，CGFJ的可视化推理提供了一定的可解释性。然而，目前本文研究未考虑CGFJ在多分类问题中的有效性。在未来的研究工作中，将继续探索案件事实描述中的复杂情况的因果关系，例如多个被告或者多个指控的多分类问题。

参考文献:

- [1] PEARL J. The seven tools of causal inference, with reflections on machine learning[J]. Communications of the ACM, 2019, 62(3): 54–60.
- [2] RIBEIRO M T, SINGH S, GUESTRIN C. Why should I trust you?: explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135–1144.
- [3] LOFTUS J R, RUSSELL C, KUSNER M J, et al. Causal reasoning for algorithmic fairness[EB]. arXiv preprint, 2018, arXiv: 1805.05859.
- [4] BAREINBOIM E, PEARL J. Causal inference and the data-fusion problem[J]. Proceedings of the National Academy of Sciences of the United States of America, 2016, 113(27): 7345–7352.
- [5] LIU C L, HSIEH C D. Exploring phrase-based classification of judicial documents for criminal charges in chinese[C]//International Symposium on Methodologies for Intelligent Systems. Berlin: Springer, 2006: 681–690.
- [6] LIU C L, CHANG C T, HO J H. Case instance generation and refinement for case-based criminal summary judgments in Chinese[J]. Journal of Information Science and Engineering, 2004, 20(4): 783–800.
- [7] LUO B, FENG Y, XU J, et al. Learning to predict charges for criminal cases with legal basis[EB]. arXiv preprint, 2017, arXiv: 1707.09168.
- [8] HU Z, LI X, TU C, et al. Few-shot charge prediction with discriminative legal attributes[C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018: 487–498.
- [9] ZHONG H X, GUO Z P, TU C C, et al. Legal judgment prediction via topological learning[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 3540–3549.
- [10] YE H, JIANG X, LUO Z, et al. Interpretable charge predictions for criminal cases: learning to generate court views from fact descriptions[EB]. arXiv preprint, 2018, arXiv: 1802.08504.
- [11] CROCE D, ROSSINI D, BASILI R. Auditing deep learning processes through kernel-based explanatory models[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: EMNLP-IJCNLP, 2019: 4037–4046.

- [12] SUN Z, FAN C, HAN Q, et al. Self-explaining structures improve NLP models[EB]. arXiv preprint, 2020, arXiv: 2012.01786.
- [13] HAN X C, WALLACE B C, TSVETKOV Y. Explaining black box predictions and unveiling data artifacts through influence functions[EB]. arXiv preprint, 2020, arXiv: 2005.06676.
- [14] WALLACE E, FENG S, BOYD-GRABER J. Interpreting neural networks with nearest neighbors[EB]. arXiv preprint, 2018, arXiv: 1809.028478.
- [15] LEI T, BARZILAY R, JAAKKOLA T. Rationalizing Neural Predictions[EB]. arXiv preprint, 2016, arXiv: 1606.04155.
- [16] NARANG S, RAFFEL C, LEE K, et al. WT5? ! training text-to-text models to explain their predictions[EB]. arXiv preprint, 2020, arXiv: 2004.14546.
- [17] JIANG Y C, BANSAL M. Self-assembling modular networks for interpretable multi-hop reasoning[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). Stroudsburg: Association for Computational Linguistics, 2019: 4474–4484.
- [18] JIANG Y C, JOSHI N, CHEN Y C, et al. Explore, propose, and assemble: an interpretable model for multi-hop reading comprehension[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 2714–2725.
- [19] ZHONG H X, WANG Y Z, TU C C, et al. Iteratively questioning and answering for interpretable legal judgment prediction[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 1250–1257.
- [20] LANDEIRO V, CULOTTA A. Robust text classification in the presence of confounding bias[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2016, 30(1): 186–193.
- [21] PONTI E M, KORHONEN A. Event-related features in feedforward neural networks contribute to identifying causal relations in discourse[C]//Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics. Stroudsburg: Association for Computational Linguistics, 2017: 25–30.
- [22] PAUL M J. Feature selection as causal inference: experiments with text classification[C]//Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Stroudsburg: Association for Computational Linguistics, 2017: 163–172.
- [23] LIU X, YIN D, FENG Y S, et al. Everything has a cause: leveraging causal inference in legal text analysis[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2021: 1928–1941.
- [24] PEARL J, MACKENZIE D. The book of why: the new science of cause and effect[M]. London: Penguin, 2018.
- [25] SUN M, CHEN X, ZHANG K, et al. Thulac: an efficient lexical analyzer for Chinese[J]. Retrieved Jan, 2016, 10: 2022.
- [26] CAMPOS R, MANGARAVITE V, PASQUALI A, et al. YAKE! Keyword extraction from single documents using multiple local features[J]. Information Sciences, 2020, 509: 257–289.
- [27] OGARRIO J M, SPIRITES P, RAMSEY J. A hybrid causal search algorithm for latent variable models[C]//The Eighth Edition of the International Conference on Probabilistic Graphical Models, Lugano: PGM, 2016, 52: 368–379.

作者简介



张虎 (1979-), 男, 博士, 山西大学计算机与信息技术学院教授, 山西大学计算机与信息技术学院 (大数据学院) 副院长, 主要研究方向为自然语言处理、大数据挖掘与分析。



张振 (1998-), 男, 山西大学计算机与信息技术学院硕士生, 主要研究方向为自然语言处理。



范越 (1997-), 男, 山西大学计算机与信息技术学院博士生, 主要研究方向为自然语言处理。



郭佳钰 (1996-), 女, 山西大学计算机与信息技术学院硕士生, 主要研究方向为自然语言处理。

收稿日期: 2023-06-09

通信作者: 张虎, zhanghu@sxu.edu.cn

基金项目: 国家自然科学基金资助项目 (No.62176145)

Foundation Item: The National Natural Science Foundation of China (No.62176145)