摘　要

机器学习中的虚假因果关联性研究

机器学习模型已经被广泛应用于解决社会问题以及进行高风险的决策。在许多的真实应用场景中，不仅要求机器学习模型具有较高准确率，还要求其具备良好的鲁棒性和安全性。然而，对抗攻击等技术的出现引起了人们对机器学习模型的安全性和鲁棒性的担忧。为了改善机器学习模型的安全性、鲁棒性等各方面表现，已经有许多研究做出了卓越贡献。然而，现有的研究较少从虚假的因果关联性的角度切入。其中，因果关联性指一个事件的发生或存在会导致另一个事件的发生或存在，虚假的因果关联性则指不是一个事件直接影响了另一个事件。虚假的因果关联性可能被恶意攻击者利用，通过针对模型的输入进行精心设计的扰动来欺骗模型。因此，虚假因果关联性与机器学习模型的安全性、鲁棒性等息息相关。为了更好地改善上述问题，本文从验证虚假因果关联性存在，说明其对机器学习模型的负面影响，以及分析虚假因果关联性特点，并总结规避虚假因果关联性的方法两方面出发，主要进行以下两方面工作：

1. 基于结构因果图，分析虚假因果关联性对线性回归模型、支持向量机回归模型、贝叶斯岭回归模型的影响。首先，基于小样本数据集 AutoMPG 与 MOP，通过虚假因果关联性分析方法，分析特征间是否存在虚假因果关联性。然后设计消融实验，研究虚假因果关联性对模型性能的影响，并分析出现此影响的原因。最后，通过比较模型在不同模拟场景下的表现，说明虚假因果关联性对模型鲁棒性、安全性的影响。实验结果表明，在小样本数据集 AutoMPG 与 MOP 的各个特征之间，的确存在虚假因果关联性。剔除此类特征虽然不会直接改善模型的性能，但是可以改善模型的安全性、鲁棒性。

2. 基于潜在结果框架，分析虚假因果关联性对 Bert、WordCNN 等深度学习模型在文本分类任务中的影响。首先，使用对抗文本生成算法生成对抗文本，验证对抗文本中是否存在虚假的因果关联性，同时分析了文本的内容、结构特点对生成对抗文本的影响。之后，比较在小规模原始数据、大规模原始数据、不同对抗样本中

训练或微调模型后，模型的性能、安全性和鲁棒性变化，研究虚假因果关联性对模型性能、安全性和鲁棒性的影响，并分析原因。随后，对对抗样本在模型特征空间中的分布情况进行可视化，从几何角度观察对抗样本与原始样本在模型中的聚类变化情况。最后，比较 WordCNN 模型在有无注意力层情况下进行分类任务的性能、安全性与鲁棒性，说明注意力机制对虚假因果关联性的影响。实验结果表明：对抗样本中的确存在虚假因果关联性，且在生成文本数据对抗样本时，文本数据的特征不是影响对抗样本生成的主要因素。通过比对模型在不同数据集上进行训练和微调后的实验结果，可以发现对抗样本中的虚假因果关联性会损害模型的性能、安全性与鲁棒性，同时，对抗样本可能导致模型特征空间中的聚类情况发生变化。具体到文本数据中，虚假因果关联性主要由一词多义现象引起。最后，使用注意力机制可以改善模型在对抗样本中的安全性与鲁棒性，但应该针对不同数据集与任务的特点进行调整。

关键词：

机器学习，深度学习，因果推断，对抗样本，自然语言处理

Abstract

Research on Spurious Causal Relationships in Machine Learning

Machine learning models have been widely applied to address societal issues and make high-stakes decisions. In many real-world scenarios, they are required to exhibit not only high accuracy but also robustness and security. However, the advent of techniques like adversarial attacks has sparked concerns about the security and robustness of machine learning models. To enhance various facets of machine learning models, including security and robustness, substantial contributions have been made through numerous research efforts. Nevertheless, existing research often lacks an examination from the perspective of spurious causal relationships. Causal relationships denote events leading to other events' occurrence or existence, whereas spurious causal relationships indicate that one event does not directly impact another event. Malicious actors may exploit spurious causal relationships by meticulously crafting perturbations on model inputs to deceive the model. Thus, spurious causal relationships are intricately linked with the security and robustness of machine learning models. To address these concerns comprehensively, this paper focuses on verifying the existence of spurious causal relationships, elucidating their adverse influence on machine learning models, analyzing the attributes of spurious causal relationships, and presenting methods to mitigate them. The paper primarily undertakes the following two tasks:

1. Using a structural causal graph, analyze the influence of spurious causal relationships on linear regression models, support vector machine regression models, and Bayesian ridge regression models. Firstly, based on small sample datasets AutoMPG and MOP, the presence of spurious causal relationships among features is analyzed through methods focusing on spurious causal relationships. Subsequently, degradation

experiments are designed to investigate the impact of spurious causal relationships on model performance and to analyze the reasons behind this impact. Finally, by comparing model performance under different simulation scenarios, the effects of spurious causal relationships on model robustness and security are demonstrated. Experimental results indicate that spurious causal relationships indeed exist among various features of the small sample datasets AutoMPG and MOP. Although removing such features may not directly enhance model performance, it can improve model security and robustness.

2. Based on a potential outcomes framework, analyze the impact of spurious causal relationships on deep learning models such as Bert and WordCNN in text classification tasks. Firstly, adversarial text generation algorithms are employed to generate adversarial texts, verifying the presence of spurious causal relationships in adversarial texts. Furthermore, the impact of text content and structural characteristics on the generation of adversarial texts is analyzed. Subsequently, the performance and robustness changes of models are compared on small-scale original data, large-scale original data, and different adversarial samples after training or fine-tuning, in order to study the effects of spurious causal relationships on model performance and robustness, and to analyze the reasons behind these effects. Then, the distribution of adversarial samples in the model's feature space is visualized to observe changes in the clustering of adversarial samples and original samples from a geometrical perspective. Finally, the performance and robustness of the WordCNN model are compared with and without an attention layer in classification tasks, elucidating the impact of attention mechanisms on spurious causal relationships. Experimental results demonstrate the existence of spurious causal relationships in adversarial samples and reveal that text features are not the primary factors influencing the generation of adversarial samples. By comparing experimental results of model training and fine-tuning on different datasets, it is evident that spurious causal relationships in adversarial samples can harm model performance, security, and robustness. Moreover, adversarial samples can cause changes in the clustering of features in the model's feature space. Specifically in text data, spurious causal relationships are

mainly caused by polysemy. Finally, the use of attention mechanisms can enhance the security and robustness of models in adversarial samples, though adjustments should be made according to the characteristics of different datasets and tasks.

Key words:

Machine learning, deep learning, causal inference, adversarial samples, natural language processing.

目 录

关于学位论文使用授权的声明

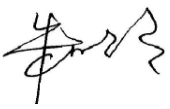
本人完全了解吉林大学有关保留、使用学位论文的规定，同意吉林大学保留或向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权吉林大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文。

（保密论文在解密后应遵守此规定）

论文级别：☑硕士 □博士

学科专业： 软件工程

论文题目： 机器学习中的虚假因果关联性研究

作者签名：  指导教师签名： 

2023 年 9 月 1 日

第 1 章 绪论

1.1 课题背景与研究意义

机器学习模型（如线性回归、支持向量机、贝叶斯岭回归、神经网络等）已经被广泛应用于解决社会问题以及进行高风险的决策[1][2][3]，在许多的真实应用场景中，不仅要求机器学习模型具有较高准确率，还要求其具备良好的鲁棒性和安全性。其中，鲁棒性指模型在面对各种数据变化和不确定性时的表现，安全性关注的是模型在面对各种恶意攻击和威胁时的表现。

为了改善机器学习模型的各方面表现，更好地服务各行各业的应用，已经有大量的研究为实现这一目标做出了卓越贡献[4]。然而，少有研究从虚假的因果关联性角度出发，解决机器学习模型性能、鲁棒性、安全性等方面的问题。因果关联性指的是两个或多个事件、现象之间存在的一种因果联系，其中一个事件（因）的发生或存在会导致另一个事件（果）的发生或存在。虚假的因果关联性正与此相反，即一个事件的发生可能是由于其他因素引起的事件同时发生，并非是当前事件作用的结果[5][6]。

下面这个例子可以更通俗地解释何谓“虚假的因果关联性”。图 1.1 展示了一个电影推荐场景中的虚假因果关联性的例子[7]。在这个场景中，汤姆看了很多英文电影，因此推荐系统倾向于推断汤姆喜欢语言为英文的电影，并继续推荐更多英文电影。然而，汤姆看这些英文电影是因为他喜欢那些科幻和冒险类型的故事，影片语言并不是这些汤姆选择电影时的主要原因。在这个例子中，汤姆观看的影片语言与推荐结果之间的关联性是虚假的因果关联性。

图 1.1 本例展示了推荐系统中的虚假因果关联性

通过在机器学习模型训练阶段，分析训练数据集中的虚假因果关联关系，并剔除与目标任务间存在虚假因果关联性的信息，或进行数据增强，吸收不同分布情况的数据[8]，学习其他数据中的因果关联性，可以减轻虚假因果关联性对模型的危害。如忽视虚假因果关联性的存在，则可能影响模型的性能、安全性或鲁棒性。以图像识别领域的机器学习模型为例，Bengio[9]等人发现机器学习模型经常通过学习图像背景中的线索来识别前景物体，如模型通过学习照片的背景中有很多海鸥，从而推断前景物体是一片海滩。表面上看，机器学习模型完成了任务，识别出了目标物体“海滩”，但这样的模型是脆弱的，当输入数据中不再存在海鸥时，模型将再也无法识别出海滩。同时，这一例子也说明了虚假因果关联性时常是隐晦的，需要对数据进行仔细地研究与分析，才能识别其中的虚假因果关联性。

因此，针对虚假因果关联关系的隐晦性，以及对机器学习模型性能、安全性、鲁棒性等方面的危害，本文认为，应该针对不同领域任务中的不同数据集，验证其中是否存在诱导机器学习模型的虚假因果关联性，具体分析各类数据中的虚假因果关联性将以何种方式何种程度损害机器学习模型，并总结虚假因果关联性的特点，进而帮助广大研究人员在训练模型时更好地规避数据中的虚假因果关联性，改善模型的表现。

综上所述，通过研究机器学习中的虚假因果关联性，可以更好地避免模型受到数据的诱导，学习到错误的规则，从而提升模型的性能、安全性或鲁棒性等，具有必要性与现实意义。

1.2 虚假因果关联性研究现状

1.2.1 小样本学习中的虚假因果关联性研究

根据霍夫丁不等式[10]，要想有效地训练机器学习模型，则应采用使经验误差最小化的原则。这种技术被称为经验风险最小化，其前提是假设空间要足够大，这依赖于大规模的训练数据。但对很多领域，仍然缺乏足量数据以训练表现优异的机器

学习模型。为改善机器学习模型在这些领域内的表现，近年来，小样本学习成为了一个研究热点[11][12]。

传统的小样本学习工作主要包括模型微调，数据增强等[13]。上述工作虽然可以在一定程度上改善机器学习模型的性能、鲁棒性和安全性。然而，仅通过学习数据属性间的相关性来拟合函数，仍然无法规避虚假因果关联性对模型的诸多负面影响。

为了进一步提升模型的综合表现，已经有一些研究基于虚假因果关联性领域的方法或理论改善机器学习模型的性能、安全性和鲁棒性[14][15]。例如，Zhongqi Yue[16]等基于因果图模型揭示了在当前的少样本学习方法中，缺乏预训练知识是限制模型性能的关键因素。并提出了一种新的 FSL 范式：干预式少样本学习（Interventional Few-Shot Learning，IFSL）。该方法与现有的基于微调和元学习的 FSL 方法是正交的，并在多个领域上达到了最先进水平。Xiao Liu[17]等提出了一种新颖的基于图的因果推断（Graph-based Causal Inference，GCI）框架，该框架使得机器学习模型在法律领域的指控消歧任务上取得了更好的表现，模型可以捕捉到多个混淆的指控之间的细微差别，并提供可解释的区分性，特别是在小样本情境中。WenWen Qiang[18]等提出了跨粒度少样本学习（Cross-Granularity Few-Shot Learning，CG-FSL），指的是在训练阶段可用足够数量的粗粒度类别样本，但在测试阶段的目标是对细粒度子类进行分类。他们基于细粒度图像数据集构建了基准测试集，通过比对基于因果图的 CG-FSL 模型和普通的 CG-FSL 模型、标准 FSL 方法发现，基于因果图的 CG-FSL 模型取得了最优的成绩。Takeshi Teshima[19]等首次基于结构因果模型中的结构方程，提出了一种新的小样本有监督领域自适应问题方法，在理论和实验两方面都验证了有效性。

此外，也有许多研究从数据分布角度出发，通过改变数据分布，降低虚假因果关联性对模型的负面影响。Xavier Garcia[20]等仅用 5 个高质量翻译数据，就让一个仅使用自监督学习训练的 Transformer 解码器模型能够与专门的监督学习最先进模型以及更通用的商业翻译系统相匹配，并且该模型具有扩展到多语言环境上的潜力。Yasaman Razeghi[21]研究了预训练模型在测试实例上的性能与这些实例的术语在预训练数据中的频率之间的相关性。具体地，他们测量了基于 GPT 的多个语言

模型在各种数值推理任务上的相关性强度，结果表明，模型对那些术语更常见的实例更准确，在某些情况下，与频率最低的 10％相比，模型对于频率最高的 10％术语的准确性提高了 70％以上。Xi Victoria Lin[22]等基于多语言数据扩充了 GPT-3 模型的训练语料，在 FLORES-101 机器翻译基准测试中，他们的模型在超过 182 个方向中的 171 个方向上超过 GPT-3，并在 45 个方向上超过了官方的有监督基线。

前人的研究已经在改善机器学习模型的性能、安全性和鲁棒性方面取得了优异的表现。然而，少有研究结合各领域的垂类知识，统计和分析小样本数据集诱导模型，损害其性能、安全性或鲁棒性的真正原因。

1.2.2 对抗攻击中的虚假因果关联性研究

对抗攻击技术的出现对深度学习模型的安全性和鲁棒性带来严峻的挑战。与对抗攻击相关的工作包括，Szegedy[23]等人首次提出在数字图像中加入轻微扰动，就能够误导神经网络模型做出错误的分类。受 Szegedy 等人的启发，Goodfellow[24]等人发现在训练过程中加入对抗性样本，能够提升模型的安全性、鲁棒性。为提升对抗样本的计算效率，JSMA 算法通过计算 Jacobian 矩阵来度量影响模型输入结果的关键像素，并迭代的对其进行修改，以达到误导模型分类的目的[25]。Papernot 等人提出通过 Substitute model 的方式生成对抗样本。Zhao 等人提出使用生成对抗网络（Generative Adversarial Network， GAN）制造对抗样本[26]。

对抗攻击技术揭示了现有神经网络技术在安全性和鲁棒性方面仍存在缺陷。尽管大量的研究工作试图揭示对抗攻击技术产生效果的原因，如设计方法来发现输入数据中的对抗样本，或通过对抗性训练来加强神经网络模型本身的鲁棒性和安全性[27][28]。但不幸的是，当前还未有明确的结论来解释对抗攻击产生效果的原因。

为了提高对抗样本的质量，改善机器学习模型的可解释性，已经有部分研究者结合虚假因果关联性领域的理论与方法改进已有的对抗样本领域研究。Chao-Han Huck Yang[29]提出了一个基于 do 算子的视觉推理因果关联性分析框架。为了研究虚假因果关联性中像素级特征的干预效果，他们引入了像素级遮蔽和对抗扰动，使用潜在空间中的特征和基于 DNN 模型的扰动预测来计算因果关联性，并进一步研究