采用反事实数据增强方法的储粮害虫事件因果强度计算

肖 乐，赵 婧，徐云飞
（河南工业大学信息科学与工程学院，郑州 450000）

摘 要：储粮害虫是影响粮食安全的重要因素，深入研究储粮害虫事件的发展过程及其因果关系极为关键。通过量化分析储粮害虫事件之间的因果强度，能够更加准确地评估潜在风险，帮助相关工作人员制定防控措施，减少不必要的损失。为解决储粮害虫领域数据中存在的数据偏差而造成模型过分依赖数据集中的表面特征，在应对泛化数据时效果不佳的问题，该研究提出一种反事实数据增强的因果强度计算方法，旨在准确量化事件之间的因果强度。设计了一个反事实数据增强的因果强度计算框架（counterfactual data augmentation-event causal strength，CDA-ECS），通过利用大语言模型（large language model，LLM）生成反事实实例，对原始数据进行扩展，将去偏的因果知识整合进预训练语言模型中，帮助其更深入地理解和学习句子的因果关系，提高模型的泛化能力。在公共数据集和领域数据集上的试验表明，所提方法能够训练出更加稳健的模型，在领域泛化数据的推理任务上准确率提高了 2.4 个百分点，能有效应用于储粮害虫事件的因果强度计算。在储粮害虫领域，反事实数据增强方法的引入为解决数据偏差提供了一种新的视角，增强后数据的多样性和复杂性使得模型能够更加深入地理解害虫行为与环境因素之间的复杂联系，进一步帮助实现储粮害虫事件的风险分析。该研究证明了反事实数据增强方法的可行性和有效性，为实现储粮害虫事件的因果强度计算提供了一定的参考。

关键词：储粮害虫；数据增强；大语言模型；反事实生成；因果强度
doi：10.11975/j.issn.1002-6819.202408083
中图分类号：TP391.1 **文献标志码**：A **文章编号**：1002-6819(2024)-24-0190-09

肖乐，赵婧，徐云飞. 采用反事实数据增强方法的储粮害虫事件因果强度计算[J]. 农业工程学报，2024，40(24)：190-198. doi: 10.11975/j.issn.1002-6819.202408083 http://www.tcsae.org
XIAO Le, ZHAO Jing, XU Yunfei. Calculation of the causal strength of stored grain pest events augmented using counterfactual data method[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(24): 190-198. (in Chinese with English abstract) doi: 10.11975/j.issn.1002-6819.202408083 http://www.tcsae.org

0 引 言

粮食安全是国家安全的根基，而储粮害虫是危害粮食安全的一个主要因素。据相关部门统计，全球每年因储粮害虫等原因，就有 8%～14%的粮食受到损失[1]，这不仅影响农民收入和粮食产出，更对国家粮食安全构成了潜在风险。粮食的损失意味着农业投入的浪费，同时也增加了国家粮食储备的成本，进一步加剧经济损失。因此，明确粮堆内储粮害虫的发展规律并推理储粮害虫事件可能造成的后果，相关工作人员能够及时做出防治措施，避免粮食重大损失，保障粮食安全。近年来，知识图谱的应用为储粮害虫知识的管理提供了新的视角，它能够以结构化的方式表示和组织储粮害虫相关的知识，包括害虫的种类、属性特征、生命周期、偏好环境、危害粮种、防治措施等。目前，储粮害虫知识图谱已被用于辅助粮情决策支持[2]。进一步，肖乐等[3]构建了储粮害虫事理图谱，以更深入地理解储粮害虫的行为模式，为害虫的监测、预警和防控提供了有力的工具。为了更深入地剖析害虫事件之间的内在联系，需充分整合与利用储粮害虫相关研究中害虫活动规律、发展趋势以及与环境相互作用等涉及的重要信息，以揭示更为复杂的害虫事件因果关系。

目前，国内外学者针对储粮害虫的检测、防治和预测等环节进行了各种研究，并取得显著进展。储粮害虫实时检测[4]是防控的重要前提，以深度学习[5]为基础的目标检测技术随之兴起。然而，某些难以识别的害虫却依然只能依赖经验丰富的专家进行人工识别。为了突破这种局限， AGARWAL 等[6]尝试结合可见近红外高光谱技术和深度学习技术帮助识别难以分辨的害虫。储粮害虫防治是防控的主要手段，一般有物理防治、化学防治、生物防治等方法[7]，中国基本采用以熏蒸剂、防护剂为主的化学防治方法。但是，由于长期使用化学杀虫剂，不仅使得害虫产生严重的抗药性，环境也受到了严重污染。面临这些挑战，物理防治方法因其环保性，易操作性，逐渐成为了储粮害虫防治体系中不可或缺的一部分。邓树华等[8]总结了惰性粉防治、气调防治等物理防治方法方面的技术成果，发现惰性粉防治防虫效果显著，而且绿色环保，操作简便成本低；气调防治在粮食储存与植物检疫方面发挥着关键作用，二者相辅相成。近年来，储粮害虫生物防治也逐渐受到重视，任剑豪等[9]给出了关于昆虫生长调节剂、信息素等生物防治的最新进展。当前，国家粮食和物资储备局提倡采用综合防治技术，强调在治理过程中要综合运用多种防治手段，以实现"绿

收稿日期：2024-08-10 修订日期：2024-11-01
基金项目：国家自然科学基金资助项目（62271191）；河南省重点研发专项（241111211100）
作者简介：肖乐，副教授，硕士生导师，研究方向为粮食信息化、知识图谱。Email：xiaole@haut.edu.cn

色、生态、经济、高效"的治理理念。进一步，为了及时有效的进行害虫防治，LIAN 等[10] 提出一种结合环境因子的元胞自动机（cellular automaton, CA）预测方法，能够有效预测储粮害虫的生长趋势。包成雷[11] 认为现有研究忽略了虫害爆发多因素的影响，无法从根本上解决虫害爆发问题，通过分析害虫的生物学特性、环境因素和历史数据，实现了集害虫检测、预测和预警为一体的测控系统。

基于以上研究分析，储粮害虫研究目前多集中于害虫检测和害虫防治等方面，并呈现绿色、高效、综合、精准测控等研究趋势。而这些研究均需要挖掘储粮害虫的生物学特性、行为模式以及与环境等诸多因素的内在联系。其中，深入探究害虫事件因果关系是有效防控的关键。因此，在害虫生长发育与条件、危害状况、传染途径等现有研究数据的基础上，本研究将深度挖掘储粮害虫事件发展脉络，并进一步量化事件之间的因果强度，为确保粮食安全提供科学的参考依据。

1　相关技术研究现状

1.1　因果强度计算

因果强度被用于衡量因果事件之间的相关程度。在储粮害虫领域，量化事件之间的因果强度有助于评估不同风险的影响程度，从而帮助人们制定精准的决策措施，预防粮食损失。早期的因果关系估计多基于统计方法实现，LUO 等[12] 采用数据驱动的方法，从大型语料库中获取因果事件句子对，并提出结合充分条件和必要条件来建模词级间因果强度。在此基础上，谭云等[13] 结合了同类原因之间的相似度量来补充原始因果强度，能够实现计算不同原因对同一结果的因果强度。然而，以往研究仅用一个单词来表示事件，存在一定的缺陷，SASAKI 等[14] 构造多词表达式扩展现有的事件表示方法，这一改进在事件因果关系估计问题上展现出了更优越的性能。为准确衡量时序节点间的因果强度，郝志峰等[15] 提出了一种基于信息熵的衡量标准，能够筛选强关系的节点进而形成完整的因果图。然而，以上基于统计的方法无法学习事件中丰富的语义知识，可能无法深入理解事件之间的因果机制。预训练语言模型因其强大的语义表示能力，在各项自然语言处理（NLP）任务中取得了显著的成果。LI 等[16] 将因果知识整合到预训练语言模型中，用于因果强度计算，实现了新的因果推理基准。目前，因果强度已在各项任务中得到广泛应用。JAIMINI 等[17] 提出了一种使用知识图谱链接预测来发现因果关系的方法-CausalDisco，并将知识图谱实体之间的因果强度关系作为权重嵌入到模型中，试验评估表明将因果权重的嵌入能显著提高算法的性能。在开放对话领域，FENG 等[18] 认为当前的评估指标在评估语法正确响应时无法与人类的判断保持一致，提出了一个名为 CausalScore 的指标，通过衡量历史对话和响应之间的因果强度来评估响应的相关性，并表明这种指标显著超越了现有的先进指标。在储粮害虫领域，量化因果强度具有重要的应用价值，但由于储粮害虫因果事件的多样性和复杂性，收集到的数据集难以覆盖所有情况，可能存在一定的偏差。本文将采用数据增强技术解决数据偏差问题，进一步指导模型学习害虫行为与环境因素之间的复杂因果关系。

1.2　数据增强

数据增强也称为数据扩增，旨在让有限的数据产生更多的价值。在实际任务中，确保模型获得较好的性能通常需要依赖丰富的数据资源，然而，高质量的数据需要耗费大量的成本。因此，在数据资源受限的情况下，可以通过数据增强技术来提高任务的质量。在粮食领域，杨森等[19] 提出基于近红外光谱和深度学习的数据增强方法，为大米品种检测方案提供了新思路。针对自然语言处理领域的文本分类任务，学者们探索了基于语义替换、噪声增强、样例生成等增强方法。STALIUNAITE 等[20] 通过同义词替换生成扰动输入来执行对抗性训练，避免模型过于依赖数据中表面特征或模式。TANG 等[21] 提出一种融合字词粒度噪声的数据增强方法，获得了大规模且高质量的数据集。近年来，基于文本生成的方法因其能够提高语义的多样性并可以与预训练语言模型结合使用的优势，得到了广泛关注。BAYER 等[22] 通过转换手段人工制造训练数据，改进分类器性能，在评估短文本和长文本任务时模型性能得到了显著提升。ZHOU 等[23] 提出重平衡方法，降低在原始数据集中具有特定概念的样本数量，增加少数类样本的数量，减轻了由于概念层面的虚假关联导致的模型偏差。随着研究的深入，反事实数据增强技术已被证明能够增强模型的鲁棒性，在泛化数据上取得了不错的效果。FEDER 等[24] 利用数据因果结构的指导，通过大语言模型生成新的样本，模拟对虚假特征的干扰以得到更强大的文本分类器。QIU 等[25] 认为使用反事实数据进行训练可能导致模型过度关注修改后的特征，采用结合对比学习的策略来促进全局特征对齐，证明了所提方法在泛化数据集上超过了现有方法。然而，这些方法多针对其特定任务而设计，难以满足本领域需求。本研究将根据储粮害虫事件数据，利用 LLM（large language model）实现基于生成方法的反事实数据文本，帮助模型学习到更加全面的领域因果知识，以实现储粮害虫事件因果推理任务。为筛选出最适配的数据优化路径，表 1 对 3 种数据增强方法进行了对比分析。

表 1　自然语言处理任务中的数据增强技术分析

Table 1　Analysis of data augmentation techniques in natural language processing tasks

相关研究 Related research	特点 Features	不足 Deficiency
基于语义替换的方法[20]	通过同义词替换、语义嵌入等技术对原始语料进行修改	缺乏上下文情况下可能改变句子的语义
基于噪声增强的方法[21]	通过随机删除、交换单词顺序等技术对为原始句子添加微弱噪声	可能失去原始句子的语义信息
基于文本生成的方法[22-23]	面向特定任务和文本特征实现新数据的生成	根据任务需求进行设计，灵活性强，生成文本的质量评估较难

2 反事实数据增强的事件因果强度计算方法

2.1 储粮害虫数据集建立及分析

本研究语料主要来源于储粮害虫相关科技文献（https://aismart.oversea.cnki.net/）、网络发布的储粮害虫事件（https://www.chinanews.com.cn/）以及粮食大辞典相关条目，获取的事件文本经数据清洗、事件标注等预处理后整理为储粮害虫事件语料库（stored grain pest events，SGPE），包含 4 560 条因果事件相关文本数据，涵盖害虫态势发展、害虫导致的损失事件、害虫防治事件 3 个粗粒度类别，表 2 展示了储粮害虫相关事件的类别以及样本示例。在对储粮害虫事件数据集进行随机抽样分析时，发现可能存在一定的数据偏差。在选择偏差方面，数据主要来源于文献资源和网络资源，其中，由于文献数据受限于时间采集、特定害虫种类研究等因素，覆盖范围相对较窄。这种样本不平衡分布的问题可能造成常见害虫事件被过度关注，稀有害虫事件的样本数量较少，进而导致模型在处理常见害虫事件时表现较好，而在处理稀有害虫事件时表现不佳。此外，由于储粮害虫事件语料库规模较小，在这种低资源数据情况下，无法充分模拟储粮害虫事件的多样性和复杂性，模型可能会过度拟合训练集中的有限样本。因此，针对储粮害虫事件数据集存在偏差以及语料库规模小导致模型过拟的问题，本研究通过数据增强技术，增加数据的多样性，避免模型对有限样本的依赖，提高模型处理储粮害虫事件的准确率和泛化能力。

表 2 储粮害虫事件示例

Table 2 Examples of stored grain pest events

序号 No.	事件 Events	地点 Locations	类型 Type
1	内蒙古地区外来虫源大量迁入和本地虫源叠加，三代黏虫、二代草地螟呈重发态势，草地贪夜蛾北迁扩散速度加快，主要农作物病虫害偏重发生。	内蒙古	态势
2	由于河北省某市近日以来连续阴雨，造成粮仓空气湿度上升，在 A 粮仓，工作人员于在仓中发现玉米象大量繁殖。	河北省	
3	安徽省田间小麦赤霉病菌源量较常年同期明显偏高，预计今年小麦赤霉病全省呈偏重至大发生态势。	安徽省	
4	谷蠹成虫将卵粒产在粮粒内部，从幼虫起一直在粮粒内部取食，发育至成虫时才钻出粮粒外部，这时被蛀食过的粮食几乎成为空壳，失去食用价值。	武汉市	损失
5	高温和潮湿环境下，小麦蛾的种群数量在短时间内急剧增加，导致粮储存损失增加。	台州市	
6	绿豆象 1 a 繁殖七代，世代重叠严重，虫蚀率高达 15%～20%，使豆类造成大量损失。	南京市	
7	玉米象对玉米的侵蚀和污染，造成玉米营养价值和卫生质量受到影响，总计经济损失达 500 万余元。	德州市	
8	做好种群控制工作，如春季控仓内处于低温环境，夏秋用灯光或性信息素结合陷阱诱杀等，可以有效减少印度谷蛾感染。	云南省	防治
9	磷化铝进行熏蒸处理后，粮库内印度谷螟和麦蛾数量显著下降，粮食的损坏率从 10% 降低到 2%。	成都市	
10	在储存小麦的仓库中释放瓢虫后，经过 6 个月的监测，麦蛾的密度减少了 80%，小麦的质量得到了显著改善。	湖南省	

2.2 反事实数据增强方法具体实现

针对储粮害虫事件语料库存在以上数据偏差而造成模型在泛化数据上表现不佳的问题，本文采用数据增强技术，提出了一种反事实数据增强的因果强度计算方法（CDA-ECS）。该方法一共分为 3 个阶段，第一阶段，将事件对中的前提句子输入检索器，得到与原句子风格相似，语义相反的前 k 个句子；第二阶段，本文利用检索到的句子设计了一种基于规则的提示模板，使得大语言模型可以根据样例生成符合要求的句子并改变原事件对句子的标签；第三阶段，将原训练实例与新生成的实例合并为新的语料库，一起训练预训练语言模型，使其真正学习到事件的因果特征，提高模型在泛化数据上推理的准确率，并进一步得到因果强度得分，方法图如图 1 所示。

2.2.1 储粮害虫事件检索

本研究采用密集段落检索器（dense passage retriever，DPR）[26] 从领域相关文档中匹配需要的句子。DPR 由 2 个独立的 BERT 编码器组成，段落编码器E_p将任意的文本段落映射到一个 d 维实数向量，并存入向量数据库（facebook AI similarity search，FAISS）中离线构建向量索引，用于之后的检索。问题编码器E_Q同样将输入问题（question）映射到 d 维向量，然后比较问题向量和所有的段落向量的相似性，选出前 k 个段落作为检索结果，本文使用点积计算向量之间的相似性：

$$\mathrm{sim}(\boldsymbol{q},\ \boldsymbol{p}) = E_Q(\boldsymbol{q})^{\mathrm{T}} E_P(\boldsymbol{p}) \tag{1}$$

式中 $\boldsymbol{q}$ 为输入问题；$\boldsymbol{p}$ 为文本段落；E_p为段落编码器；E_Q为问题编码器。

$$D(x) = (x_i, y_i) \tag{2}$$

式中$D(x)$为本研究的训练集，x_i为训练样本，包含一个前提句子和一个假设句子，y_i为前提句子和假设句子对应的类标号。对每一个样本x_i，能够从检索器 DPR 中得到$E(x)$。

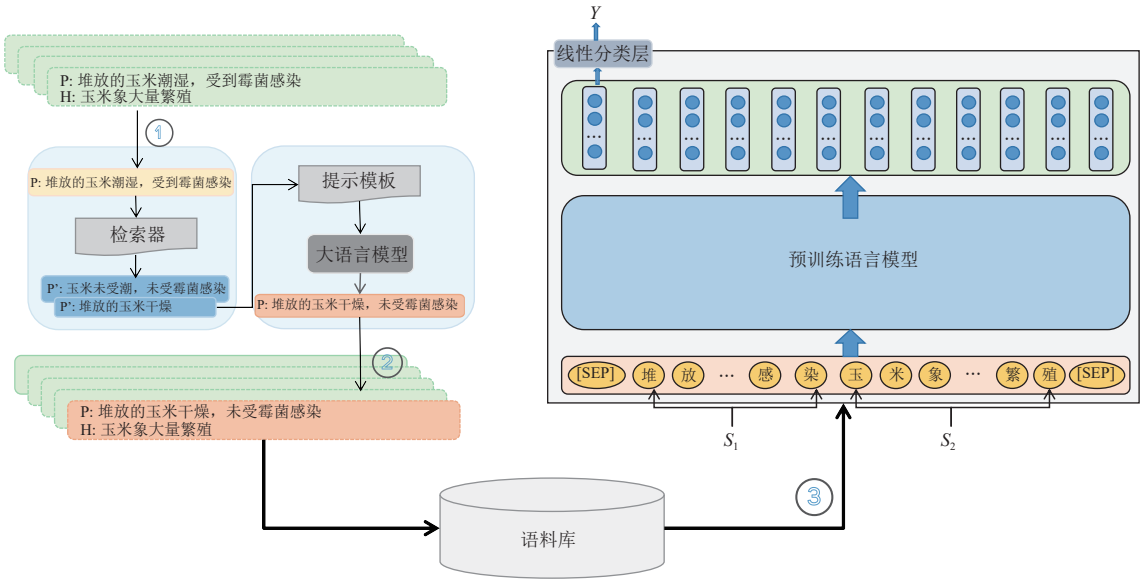
$$E(x) = (\hat{x}_1, \hat{x}_2, \cdots\cdots, \hat{x}_n) \tag{3}$$

式中$E(x)$为检索器 DPR 检索到的 n 个与原前提句子语义相反的句子集合；$\hat{x}_n$为句子集合中的第 n 个句子（n=1，2，……）。

为了使得 DPR 能够检索出更符合任务所需的句子，需要先对其进行训练，本文使用一个大小为 m 的种子数据集对 DPR 进行预训练。训练使用的超参数为：批大小为 16，最大序列长度为 64，权重衰减为 0.01，学习率为 0.000 02，训练轮数为 100。每个训练样本包含一个问题q_i，一个正样本p_i^+，以及 n 个负样本p_i^-。q_i是由连接符 [SEP] 分隔的前提句子和假设句子组成的，本文为每个样本的前提句子手动编写能使前提句子和假设句子标签改变的实例，将其作为正样本p_i^+。形式化为

$$C^+ = \{(q_1,\ p_1{}^+), (q_2,\ p_2{}^+), \cdots (q_m,\ p_m{}^+)\} \tag{4}$$

式中C^+为正样本集合；（q_m，p_m^+）为正样本集合中的第 m 个样本，m=1，2，……。



注：S_1 为输入句子对中的第一个句子，S_2 为输入句子对中的第二个句子，[SEP] 为句子之间的分隔符，P 表示前提句子，H 表示假设句子，Y 为两个句子的因果强度得分，下同。

Note：S_1 is the first sentence in the input sentence pair, S_2 is the second sentence in the input sentence pair, [SEP] is the separator between sentences, P is a premise sentence and H is a hypothesis sentence, Y is the causal strength score of the two sentences. the same below.

图 1　反事实数据增强的事件因果强度计算方法

Fig.1　Counterfactual data augmentation-event causal strength method

使用大语言模型（large language model，LLM）为每个训练样本的前提句子生成多个语义相似的不同句子，作为负样本，形式化为

$$C^- = \{(q_1，\ p_{1,1}^-,\cdots,p_{1,m}^-),\cdots,(q_n，\ p_{n,1}^-,\cdots,p_{n,m}^-)\} \quad (5)$$

式中C^-为负样本集合；（q_{nm}，p_{nm}^-）为负样本集合中的第 n 个训练样本的第 m 个负样本（$n,m=1,2,\cdots\cdots$）

本文的训练目标是使得正样本p_i^+在向量空间中和问题q_i的前提句子距离更近，而语义相似负样本$p_{i,\ j}^-$在向量空间中和问题q_i的前提句子距离更远[27]，损失函数为

$$L(q_i,p_i^+,p_{i,1}^-,\cdots,p_{i,n}^-) = -\ln\frac{e^{\mathrm{sim}(q_i,p_i^+)}}{e^{\mathrm{sim}(q_i,p_i^+)}+\sum_{j=1}^{n}e^{\mathrm{sim}(q_i,p_{i,j}^-)}} \quad (6)$$

2.2.2　储粮害虫反事实数据生成

基于上述 DPR 为每个训练实例检索到的样本，本研究设计了一种基于样例生成的提示模板，用于提示 LLM 为每个训练实例生成反事实实例，提示模板被形式化为

$$P = \{D，\ C，\ Q\} \quad (7)$$

式中 P 表示提示，D 表示一段任务描述，旨在为 LLM 明确需要解决的任务需求；C 表示为上下文信息，它为 LLM 提供了生成反事实的样例，这有助于模型更好地理解训练样本的反事实生成过程，并指导当前问题 Q 生成与任务相符的反事实。

与基于语义替换的增强方法类似[28]，样例生成使用规则和预训练模型来生成增强数据。然而，样例生成方法的独特之处在于其针对性，它专为特定任务量身定制，能够根据任务的具体需求进行调整。此外，这种方法在实施过程中需要依赖于标签信息和数据格式等与任务相关的细节。这样不仅可以确保增强数据的有效性，还可以提升数据的多样性。由于生成式预训练语言模型学习了大规模语料的知识，已被证明在样例生成中具有不错的表现[29]。因此，本文将 GPT-3.5-Turbo 用于基于样例生成的数据增强方法，在原始数据上进行扰动，生成新的可用数据，达到数据增强的目的。

具体来说，首先将从检索器中检索到的前 top k 个句子进行分词、去除停用词等预处理，并提取有助于改变标签的单词，形成一个提示词列表。进一步，本文在提示模板中设计反事实生成规则，要求 GPT-3.5-Turbo 对原始标签进行翻转，并且根据提示词列表对原始前提句子进行扰动，生成新的反事实样本。提示模板以及反事实数据增强效果如图 2 所示。

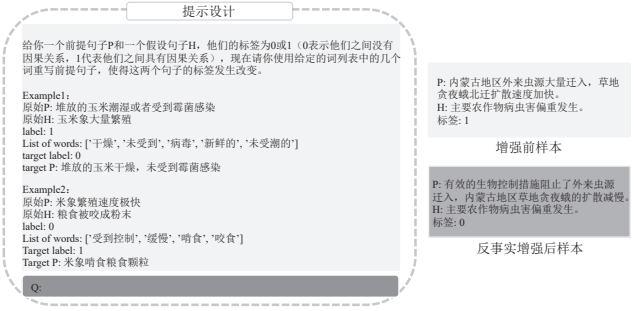


图 2　反事实生成提示模板及增强样本示例

Fig.2　Counterfactuals generate prompt templates and enhance sample examples

2.2.3　事件因果强度计算

通过反事实数据生成，得到了增强数据集D_{aug}，将

原始数据集D和增强数据集D_{aug}合并，形成新的训练数据集$\widetilde{D}$。使用增强后的领域因果知识对预训练语言模型ALBERT进行微调，使其不仅掌握文本的一般分布特性，还能学到领域事件之间因果背景，提升泛化数据上模型推理的准确率，进而提高储粮害虫事件图谱中事件之间的因果强度量化的可靠性。图3为储粮害虫事件图谱样例（部分）。

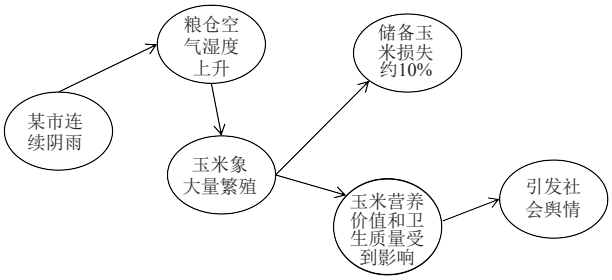


图3 储粮害虫事件图谱样例（部分）

Fig.3 Example of stored grain pest event graph (part)

本文将基于预训练语言模型的因果强度计算方法表示为$C=\mathrm{CPLM}(S_1, S_2)$，此方法分为两个阶段。在训练阶段，每条训练数据包含一个事件对及其因果关系标签，将两个事件文本拼接[30]后输入预训练语言模型ALBERT，拼接方法为：[SEP] 某市连续阴雨。粮仓空气湿度上升 [SEP]。

输入文本序列中的每个单词通过嵌入层被转换成一个上下文相关的向量表示，接着，编码后的上下文向量通过池化操作被聚合成一个固定长度的向量$\boldsymbol{H}$，全连接层通过权重和偏置对向量进行线性变换，以学习复杂的因果关系映射，其输出为因果强度得分。通过这个过程，模型能够评估输入事件对之间是否存在因果关系。

$$H_1, H_2 \ldots\ldots, H_n = \mathrm{PLM}(S_1, S_2) \tag{8}$$

式中PLM为预训练语言模型

$$H_n = \mathrm{Pooler}(H_1, H_2 \ldots\ldots, H_n) \tag{9}$$

式中Pooler为池化操作

$$C = \sigma(\boldsymbol{W}\boldsymbol{H} + b) \tag{10}$$

式中$\boldsymbol{W}$为权重矩阵，$\boldsymbol{H}$为输入向量，b为偏置项。

根据任务的目标，本文训练的是两个元素之间的因果相似关系，而非样本的类别得分。因此，本研究选择铰链损失函数（hinge loss），使得正样本的预测值大于某个阈值，而负样本的预测值小于某个阈值。

$$L = \max(0, \alpha + \mathrm{CPLM}(S_1^-, S_2^-) - \mathrm{CPLM}(S_1^+, S_2^+)) \tag{11}$$

式中L为损失函数；α为超参数，$\mathrm{CPLM}(S_1^-, S_2^-)$表示负样本因果强度得分，$\mathrm{CPLM}(S_1^+, S_2^+)$表示正样本因果强度得分。

在预测阶段，经预训练的模型能够对新输入的事件对进行处理，输出一个分数$y' \epsilon [0,1]$，为事件对的因果强度得分，将其作为事件图谱边上的权重。因此，能够得到带权重的储粮害虫事件图谱，w_1=0.35，w_2=0.31，w_3=0.19，w_4=0.13，w_5=0.21，如图4所示。

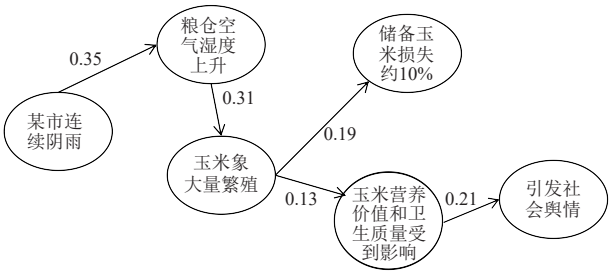


图4 带权重的储粮害虫事件图谱样例（部分）

Fig.4 Sample of stored grain pest event graph with weights (part)

2.3 训练平台及参数

模型训练的平台硬件配置为I7-10 700 CPU，NVIDIA RTX 3090Ti，软件环境为Win 10系统，64位，Python 3.8版本，Pytorch 1.7框架，Adam优化器，预训练语言模型ALBERT的训练参数如表3所示。

表3 训练参数

Table 3 Training parameters

训练参数 Training parameters	数值 Values
编码器句子长度 Encoder sentence length	64
权重衰减系数 Weight decay factor	0.01
学习率 Learning rate	0.000 02
批处理大小 Batch size	32
迭代周期数 Number of train epochs	20

2.4 试验结果与分析

在所有试验中，采取相同的试验设置，将本文方法（CDA-ECS）与朴素增强方法[24]（naive augmentations，NA）、条件增强方法[24]（conditional augmentations，CA）进行了对比。NA仅仅替换原始数据的某些关键词，不改变进句子语义，提示LLM创建新的数据，而CA使用简单的匹配策略，基于能够改变原始句子语义的特征词进行搜索，并提示LLM以匹配的词语改写原句子。

2.4.1 公共数据集试验结果

为了证明本文所提方法的适用性，本文在广泛使用且著名的常识因果推理基准测试集COPA[31]（choice of plausible alternative）上进行测试，该数据集包含了1 000个开放领域常识性因果问题。此外，本文还采用了在COPA基础上开发出的BCOPA-CE（balanced-copa test set with cause-effect）[32]测试集，该测试集包含无偏的标签分布，并增加了因果关系的难度，是一个更具挑战性的因果推理测试集。试验结果如表4所示。

由表4可知，NA方法在COPA测试集上取得了最好的表现，准确率达到81.5%，而本文方法CDA-ECS低于基线CA 1.0个百分点，表明本文方法在处理简单因果关系时不能凸显反事实数据的优势。而在面对更具挑战性的BCOPA-CE数据集时，所有模型的准确率普遍下降。与在COPA数据集上的表现相比，3种数据增强方法的准确率分别下降了25.8、21.4和19.9个百分点。但本文方法CDA-ECS相较于NA、CA方法分别高出4.9、1.4

个百分点，这突出了反事实数据多样性及复杂性在应对泛化数据时的重要性。

表 4　不同模型在公共数据集 COPA 和 BCOPA-CE 上的准确率

Table 4　Accuracy of different models on public datasets-choice of plausible alternative(COPA) and balanced-copa test set with cause-effect(BCOPA-CE)　%

方法 Method	COPA	BCOPA-CE
PMI	58.3	49.8
BERT	76.6	51.7
ALBERT	80.3	56.2
NA	81.5	55.7
CA	80.6	59.2
CDA-ECS(本研究)	80.5	60.6

注：PMI 为点互信息因果强度计算方法；BERT 与 ALBERT 为基于预训练语言模型的因果强度计算方法；NA 与 CA 为经数据增强后的预训练语言模型因果强度计算方法；CDA-ECS 为经本文所提方法增强后的预训练语言模型因果强度计算方法。

Note：PMI is a method for calculating the causal strength of point-mutual information; BERT and ALBERT are methods for calculating the causal strength of pre-trained language models;NA and CA for the calculation of causal strength of pre-trained language models augmented with data;CDA-ECS is a pre-trained language model causal strength calculation method augmented by the method proposed in this paper.

2.4.2　领域数据集试验结果

本文在储粮害虫领域数据集（stored grain pest events，SGPE）上进行试验，选用 SGPE 的一个子集训练 ALBERT[16] 模型，在两个测试集上进行评估。测试集 ID-SGPE（in distribution-stored grain pest events）与训练集具有相似的特征和分布。测试集 OOD-SGPE（out of distribution-stored grain pest events）是用于模型测试的泛化数据集，包含某个特定地区的数据，在模型训练阶段没有使用过，用于评估模型的泛化能力。表 5 显示了本文方法和和基线方法的对比结果。

表 5　不同模型在 ID-SGPE 数据集和 OOD-SGPE 数据集上的对比效果

Table 5　Comparison effects of different models on the in distribution- stored grain pest events dataset(ID-SGPE) and the out of distribution-stored grain pest events dataset(OOD-SGPE)　%

方法 Method	ID-SGPE	OOD-SGPE
PMI	68.3	61.9
BERT	75.2	67.1
ALBERT	80.5	72.2
NA	81.0	70.6
CA	80.3	73.2
CDA-ECS(本研究)	80.6	75.6

同样，选用与公共数据集相同的基线模型进行试验，由表 5 可知，在测试集 ID-SGPE 上，由于与训练数据具有良好的一致性，CDA-ECS 方法生成的反事实样本并未带来明显的性能提升，而 NA 方法由于简单的变换，生成的数据能够保持数据的原始分布，性能最好，达到了 81.0% 的准确率。然而，在面对模型未见过的泛化测试集 OOD-SGPE 时，CDA-ECS 方法由于增加了数据的多样性和复杂性，模型能够理解文本的深层结构，相较于基线 CA 方法准确率提高了 2.4 个百分点，达到 75.6%，能为储粮害虫事件因果强度计算带来帮助。因此，本方案更适用于模型应对泛化场景下的储粮害虫事件数据。

2.4.3　消融试验

为了理解每个组件的贡献度，本节通过消融试验深入分析了各个关键组件对模型性能的具体影响。分别对整个框架中检索部分和生成部分进行消融研究，并在不改变其他试验设置下观察结果的变化，试验结果如表 6 所示。

表 6　消融试验结果

Table 6　Results of ablation experiments　%

方法 Method	ID-SGPE	OOD-SGPE
DPR+ GPT-3.5-Turbo（all）	80.3	75.6
GPT-3.5 -Turbo	79.8	74.1
TF-IDF+GPT-3.5-Turbo	79.5	73.7

注：DPR 为密集段落检索器；TF-IDF 为传统的词频检索方法；GPT-3.5-Turbo 为本文选用的大型语言模型；（all）表示此种组合方式涉及了本文方法的所有组件。

Note: DPR is the dense paragraph retriever ; TF-IDF is the traditional word frequency retrieval method; GPT-3.5-Turbo is the large language model chosen in this paper; (all) denotes that such a combined approach involves all the components of the paper's method.

为了评估匹配示例在生成反事实实例扰动多样性中的作用，本文去除文档检索环节，即直接利用大型语言模型 GPT-3.5-Turbo 根据提示修改原始示例以生成反事实实例，这导致在泛化的测试集上的准确率下降了 1.5 个百分点。这是由于大模型在没有文档检索辅助的情况下，无法获取更多的相关信息，限制了生成反事实实例的质量和相关性。进一步，为了深入探究本文选用密集段落检索器 DPR 的性能优势，本研究采用一种基于 TF-IDF（term frequency-inverse document frequency）的传统检索方法替代 DPR 这种先进的双编码器检索模型，并进行试验。结果表明，无论在 ID-SGPE测试集还是在 OOD-SGPE 测试集上，采用这种传统检索方法后，其性能均出现了不同程度的下降情况。尤其是在后者这一领域泛化的测试集 OOD-SGPE 中，准确率相对于本文方法明显降低。这表明 DPR 在理解语义关系方面展现的优越性，而与之相对的是，TF-IDF 主要依赖关键词的频率及其重要性进行检索，在捕捉复杂语义信息上存在一定局限。因此，当整个过程缺少高效的检索机制，单纯依靠大语言模型来生成反事实实例时，往往会导致性能降低。同时，若是采用类似 TF-IDF 这种基础的检索器，那么其对于任务准确率所产生的负面影响会更加显著。

2.5　储粮害虫事件因果强度分析

本文对提出的因果强度计算方法进行性能测试分析。通过本文数据增强方法，引入了多样化的数据，模型将调整因果强度得分，以反映更广泛的数据分布。为验证本文方法的泛化性，选用来自某地区实地调研的储粮害虫事件数据，以测试模型应对不同地区样本的处理能力。表 7 所示为数据增强前后对应事件的因果强度结果，经人工对比分析后，表明模型在处理泛化数据的因果推理任务中展现出良好效果，这充分证实了本文方法在泛化层面的突出优势。

为更好地将事件间因果强度应用于下游任务，根据本研究方法，将数据增强后的预训练语言模型 ALBERT

输出结果提取出来，得到带权重的事件图谱如图 5 所示。通过考虑事件间的因果关联强度，能够准确反映现实世界的复杂性，从而有助于理解事件的演变趋势，帮助人们提前预防风险，减少损失。

表 7 因果强度结果

Table 7 Results of causal strength

事件句子对 Event sentence pair	因果强度增强前 Pre-causal strength enhancement	因果强度增强后 Aft-causal strength enhancement
S1：粮仓空气湿度上升 S2：玉米象大量繁殖	0.16	0.31
S1：内蒙古地区外来虫源大量迁入 S2：病虫呈重发态势	0.23	0.17
S1：玉米象对玉米侵蚀和污染 S2：玉米营养价值和卫生质量受到影响	0.28	0.35
S1：稻飞虱、稻纵卷叶螟等迁飞性害虫田间虫量激增 **S2:** 玉米粘虫局地暴发严重	0.22	0.16
S1：玉米象大量繁殖 S2：储备玉米损失约 10%	0.25	0.19
S1：谷蠹群体迅速繁殖扩散 S2：粮食库存中的损失增加	0.42	0.23

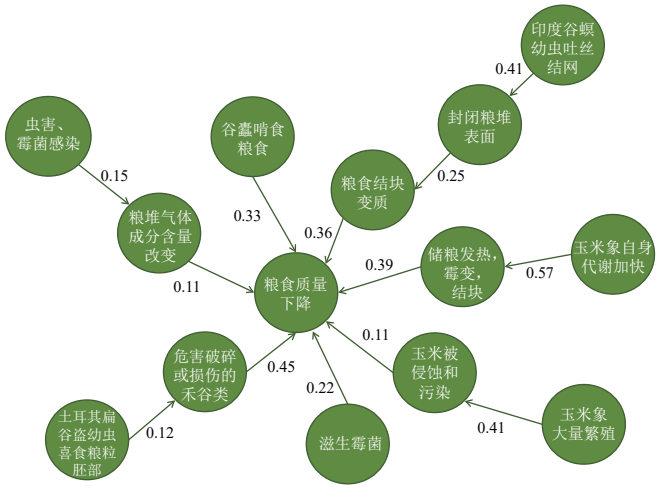


图 5 带权重的储粮害虫事件图谱

Fig.5 Stored grain pest event graph with weights

3 结 论

本文提出了反事实数据增强方法（counterfactual data augmentation-event causal strength，CDA-ECS），以储粮害虫事件为研究对象，用于实现储粮害虫事件的因果强度计算。针对数据偏差问题，对比不同的数据增强方法，为提升模型泛化能力提供优选增强方法，主要结论如下：

1）通过对原始语料数据进行基于语义的匹配，利用 LLM（large language model）进行反事实编辑实现新样例的生成，模型在泛化数据的因果推理任务上相较于基线 CA（conditional augmentations）方法准确率上升 2.4 个百分点。表明确保反事实数据与原样本语言一致性以及扰动多样性有效提高了模型应对复杂数据的泛化能力，是计算储粮害虫事件因果强度的优选方法。

2）储粮害虫事件经因果强度量化后，被用于完善事件图谱。通过这种方式，能够有效处理和分析多个因果事件交互的复杂情况，从而为粮食领域复杂因果分析提供重要支持。

本研究方法可用于金融风险控制，通过量化金融风险因素间的因果关系，以提升风控水平；也可用于医疗流行病预防领域，通过量化疾病发生和传播因素的因果强度，为公共卫生决策提供支持。目前，在 LLM 生成反事实数据过程中缺乏人工干预，可能导致生成的数据在标签翻转时出现不准确情况，从而影响模型对因果关系的理解和量化。未来，将进一步设计反事实数据生成优化方法，提高反事实数据生成质量，为量化事件因果强度提供可靠依据。

[参 考 文 献]

[1] 孙晟源，王康旭，陈二虎，等. 转录组学在储粮害虫研究中的进展 [J]. 中国粮油学报，2024，39(4)：1-8.
SUN Shengyuan, WANG Kangxu, CHEN Erhu, et al. Advances in transcriptomics in stored-product pests [J]. Journal of the Chinese Cereals and Oils Association, 2024, 39(4): 1-8. (in Chinese with English abstract)

[2] 肖乐，李家馨，葛亮，等. 面向粮情决策支持的知识图谱构建研究 [J]. 中国粮油学报，2022，37(10)：29-37.
XIAO Le, LI Jiaxin, GE Liang, et al. Knowledge graph construction for decision support of grain situation [J]. Journal of the Chinese Cereals and Oils Association, 2022, 37(10): 29-37. (in Chinese with English abstract)

[3] 肖乐，陈啸林，单昕. 面向储粮害虫的事理图谱构建研究[J]. 中国粮油学报，2023，38(10)：185-195.
XIAO Le, CHEN Xiaolin, SHAN xin. Construction of stored grain pest event evolutionary graph[J]. Journal of the Chinese Cereals and Oils Association, 2023, 38(10): 185-195. (in Chinese with English abstract)

[4] 张硕，韩少云，熊黎剑，等. 基于气敏传感器阵列特征优化的储粮害虫赤拟谷盗检测[J]. 农业工程学报，2022，38(10)：303-309.
ZHANG Shuo, HAN Shaoyun, XIONG Lijian, et al. Detection of stored grain pests Tribolium castaneum (Herbst) based on the feature optimization of gas sen-sor array[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(10): 303-309. (in Chinese with English abstract)

[5] 郭嘉璇，王蓉芳，南江华，等. 融入全局相应归一化注意力机制的 YOLOv5 农作物害虫识别模型[J]. 农业工程学报，2024，40(8)：159-170.
GUO Jiaxuan, WANG Rongfang, NAN Jianghua, et al. YOLOv5 model integrated with GRN attention mechanism for insect pest recognition[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(8): 159-170. (in Chinese with English abstract)

[6] AGARWAL M, AL-SHUWAILI T, NUGALIYADDE A, et al. Identification and diagnosis of whole body and fragments of Trogoderma granarium and Trogoderma variabile using visible near infrared hyperspectral imaging technique coupled with deep learning[J]. Comput Electron Agric, 2020, 173: 105438.

[7] 国家粮食局，国家粮食储备局成都粮食储藏科学研究所. 粮油储藏技术规范：GB/T 29890-2013 [S]. 北京: 中国标准出版社.2013.

[8] 邓树华，吴树会，潘琴，等. 储粮害虫物理防治技术研究[J]. 粮食与油脂，2019，32(1)：10-12.
DENG Shuhua, WU Shuhui, PAN Qin, et al. Study on the physical control technologies of stored grain pest[J]. Cereals & Oils, 2019, 32(1): 10-12. (in Chinese with English abstract)

[9] 任剑豪，吴卫国，宗平，等. 储粮害虫生物防治技术研究进展[J]. 粮油食品科技，2020，28(6)：218-222.
REN Jianhao, WU Weiguo, ZONG ping, et al. Research progress on bio-control technology of stored-grain pests[J]. Science and Technology of Cereals, Oils and Foods, 2020, 28(6): 218-222. (in Chinese with English abstract)

[10] LIAN F, GE H, JIANG Y. A prediction model for stored grain pests based on cellular automaton [C]// 2015 International Conference on Computational Intelligence and Communication Networks (CICN).Jabalpur, India:IEEE,2015:1325-1330.

[11] 包成雷. 基于储粮害虫预测的粮库测控系统研究 [D]. 杭州：浙江大学，2020.
BAO Chenglei. Research on Grain Storage Measurement and Control System Based on Grain Storage Pest Prediction [D]. Hangzhou: Zhejiang University, 2020. (in Chinese with English abstract)

[12] LUO Z, SHA Y, ZHU K, et al. Commonsense causal reasoning between short texts[C]// Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning. Cape Town, South Africa: AAAI, 2016: 421-430.

[13] 谭云，彭海阔，秦姣华，等. 基于权重计算的中文因果关系分析[J]. 华中科技大学学报（自然科学版），2022，50(2)：112-117.
TAN Yun, PENG Haikuo, QIN Jiaohua, et al. Chinese causality analysis based on weight calculation[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2022, 50(2): 112-117. (in Chinese with English abstract)

[14] SASAKI S, TAKASE S, INOUE N, et al. Handling Multiword Expressions in Causality Estimation[C]//Proceedings of the 12th International Conference on Computational Semantics (IWCS). Montpellier, France: Acl, 2017: 298-307.

[15] 郝志峰，谢蔚涛，蔡瑞初，等. 基于因果强度的时序因果关系发现算法[J]. 计算机工程与设计，2017，38(1)：132-137.
HAO Zhifeng , XIE Weitao , CAI Ruichu , at al. Causal inference on time series using causal strength [J]. Computer Engineering and Design, 2017, 38(1): 132-137. (in Chinese with English abstract)

[16] LI Z. DING X. LIAO K, et al. CausalBERT: Injecting causal knowledge into pre-trained models with minimal supervision. [EB/OL]. [2024-08-23]. https://doi.org/10.48550/arXiv.2107.09852.

[17] JAIMINI U, HENSON C A, SHETH A P.CausalDisco: Causal discovery using knowledge graph link prediction [EB/OL]. [2024-08-23]. https://arxiv.org/pdf/2405.02327v1.

[18] FENG T, QU L, KANG X, et al. CausalScore: An automatic reference-free metric for assessing response relevance in open-domain dialogue systems [EB/OL]. [2024-08-26]. https://doi.org/10.48550/arXiv.2406.17300.

[19] 杨森，张新昊，王振民，等. 基于近红外光谱和深度学习数据增强的大米品种检测[J]. 农业工程学报，2023，39(19)：250-257.
YANG Sen, ZHANG Xin'ao, WANG Zhenmin, et al. Rice variety detection based on near-infrared spectroscopy and deep learning data augmentation[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2023, 39(19): 250-257. (in Chinese with English abstract)

[20] STALIUNAITE I, GORINSKI P J, IACOBACCI I. Improving commonsense causal reasoning by adversarial training and data augmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, California, USA: AAAI 2021, 35(15): 13834-13842.

[21] TANG Z , JI Y, ZHAO Y, et al. Chinese grammatical error correction enhanced by data augmentation from word and character levels[C]//Proceedings of the 20th Chinese National Conference on Computational Linguistics. Hohhot, China: CIPSC, 2021: 13-15.

[22] BAYER M, KAUFHOLD M A, BUCHHOLD B, et al. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers[J]. International Journal of Machine Learning and Cybernetics, 2021, 14: 135-150.

[23] ZHOU Y, XU P, LIU X, et al. Explore spurious correlations at the concept level in language models for text classification[EB/OL]. [2024-08-23]. https://doi.org/10.48550/arXiv.2311.08648.

[24] FEDER A, WALD Y, SHI C, et al. Data augmentations for improved (large) language model generalization [J]. Advances in Neural Information Processing Systems, 2023, 36:70638-70653

[25] QIU X, WANG Y, GUO X, et al. PairCFR: Enhancing model training on paired counterfactually augmented data through contrastive learning[C] // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand: ACL, 2024: 11955-11971.

[26] KARPUKIHIN V, OGUZ B, MIN S, et al. Dense passage retrieval for open-domain question answering [EB/OL]. [2024-08-23]. https://doi.org/10.48550/arXiv.2004.04906.

[27] IZACARD G, CARON M, HOSSEINI L, et al. Unsupervised dense information retrieval with contrastive learning[EB/OL]. [2024-08-29]. https://doi.org/10.48550/arXiv.2112.09118.

[28] 利建鑫，任江涛. 一种基于句子语义替换的电子病历文本数据增强方法：CN112836047B[P]. 2022-05-27.

[29] NG N, CHO K, GHASSEMI M. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.Punta Cana, Dominican Republic:ACL, 2020:1268-1283.

[30] KAVUMBA P, INOUE N, HEINZERLING B, et al. Balanced COPA: Countering superficial cues in causal reasoning[C]// Proceedings of the 26th Annual Meeting of the Association for Natural Language Processing. Tokyo, Japan: ANLP, 2020:

1105-1108.
[31] ROEMMELE M, BEJAN C A, GORDON A S. Choice of plausible alternatives: An evaluation of commonsense causal reasoning[C]//Proceedings of the 2011 AAAI Spring Symposium Series. California, USA: AAAI: 2011: 394-398.
[32] HOSSEINI P, BRONIATOWAKI D A, DIAB M. Knowledge-augmented language models for cause-effect relation classification[C]// Proceedings of the First Workshop on Common Sense Representation and Reasoning (CSRR 2022). Dublin, Ireland: ACL, 2022: 43-48.

Calculation of the causal strength of stored grain pest events augmented using counterfactual data method

XIAO Le , ZHAO Jing , XU Yunfei

(*College of Information Science and Engineering*, *Henan University of Technology*, *Zhengzhou* 450000, *China*)

Abstract: Stored grain pests have been one of the most important influencing factors on food security in recent years. It is extremely critical to explore the grain storage pest events and their causal relationships. Furthermore, the causal strength among grain storage pest events can be expected to more accurately assess the potential risks, in order to formulate the preventive and control measures. However, the data bias in the grain storage pest domain can often rely overly on the surface features in the dataset, leading to low efficiency with generalized data. In this study, the causal strength among events was accurately computed and quantified using counterfactually augmented data. As such, the counterfactual data augmentation-event causal strength computation framework (CDA-ECS) was designed to generate the counterfactual instances using a large language model (LLM). The original data was then extended to integrate the debiased causal knowledge into the pre-trained language model. The causal relationships of sentences were learned more deeply to improve the generalization of the model. Specifically, three stages were divided: In the first stage, the premise sentences in the event pairs were inputted into a retriever to obtain the top k sentences that were similar in style and opposite in semantics to the original sentences; In the second stage, a rule-based cueing template was designed using the retrieved sentences. The large language model was utilized to generate the compliant sentences, and then adjust the labels of the original event pair sentences using the samples; In the third stage, the original training and the newly generated instances were merged into a new corpus to train together the pre-trained language model. The causal features of the events were learned to improve the accuracy of the reasoning on the generalized data, in order to obtain the causal strength score. Experiments on the public and domain datasets demonstrated that the more robust models were trained with 2.4 percentage points higher accuracy on the inference task on generalized data, which was effectively applied to calculate the causal intensity of grain storage pest events. The counterfactual data augmentation was introduced to represent the data bias in the field of grain storage pests. The diversity and complexity of the augmented data were utilized to more deeply understand the complex links among pest behavior and environmental factors, in order to achieve the risk analysis of grain storage pest events. Nevertheless, it was still lacking human intervention in the process of counterfactual data using LLM, particularly when the labels were flipped. The quantification of causal relationships can also be expected to improve in the future. The counterfactual data generation can be optimized to further improve the quality of counterfactual data generation. The finding can provide a reliable basis to quantify the causal intensity of events. In conclusion, an effective solution can be proposed to improve the performance of causal analysis models in the field of grain storage pests. It is also expected to serve as the more accurate decision-making in risk assessment and management.

Keywords: stored grain pests; data augmentation; large language model; counterfactual generation; causal strength