计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.20210780

因果机器学习的前沿进展综述

李家宁[1,2] 熊睿彬[1,2] 兰艳艳[3] 庞亮[4] 郭嘉丰[1,2] 程学旗[1,2]

[1]（中国科学院网络数据科学与技术重点实验室（中国科学院计算技术研究所） 北京 100190）

[2]（中国科学院大学 北京 100049）

[3]（清华大学智能产业研究院 北京 100086）

[4]（中国科学院计算技术研究所数据智能系统研究中心 北京 100190）

（lijianing@ict.ac.cn）

Overview of the Frontier Progress of Causal Machine Learning

Li Jianing[1,2], Xiong Ruibin[1,2], Lan Yanyan[3], Pang Liang[4], Guo Jiafeng[1,2] and Cheng Xueqi[1,2]

[1]（*CAS Key Laboratory of Network Data Science and Technology（Institute of Computing Technology, Chinese Academy of Sciences）, Beijing 100190*）

[2]（*University of Chinese Academy of Sciences, Beijing 100049*）

[3]（*Institute for AI Industry Research, Tsinghua University, Beijing 100086*）

[4]（*Data Intelligence System Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190*）

Abstract Machine learning is one of the important technical means to realize artificial intelligence, and it has important applications in the fields of computer vision, natural language processing, search engines and recommendation systems. Existing machine learning methods often focus on the correlations in the data and ignore the causality. With the increase in application requirements, their drawbacks have gradually begun to appear, facing a series of urgent problems in terms of interpretability, transferability, robustness, and fairness. In order to solve these problems, researchers have begun to re-examine the necessity of modeling causal relationship, and related methods have become one of the recent research hotspots. This paper organizes and summarizes the work of applying causal techniques and ideas to solve practical problems in the field of machine learning in recent years, and sorts out the development venation of this emerging research direction. First, we briefly introduce the closely related causal theory to machine learning. Then, we classify and introduce each work based on the needs of different problems in machine learning, explain their differences and connections from the perspective of solution ideas and technical means. Finally, we summarize the current situation of causal machine learning, and make predictions and prospects for future development trends.

Key words causal relationship; spurious correlation; causal inference; machine learning; deep learning; artificial intelligence

摘要 机器学习是实现人工智能的重要技术手段之一，在计算机视觉、自然语言处理、搜索引擎与推荐系统等领域有着重要应用. 现有的机器学习方法往往注重数据中的相关关系而忽视其中的因果关系，而随着应用需求的提高，其弊端也逐渐开始显现，在可解释性、可迁移性、鲁棒性和公平性等方面面临一系列亟待解决的问题. 为了解决这些问题，研究者们开始重新审视因果关系建模的必要性，相关方法也成为近期的研究热点之一. 在此对近年来在机器学习领域中应用因果技术和思想解决实际问题的工作进行整理和总结，梳理出这一新兴研究方向的发展脉络. 首先对与机器学习紧密相关的因果理论做简要介绍；然后以机器学习中的不同问题需求为划分

收稿日期：2021-07-23；**修回日期**：2021-11-15

基金项目：国家自然科学基金项目(61722211, 61773362, 61906180)；中国科学院青年创新促进会(20144310)；联想-中科院联合实验室青年科学家项目；重庆市基础科学与前沿技术研究专项项目(重点)(cstc2017jcjyBX0059)

This work was supported by the National Natural Science Foundation of China (61722211, 61773362, 61906180), the Youth Innovation Promotion Association CAS(20144310), the Lenovo-CAS Joint Lab Youth Scientist Project, and the Project of Chongqing Research Program of Basic Research and Frontier Technology (cstc2017jcyjBX0059).

通信作者：兰艳艳（lanyanyan@tsinghua.edu.cn）

依据对各工作进行分类介绍，从求解思路和技术手段的视角阐释其区别与联系；最后对因果机器学习的现状进行总结，并未来发展趋势做出预测和展望.

关键词 因果关系；伪相关关系；因果推断；机器学习；深度学习；人工智能

中图法分类号 TP181

机器学习是一门研究如何设计算法利用数据使机器在特定任务上取得更优表现的学科，其中以深度学习[1]为代表的相关技术已成为人们研究实现人工智能方法的重要手段之一. 至今机器学习研究已经取得大量令人瞩目的成就：在图像分类任务上的识别准确率超过人类水平[2]；能够生成人类无法轻易识别的逼真图像[3]和文本[4]；在围棋项目中击败人类顶尖棋手[5]；在蛋白质结构预测任务上媲美真实实验结果[6]等. 目前机器学习在计算机视觉、自然语言处理、搜索引擎与推荐系统等领域正发挥着不可替代的作用，相关应用涉及互联网、安防、医疗、交通和金融等众多行业，对社会发展起到了有力的促进作用.

尽管机器学习研究获得了一系列丰硕的成果，其自身的问题却随着应用需求的提高而日益凸显. 机器学习模型往往在给出预测结果的同时却不会解释其中的理由，以至于其行为难以被人理解[7]；同时机器学习模型还十分脆弱，在输入数据受到扰动时可能完全改变其预测结果，即使这些扰动在人看来是难以察觉的[8]；机器学习模型还容易产生歧视行为，对不同性别或种族的人群给予不同的预测倾向，即使这些敏感特征不应当成为决策的原因[9]. 这些问题严重限制了机器学习在实际应用中发挥进一步的作用.

造成这一系列问题的一个关键原因是对因果关系的忽视. 因果关系，指的是 2 个事物之间，改变一者将会影响另一者的关系. 然而其与相关关系有所不同，即使 2 个事物之间存在相关关系，也未必意味着它们之间存在因果关系. 例如图像中草地与牛由于常在一起出现而存在正相关关系，然而两者之间却没有必然的因果关系，单纯将草地改为沙地并不会改变图像中物体为牛的本质. 机器学习的问题在于其模型的训练过程仅仅是在建模输入与输出变量之间的相关关系，例如一个识别图像中物体类别的机器学习模型容易将沙地上的牛识别为骆驼，是因为训练数据中的牛一般出现在草地上而沙地上更常见的是骆驼. 这种具备统计意义上的相关性却不符合客观的因果规律的情况也被称为伪相关（spurious correlation）. 伪相关问题的存在对只考虑相关性的机器学习模型带来了灾难性的影响：利用伪相关特征进行推断的过程与人的理解不相符，引发可解释性问题；在伪相关特征发生变化时模型预测结果会随之改变从而导致预测错误，引发可迁移性和鲁棒性问题；如果伪相关特征恰好是性别和肤色等敏感特征，则模型决策还会受到敏感特征的影响，引发公平性问题. 忽视因果关系导致的这些问题限制了机器学习在高风险领域及各类社会决策中的应用. 图灵奖得主 Bengio 指出，除非机器学习能够超越模式识别并对因果有更多的认识，否则无法发挥全部的潜力，也不会带来真正的人工智能革命. 因此，因果关系的建模对机器学习是必要的，需求也是十分迫切的.

因果理论即是描述、判别和度量因果关系的理论，由统计学发展而来. 长期以来，由于缺乏描述因果关系的数学语言，因果理论在统计学中的发展十分缓慢. 直到 20 世纪末因果模型被提出后，相关研究才开始蓬勃兴起，为自然科学和社会科学领域提供了重要的数据分析手段，同时也使得在机器学习中应用因果相关的技术和思想成为可能. 图灵奖得主 Pearl 将这一发展历程称为“因果革命”[10]，并列举了因果革命将为机器学习带来的 7 个方面的帮助[11]. 我们将在机器学习中引入因果技术和思想的研究方向称为因果机器学习（causal machine learning）. 目前机器学习领域正处于因果革命的起步阶段，研究者们逐渐认识到了因果关系建模的必要性和紧迫性，而因果机器学习的跨领域交叉特点却限制了其自身的前进步伐. 本文希望通过对因果理论和因果机器学习前沿进展的介绍，为相关研究者扫清障碍，促进因果机器学习方向的快速发展. 目前针对因果本身的研究已有大量相关综述文献[12-14]，内容主要涵盖因果发现和因果效应估计的相关方法，但很少涉及在机器学习任务上的应用. 综述文献[15-16]详细地介绍了因果理论对机器学习发展的指导作用，着重阐述现有机器学习方法的缺陷和因果理论将如何发挥作用，但缺少对这一方向最前沿工作进展的整理和介绍，而这正是本文将重点介绍的内容.

1 因果理论简介

因果理论发展至今已成为统计学中的一个重要分支，具有独有的概念、描述语言和方法体系. 对于因果关系的理解也已经不再仅停留在哲学概念的层面，而是有着明确的数学语言表述和清晰的判定准则. 当前广泛被认可和使用的因果模型有 2 种：潜在结果框架（potential outcome framework）和结构因果模型（structural causal model, SCM）. Neyman 等人[17]和 Rubin[18]提出的潜在结果框架又被称为鲁宾因果模

型（Rubin causal model, RCM），主要研究 2 个变量的平均因果效应问题；Pearl[19]提出的结构因果模型使用图结构建模 1 组变量关系，除了效应估计也会关注结构发现问题. RCM 与 SCM 对因果的理解一致，均描述为改变一个变量是否能够影响另一个变量，这也是本文所考虑的因果范畴. 两者的主要区别在于表述方法不同，RCM 更加简洁直白，相关研究更为丰富；而 SCM 表达能力更强，更擅长描述复杂的问题. 虽然目前依然存在对因果的其他不同理解，这些理解通常不被视为真正的因果，例如格兰杰因果（Granger causality）[20]描述的是引入一个变量是否对另一个变量的预测有促进作用，本质上仍是一种相关关系.

本节将对因果相关概念以及 RCM 与 SCM 的相关理论和技术进行简要介绍. 由于本文关注的主要内容是因果机器学习而不是因果本身，本节将侧重于介绍机器学习中所使用的因果的概念和思想，而不会过多关注因果领域自身的前沿研究.

1.1 因果概念

统计学中对于因果关系的定义符合人们直觉上的认知. 在一个数据系统中，用于分析的数据通常会表述为一组变量，每个变量都对应一种已知或未知的产生机制. 对于 2 个给定的变量，如果在保持其他机制不变的情况下，改变一个变量会使得另一个变量也发生改变，则称前者为因，后者为果，同时称两者之间存在因果关系（causal relationship），因变量对果变量的影响称为因果效应（causal effect）. 求解 1 对或多对变量是否存在因果关系以及因果效应强度的任务称为因果推断（causal inference）. 通常而言，如果对因果效应强度的定量研究是显著的，则认为因果关系存在. 判定因果关系的存在性将不可避免地涉及到对原始变量系统的改变，即需要改变目标变量的产生机制，这也是其区别于相关关系（correlation）的关键点. 相对而言，判定 2 个变量 X 和 Y 是否存在相关关系则不需要改变系统，只需检验观测变量的边际分布与条件分布是否一致，即判定 $P(X \mid Y)=P(X)$ 是否成立. Pearl 等人[10]在阐述相关和因果之间的差异时提出了“因果之梯（ladder of causation）”的概念，自下而上将问题划分为关联、干预和反事实 3 个层次，分别对应于观察、行动和想象 3 类活动. 通常而言，回答因果问题需要借助反事实或者干预，若希望仅借助关联来判定因果关系则必须处理好混杂因素，这些都是研究因果理论所需的重要概念. 下面将从回答因果关系判定问题的角度出发，对反事实、干预和混杂因素 3 个概念进行介绍：

反事实（counterfactual）指的是在已经观测到 1 组变量的情况下，假设其中部分变量具有另外的取值的操作. 例如人在反思自己的行为时，往往会考虑“如果我当时没有做某事而是做了其他某事，那么结果将会怎样”，这是典型的基于反事实的思考，是根据结果溯源寻找原因的有效手段. 如果发现某个变量改变取值后会导致结果改变，该变量即是结果的原因之一. 反事实考虑的是一种实际并未发生过也难以再次观测到的情景，因为它假定 2 次观测之间除了需要研究的变量有所改变外，其他外部变量取值和作用机制需完全保持一致. 尽管反事实操作的结果直接反映了变量之间的因果关系，由于通常无法针对同一个体平行地实施 2 种不同操作，使得在实际应用中几乎无法用于因果判定，更多情况下只是作为一种指导性思想使用. 想要判断因果关系的存在性，人们只能诉诸群体层面上的平均观测结果，即采用干预操作.

干预（intervention）指的是改变部分变量产生机制并维持其余机制不变的操作，是因果关系判定和度量的关键操作. 如果对一个变量的干预改变了另一个变量的概率分布，则意味着前者是后者的因. 例如，通常认为海拔高度是气温的因，这是因为海拔高度通过特定的物理机制对气温产生了影响. 如果对海拔高度进行干预，即调整地理位置来改变海拔，气温也会随之产生变化，因为背后的物理机制仍然能够生效；相反，如果对气温进行干预，例如提供额外的热源对空气进行加热，这改变了气温的产生机制却保持海拔的产生机制不变，最终海拔并不会因此而改变. 可见通过干预操作可以对因果关系的存在性和方向性做出清晰的判断，事实上这也是科学研究中最常用的手段，随机对照实验即属于这一思路. 干预不同于反事实，不要求外部变量的取值严格一致，只需要满足概率分布不变的假设即可，这在一般的应用场景中通常可以满足，因此更常用于因果关系的判定. 然而这种通过干预观测系统的改变来判断因果关系的做法并不能解决实际中所有的因果问题，在许多情况下干预操作的成本过高或实施风险过大，甚至可能因为违反伦理道德而无法实际实施，如研究吸烟对肺癌的影响时不能强制要求普通人群吸烟. 这种情况下就需要避免对目标变量进行干预，而仅仅通过观测原有机制产生的数据来估计干预的效果，这类研究问题也成为了因果推断领域重点关注的问题.

混杂因素（confounder）指的是一类变量，如果不对它们的取值进行控制，通过观测数据得到的干预结果的估计就会产生偏差. 通常来说，混杂因素指的是那些能够对所研究的 1 对变量同时产生影响的因素. 例如对于儿童穿鞋尺码与阅读能力呈正相关的现象，年龄即是一个混杂因素，如果不控制年龄则会得

出“儿童穿更大尺码的鞋子能提升其阅读能力”的错误结论，相反若控制年龄变量，即针对不同年龄的儿童分组考察他们鞋子尺码与阅读能力的关系，则会发现两者之间不存在相关关系. 理论上如果可以发现并控制所有的混杂因素，那么因果关系的判定就等价于该条件下相关性的判定. 然而寻找一个充分的变量集合以囊括所有的混杂因素是十分困难的，也不可能在不做任何假设的情况下判断已有变量集合是否充分. 另外，简单地将所有其他变量都视为混杂因素的做法也不可取，例如研究一个人才华和外貌的关系时，对其是否是名人这一变量进行控制就是错误的. 因为一个人成名需要好的才华或者好的外貌，两者都不好的人很难成为名人，所以如果一个名人的外貌不好那么他就更可能有好的才华. 在这种受控条件下两者呈现一种负相关，即使原本两者是不相关的. 如何鉴别和处理混杂因素始终是因果推断领域的核心问题之一.

1.2 因果模型

记待研究的变量为 X 和 Y，其他协变量（covariate）构成的向量为 $\boldsymbol{Z}=(Z_1,Z_2,\cdots)$. 为简化考虑，假设 X 是二值变量，即取值只能为 0 或 1. 现在观测到 1 组数据 $\mathcal{D}=(X^{(i)},Y^{(i)},\boldsymbol{Z}^{(i)})$，需要估计 X 取值由 0 变为 1 时对 Y 的因果效应. 由于 $\boldsymbol{Z}$ 中可能存在混杂因素，直接使用条件期望差值 $\mathbb{E}[Y\mid X=1]-\mathbb{E}[Y\mid X=0]$ 作为估计值可能导致偏差. 在这种情况下想要准确进行因果效应估计，需要做出适当的假设构建模型. 本节将对潜在结果框架 RCM 和结构方程模型 SCM 2 种因果模型的概念理论内容进行简要介绍.

1.2.1 潜在结果框架

潜在结果指的是一个个体如果接受了某种处理会怎样，也就是指如果 $X^{(i)}$ 取某种值时对应 $Y^{(i)}$ 取值会如何. 对于个体 i 来说，采取 $X=x$ 的处理的潜在结果记作 $Y_x^{(i)}$，$X^{(i)}$ 对 $Y^{(i)}$ 带来的因果效应可由 $X^{(i)}$ 的不同取值对应的潜在结果差值来计算，即个体处理效应（individual treatment effect, ITE），定义为 $V_{\mathrm{ITE}}^{(i)}=Y_1^{(i)}-Y_0^{(i)}$. 由于同一个个体通常不可能既采取 $X=0$ 的处理同时也采取 $X=1$ 的处理，实际最多只能观测到 1 个结果，另一个结果则是反事实的，这也是被称为“潜在结果”的原因. X 对 Y 的总体因果效应记为个体处理效应的期望，称为平均处理效应（average treatment effect, ATE）：

$$V_{\mathrm{ATE}}=\mathbb{E}[V_{\mathrm{ITE}}]=\mathbb{E}[Y_1]-\mathbb{E}[Y_0]. \tag{1}$$

平均处理效应等同于对 X 的不同干预所得结果之差. 如果这种干预是实际可行的，那么可以直接通过干预操作获得潜在结果的平均取值，从而计算 ATE. 干预意味着 X 的取值不再由观测决定，而是由实验者确定，这种方式通常称为随机对照实验，$X=1$ 的群体称为处理组，$X=0$ 的群体称为控制组.

然而如 1.1 节所述，干预在许多情况下是不可行的，只能使用观测数据对 ATE 进行估计. 基于潜在结果框架研究使用观测数据研究因果效应的做法最早由 Rubin[18]提出，因此该模型也称作鲁宾因果模型，即 RCM. RCM 对因果的描述较为简洁，除了要研究因果效应的 1 对变量以外，对其他变量的相互作用机制不做假设，因此经常在进行因果效应估计的场景使用. 这种情况下需要考虑混杂因素，真实的 ATE 可以由通过控制全部混杂因素获得. 对变量进行控制指的是按照该变量的不同取值分组，组内计算效应期望之后再在组间计算期望. 如果 $\boldsymbol{Z}$ 是全部混杂因素的集合，那么

$$V_{\mathrm{ATE}}=\mathbb{E}_{\boldsymbol{z}}[\mathbb{E}[Y\mid X=1,\boldsymbol{Z}=\boldsymbol{z}]-\mathbb{E}[Y\mid X=0,\boldsymbol{Z}=\boldsymbol{z}]]. \tag{2}$$

在 RCM 中，如果满足一定的假设，上述计算得到的 ATE 即是 X 对 Y 的真实因果效应. 这些假设包括：

1)个体处理值稳定假设（stable unit treatment value assumption, SUTVA）[21]，指的是一个个体的潜在结果不受其他个体处理的影响. 例如一个人服用药物获得的治疗效果不受其他人是否服用药物的影响.

2)处理分配机制可忽略性（ignorability of treatment assignment mechanism）[22]，指的是固定混杂因素后，潜在结果不受处理方式的影响. 例如对于一个人是否服药导致的潜在治疗效果具有确定性，不随实际是否服药的行为而发生改变.

3)正值性（positivity）[22],指的是对于每个个体均有非零的可能性采取每种处理方式.

采用控制所有混杂因素的方法计算 ATE 在实际问题中可能会遇到困难，通常是由于混杂因素的维度很高，控制相同取值的样本可能数量很少，导致期望估计不准确. 针对这一问题，研究者们提出了多种解决方案. 常见的方法有基于倾向性得分的估计方法、

基于回归的估计方法以及两者相结合的方法. 倾向性得分（propensity score）指的是给定协变量 $\boldsymbol{Z}$ 的情况下获得处理 $X=1$ 的概率，即 $P(X=1|\boldsymbol{Z})$，可以使用机器学习模型进行建模. 文献[22]指出，在 ATE 的表达式中使用倾向性得分代替协变量 $\boldsymbol{Z}$ 仍能够保证估计的正确性，因此可以通过控制倾向性得分计算分组期望的方式来计算 ATE. 一种做法称为倾向性得分匹配（propensity score matching）[22]，为处理组中的每个个体选择得分最接近的 1 个或 1 组对照组个体进行匹配，计算它们结果的平均差值，然后在整个处理组上取平均，即可得到 ATE 的估计. 另一种做法称为逆处理概率加权（inverse probability of treatment weighting, IPTW）[23]，也称为 IPW 或 IPS，通过将每个样本的结果除以倾向性得分后再取平均，即可得到预结果的估计，从而计算 ATE:

$$V_{\mathrm{IPTW}}=\frac{1}{n}\sum_i\frac{X^{(i)}\cdot Y^{(i)}}{P(X=1|\boldsymbol{Z}^{(i)})}-\frac{1}{n}\sum_i\frac{(1-X^{(i)})\cdot Y^{(i)}}{1-P(X=1|\boldsymbol{Z}^{(i)})}. \tag{3}$$

基于回归的估计方法简称回归估计[24]，其思想是使用机器学习模型建模给定处理 X 和协变量 $\boldsymbol{Z}$ 时结果 Y 的期望，即 $\mathbb{E}[Y|X,\boldsymbol{Z}]$，然后用这一回归模型来模拟干预，即可得到 ATE 的估计值：

$$V_{\mathrm{REG}}=\frac{1}{n}\sum_i\mathbb{E}[Y|X=1,\boldsymbol{Z}^{(i)}]-\frac{1}{n}\sum_i\mathbb{E}[Y|X=0,\boldsymbol{Z}^{(i)}]. \tag{4}$$

回归估计方法可以和 IPTW 方法相结合得到双稳健估计（doubly robust estimation, DRE）[25]：

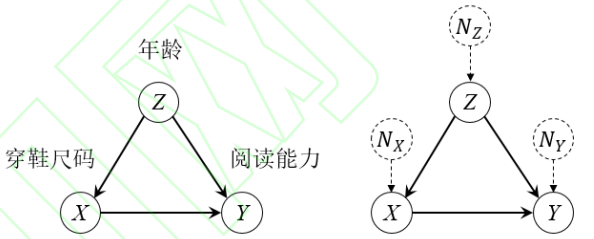
$$V_{\mathrm{DRE}}=\frac{1}{n}\sum_i\frac{X^{(i)}\cdot(Y^{(i)}-\mathbb{E}[Y|X=1,\boldsymbol{Z}^{(i)}])}{P(X=1|\boldsymbol{Z}^{(i)})}-\frac{1}{n}\sum_i\frac{(1-X^{(i)})\cdot(Y^{(i)}-\mathbb{E}[Y|X=0,\boldsymbol{Z}^{(i)}])}{1-P(X=1|\boldsymbol{Z}^{(i)})}+V_{\mathrm{REG}}. \tag{5}$$

只要 2 种估计中的 1 种是可靠的，那么 DRE 整体即是可靠的.

除以上方法外，还有混杂平衡（confounder balancing）[26]、分层（stratification）[27]等众多其他方法处理混杂因素的问题，可参考文献[28]中的介绍，在此不再详细展开. 这些方法都要求混杂因素的值是可观测的，限制了 RCM 在一些场景中的应用. 这种情况下的部分问题可以使用结构因果模型解决.

1.2.2 结构因果模型

结构因果模型 SCM 由 Pearl[19]提出，其思想是将所有需要考虑的变量组织成一个有向无环图，图的每个节点都代表 1 个变量，1 条由节点 A 指向节点 B 的有向连边代表 A 对 B 有直接的因果作用. 这种图又称为因果图（causal graph），记作 $G=(V,E)$，其中节点集合 $V=\{X,Y,Z_1,Z_2,\cdots\}$ 包含所有考虑的变量，边集合 E 包含所有对变量直接因果关系的先验假设（本节所用符号与前文无关）. 例如儿童穿鞋尺码与阅读能力关系的因果图可如图 1(a)表示（假设穿鞋尺码对阅读能力的因果效应是待研究的未知量）：



(a) 原始因果图 (b) 带外生变量的因果图

Fig. 1 Example of causal graph

图 1 因果图示例

结构因果模型中的一个重要概念是结构方程（structural equations），其假设每个节点都对应一个未观测到的外生变量（exogenous variable），节点的值由该外生变量及所有直接父节点变量通过一个方程所唯一确定，例如：

$$X=f_X(\boldsymbol{PA}_X,N_X),$$

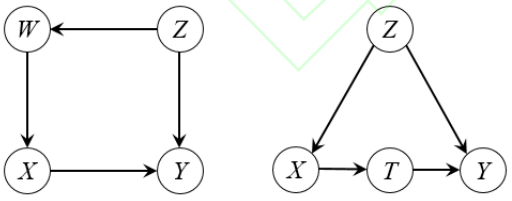
其中 $\boldsymbol{PA}_X$ 指的是节点 X 的所有父节点，N_X 是 X 对应的外生变量. 上面例子所对应的完整结构方程为

$$Z=N_Z,\quad X=f_X(Z,N_X),\quad Y=f_Y(Z,X,N_Y).$$

之所以称为这些方程是“结构方程”，是因为其代表着变量的生成机制，只能由等式右边对左边赋值，而不能随意变换方向. 外生变量描述的是对应节点变量的所有随机因素，其自身具有确定性的概率分布，通常未被观测也无法进行控制，而且 SCM 中假设所有外生变量之间相互独立，图 1(b)展示了一个外生变量的例子. 通过结构方程和外生变量，SCM 能够很清晰地定义干预和反事实操作. 其中干预操作是将干预节点的结构方程替换掉，对应在因果图中即是去掉所有指向干预节点的箭头. 这在 SCM 中也称为 do 操

作，例如将通过干预将节点 X 的取值置为1记作 $do(X=1)$，X 的结构方程也对应修改为 $X=1$，意味着 X 不再受其父节点和外生变量的影响．反事实操作同样由 do 操作给出，但同时会限制所有外生变量取值不变．

在 SCM 中，混杂因素识别可以直接借助因果图结构完成，一个变量成为混杂因素当且仅当存在由该节点指向 X 和 Y 的各 1 条有向路径（指向 Y 的路径不能通过 X ）．X 对 Y 的因果效应仍然可以像 RCM 中一样在识别混杂因素后计算 ATE 得到，不过在 SCM 中可以由干预操作直接给出，即 $\mathbb{E}[Y\mid do(X=1)]-\mathbb{E}[Y\mid do(X=0)]$．这种方法的关键是计算 $P(Y\mid do(X=x))$，这可以通过将因果图视为贝叶斯网络（Bayesian network）进行概率分解得到．然而由 do 操作定义直接给出的求解方法面对稍复杂的因果图时也会变得很复杂，因此一般不会直接使用．更常用的方法称为后门调整（backdoor adjustment）：一条指向 X 并连接 Y 的路径称为 X 到 Y 的后门路径，通过控制路径上的某些节点使得所有后门路径被关闭的方法称为后门调整．路径上的边均指向自身的节点称为对撞节点（collider）．一条路径是关闭的，当且仅当某个对撞节点没有被控制或者某个非对撞节点被控制．RCM 中控制所有混杂因素而不控制其他节点的做法恰恰是后门调整中的一个特例．例如图 2(a)中的因果图，Z 是一个混杂因素，$X\leftarrow W\leftarrow Z\rightarrow Y$ 是一条后门路径，W 和 Z 均不是对撞节点，所以单独控制 Z 或 W，或者同时控制两者都是可以的．



(a) 前门调整 (b) 后门调整

Fig. 2 Example of frontdoor/backdoor adjustment

图 2 前门/后门调整示例

使用 SCM 相对于 RCM 的优势最主要体现在混杂因素无法观测的场景．这种情形下 RCM 将无法使用，而 SCM 可以通过一种称为 do 演算（do-calculus）的方法将因果效应的计算转化为仅在可观测变量上的计算，从而解决部分问题．do 演算包含 3 条规则，这些规则已被证明是完备的，即如果存在一种仅通过可观测变量的观测分布计算因果效应的方法，那么这种方法一定能由 do 演算推导得到，由于篇幅所限不在此展开详细介绍．do 演算的一个常见实例是前门调整（frontdoor adjustment）[29]，如图 2(b)中的因果图，变量 T 称为前门变量，因为其不受 Z 的直接影响，且 X 对 Y 的效应仅仅通过 T 生效．通过前门变量可以在不观测 Z 的情况下计算因果效应：

$$P(Y\mid do(X=x))=\sum_{x'}\sum_{t}P(Y\mid T=t,X=x')\cdot P(X=x')P(T=t\mid X=x). \tag{6}$$

在因果推断及因果机器学习任务中，因果图通常是未知的．一种方式是根据具体问题结合领域知识给出先验的因果图结构，另一种方式是从数据中学习部分因果图信息．后者又被称为因果发现（casual discovery）任务，目的是从一系列变量的观测结果中推断因果图结构．因果发现有以下几类主要方法：基于约束的方法、基于评分的方法和基于结构方程的方法．基于约束的方法主要考虑数据中的条件独立性，通过检验各个变量之间是否条件独立，给出可能的因果图的等价类，即确定部分连边及其方向．这类方法包括 PC（Peter and Clark）[30]，IC（inductive causation）[31]，FCI（fast causal inference）[32]方法等．基于评分的方法思路是利用评分函数来求解得分最高的因果图，常见的评分为贝叶斯信息准则（Bayesian information criterion, BIC）[33]，即联合考虑样本似然和因果图的复杂度，代表性方法是 GES（greedy equivalence search）[34]．基于结构方程的方法是对结构方程的形式做一定的假设，从而可以求解完整的因果图，但同时适用范围也受到方程形式的限制，常见方法包括 LiNGAM（linear non-gaussian acyclic model）[35]和后非线性模型（post-nonlinear model）[36]等．因果发现在实际应用中面临的最大问题是可识别性（identifiability），即能否从观测数据中识别唯一确定的因果图．

因果图的出现还催生了中介分析[37-38]的研究方向，即在有中介变量（mediator）存在的情况下将 X 对 Y 的因果效应借助分解为直接效应和间接效应．如图 3 所示，X 对 Y 产生的因果效应由 2 条路径共同决定，一条是经由中介变量 M 间接影响 Y，一条是直接对 Y 产生影响．

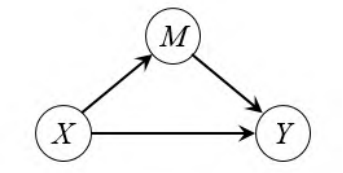


Fig. 3 Example of mediation analysis

图 3 中介分析示例

假设已观测到 $X=x, M=m$ 时有 $Y=Y_{xm}$，这一观测相对于参考情况 $X=x^*$ 下的期望 $Y=\mathbb{E}[Y_{x^*}]$ 之间的差距称为全效应（total effect, TE），即 $V_{\mathrm{TE}}=Y_{xm}-\mathbb{E}[Y_{x^*}]$. 直接效应和间接效应需要依靠反事实来定义，例如直接效应可以视为在观测样本上缺少 $X=x$ 造成的差距或者在参考情况下添加 $X=x$ 造成的差距，前者称为全直接效应（total direct effect, TDE），后者称为自然直接效应（natural direct effect, NDE），分别有 $V_{\mathrm{TDE}}=Y_{xm}-Y_{x^*m}$，$V_{\mathrm{NDE}}=\mathbb{E}[Y_x]-\mathbb{E}[Y_{x^*}]$. 同样的，间接效应也分为 2 种，全间接效应（total indirect effect, TIE）与自然间接效应（natural indirect effect, NIE），分别有 $V_{\mathrm{TIE}}=Y_{xm}-\mathbb{E}[Y_x]$，$V_{\mathrm{NIE}}=Y_{x^*m}-\mathbb{E}[Y_{x^*}]$. 以上效应之间满足关系 $V_{\mathrm{TE}}=V_{\mathrm{TDE}}+V_{\mathrm{NIE}}=V_{\mathrm{TIE}}+V_{\mathrm{NDE}}$.

2 因果机器学习相关工作介绍

近年来随着因果理论和技术的成熟，机器学习领域开始借助因果相关技术和思想解决自身的问题，这一研究方向逐渐受到研究者越来越多的关注. 至今，因果问题被认为是机器学习领域亟待解决的重要问题，已成为当下研究的前沿热点之一. 机器学习可以从因果技术和思想中获得多个方面的益处. 首先，因果理论是一种针对数据中规律的普适分析工具，借助因果图等语言可以对研究的问题做出细致的分析，有利于对机器学习模型的目标进行形式化以及对问题假设的表述. 其次，因果推断提供了消除混杂因素以及进行中介分析的手段，对于机器学习任务中需要准确评估因果效应及区分直接与间接效应的场景有十分重要的应用价值. 再者，反事实作为因果中的重要概念，也是人在思考求解问题时的常用手段，对于机器学习模型的构建和问题的分析求解有一定的指导意义.

本节将对近年来因果机器学习的相关工作进行整理介绍，涉及应用领域包括计算机视觉、自然语言处理、搜索引擎和推荐系统等. 按照所解决问题的类型进行划分，因果机器学习主要包括以下内容：可解释性问题主要研究如何对已有机器学习模型的运作机制进行解释；可迁移性问题主要研究如何将模型在特定训练数据上学到的规律迁移到新的特定环境；鲁棒性问题主要研究寻找普适存在的规律使模型能够应对各种未知的环境；公平性问题主要研究公平性度量指标并设计算法避免歧视；反事实评估问题主要研究如何在存在数据缺失的场景中进行反事实学习. 这些问题与因果理论的关系如图 4 所示，下面针对这些问题分别展开介绍.

2.1 可解释性问题

机器学习模型会根据给定输入计算得到对应的输出，但一般不会给出关于“为什么会得到此输出”的解释. 然而这种解释有助于人们理解模型的运作机制，合理的解释能够使结果更具有说服力. 因此近年来涌现出许多致力于为现有模型提供解释方法的工作，为模型的诊断分析提供了有效手段[39]. 解释的核心在于“模型得到此输出，是因为输入具有什么样的特征”，这本质上是在探讨在此模型参与过程中输入特征与输出结果之间的因果关系，例如估计特征对输出变量的因果效应强度.

由于机器学习模型对输入数据的处理过程是一个独立而完整的过程，输入与输出变量之间一般不会受到混杂因素的影响，因此即使不使用因果术语也可以对任务进行描述. 这体现为早期的模型解释方法并不强调因果，少数强调因果的方法也并不一定依赖因果术语. 因果理论的引入为可解释性问题领域带来的贡献主要有 2 个方面：一是在基于归因分析的解释方法中建模特征内部的因果关系，二是引入一类新的解释方法即基于反事实的解释. 基于归因分析和基于反事实的解释构成了当前最主要的 2 大类模型解释方法，以下分别展开介绍.

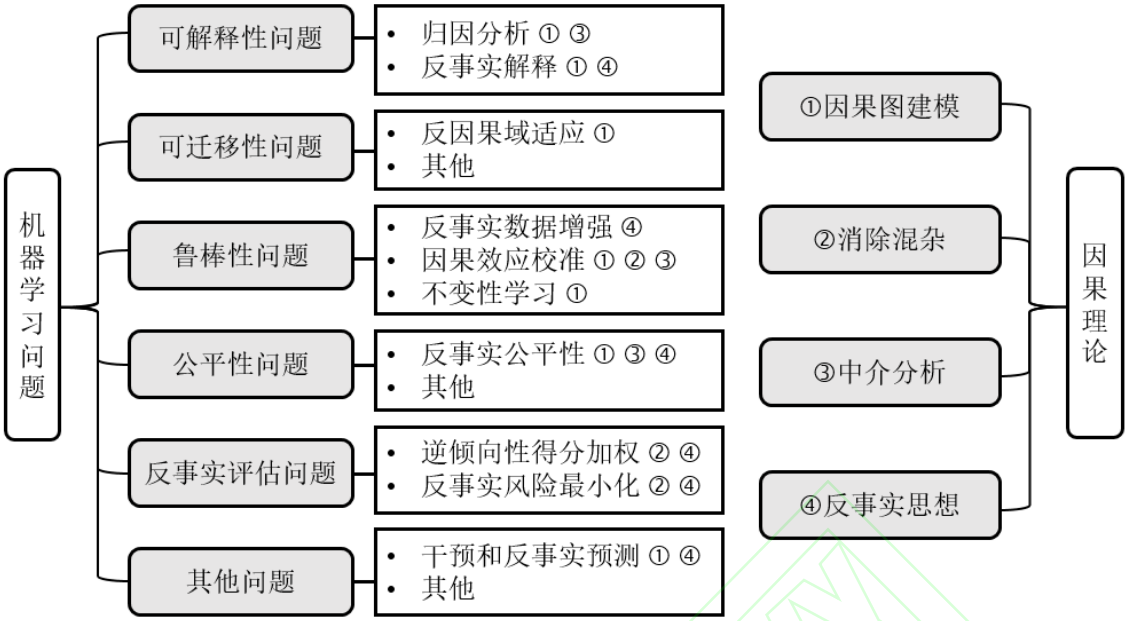


Fig. 4 Overview of main research problems in causal machine learning

图 4 因果机器学习的主要研究问题总览

Table 1 Application of Causal Methods on Interpretability Problems

表 1 因果方法在可解释性问题上的应用

分类		典型思路和方法
基于归因分析	忽略特征间结构	直接计算每个输入特征对模型输出的因果效应[40-46]
	考虑特征间结构	引入输入特征间的先验因果图结构，调整特征对模型输出的因果效应[47-48]
基于反事实	输入数据反事实	在模型输入空间构造反事实样本[49-61]
	输出数据反事实	对生成模型的中间节点进行反事实，构造反事实生成样本[62]
	反事实可行性	对反事实操作的约束条件进行额外建模[63-66]

2.1.1 基于归因分析的解释方法

基于归因分析（attribution）的方法是机器学习模型解释方法中最早出现也是最为成熟的方法. 对于一个具有 n 个特征的样本 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$，模型将其映射为输出 $Y=y$，归因分析指的是为每个特征分配一个归因值，即构造一个归因向量 $\boldsymbol{\Phi}=(\phi_1,\phi_2,\cdots,\phi_n)$，其中 ϕ_i 代表特征 $X_i=x_i$ 对结果 $Y=y$ 的贡献大小（本节所用符号与前文无关）. 基于归因分析的常见解释方法主要包括：LIME[40]，Grad-CAM[41]，Integrated Gradient(IG)[42]，Shapley Values(SHAP)[43]等.

以 SHAP 方法为例，SHAP 方法认为一个特征对于输出变量的效应强度应该为：使用该特征的预测结果与不使用该特征的预测结果之差. 将整个特征集合记作 $\mathcal{F}=\{1,2,\cdots,n\}$，预测输出结果需要选择一个特征子集，计算特征 i 的效应需要对比不含 i 的所有子集与对应添加 i 的子集的差别，即 $f_{S\cup\{i\}}(\boldsymbol{X}_{S\cup\{i\}})-f_S(\boldsymbol{X}_S)$. 在所有满足条件的子集上

取加权平均的结果即为特征 i 的 Shapley 值．SHAP 方法将 Shapley 值作为特征的归因值，其他归因方法也会得到这样的归因向量．

基于归因分析的解释方法虽然描述的是因果关系，但一般不依赖因果术语，一些文献采用了因果的表述，本质上仍属于归因解释的框架．例如文献[44]提出一种针对端到端文本生成模型的因果解释框架，预测源文本中的单词对目标文本中的单词的影响强度，相当于将源文本单词视为特征集合，针对每个目标单词的预测都给出 1 个对应的归因向量．文献[45]提出一种在不确定因素下图像分类模型的因果解释方法，主要贡献在于对每个特征除了计算归因值以外还会计算其置信度．文献[46]提出将机器学习模型整体视为一个 SCM 模型，然后计算每个特征对输出结果的平均处理效应，相当于将解释问题重新使用因果语言进行形式化，但在做法上与其他归因解释方法并无本质不同．

基于归因分析的解释方法一般将每个特征视为独立的变量进行考虑，而当特征之间存在相互影响时就必须借助因果理论进行刻画和求解．文献[47]基于 SHAP 方法将先验因果知识引入，提出非对称 SHAP 方法，其核心思想在于，原始 Shapley 值计算方法会将所有特征序列的置换平等看待，而非对称 SHAP 会调整这些置换的权重，例如将不符合因果顺序的置换的权重置为 0，从而将子节点的因果效应汇总归于祖先节点的因果效应．文献[48]同样基于 SHAP 方法，从另一个角度提出了引入因果知识的方式．SHAP 方法需要计算特征子集 S 下模型的期望输出 $v(S)$，为保持样本位于数据流形之上，一般选择计算以 $\boldsymbol{X}_S=\boldsymbol{x}_S$ 为条件下的期望．该文献认为，在给定因果图结构的情况下应使用 do 操作而非取条件的操作，即 $do(\boldsymbol{X}_S=\boldsymbol{x}_S)$，由该方法得到的归因值称为因果 Shapley 值．同时，该文献利用中介分析将总体效应 $v(S\cup\{i\})-v(S)$ 拆解为直接效应与间接效应，展示了在不同因果结构下对于相同观测数据的解释存在的差异．

2.1.2 基于反事实的解释方法

基于反事实的解释方法是近年来新兴的一类模型解释方法，其中“反事实”作为一种因果术语指的是如果样本的部分特征发生了改变而其他特征不变将会怎样．一般而言，反事实解释方法会寻找一种样本特征处理方法使样本的预测结果发生显著改变，例如对图像的局部进行替换或遮挡从而改变分类类别等．与归因分析不同，反事实解释并不会提供每个特征的重要度，而是直接给出改变预测结果的途径，相当于给出信息“模型对样本 X 的输出为 A 而不是 B，是因为 X 具有特征 $\boldsymbol{f}$，如果该特征变为 $\boldsymbol{g}$ 则其输出会变为 B”（本节所用符号与前文无关）．

文献[49-50]提供了 1 类典型的反事实解释方法．针对图像分类任务，需要从给定原始图像中选择 1 块区域使其替换为其他内容后变为目标类别．所替换内容为目标类别的 1 幅干扰图像的某一块区域．修改后的复合图像构成了原样本的一个反事实解释，如图 5 所示：

(a) 原始图像 (b) 干扰图像 (c) 复合图像

Fig. 5 Example of counterfactual explanation [49]

图 5 反事实解释示例[49]

文献[51]在为图像分类模型构造反事实解释时避开了图像的修改合成过程，直接生成可读的文本解释，例如“它不是猩红丽唐纳雀，因为它没有黑色的翅膀”．文献[52]通过优化的方式求解图像的掩码，使得遮挡该区域后模型不再将其分类为原始类别．文献[53]在视频分类上应用反事实解释，选取视频中关键片段的关键矩形区域，并通过预测该区域的语言学属性为其搭配简单的文本解释，如“是骑行而不是滑板运动，因为姿势是坐着”．文献[54]利用局部语义纹理特征作为解释工具，称为断层线（fault-line），解释原始图像需要增减哪些语义特征才能改变为目标类别．文献[55]在强化学习中将行动影响建模为 SCM，为智能体的行为做模板式的反事实解释，例如“智能体选择建造供应站而不是兵营，因为可以拥有更多供应站，有利于破坏对手更多的单位和建筑”．文献[56]提出反事实解释需满足可行性和多样性，并采用优化

的方式求解反事实解释的集合．文献[57]为贝叶斯网络分类器构造反事实解释，求解值改变即引起结果改变的变量集合．文献[58]在反事实解释的基础上提出半事实（semi-factual）解释的概念，与反事实解释的区别在于其对于样本的修改接近改变输出但实际并未真正改变．文献[59]为针对图（graph）数据的分类器设计反事实解释方法，提出一种基于搜索的方法寻找反事实图．文献[60]针对以往基于算法的反事实样本构造方法过于耗时的问题，提出一种基于模型的反事实样本生成方法．文献[61]为集成树（tree ensemble）模型设计了反事实解释方法，建模为混合整数规划问题并进行求解．

文献[62]针对图像生成模型研究了一种特殊的反事实解释方法．由于图像生成模型的输入为无直观意义的噪声，一般的反事实研究不易产生有价值的解释．因此该方法不再针对输入特征进行反事实，而是将神经网络模型视为白盒 SCM，在其内部表达节点上进行反事实，其目的是寻找模型中的独立生成机制，从而有助于对模型的理解．具体方法是寻找一些网络内部节点集合，使得在 2 幅图像上做数值交换后输出差异尽可能大，这些节点即反映了图像的关键生成机制．图 6 展示了该文献方法可通过 2 幅图像在关键内部节点上的数值交换实现反事实的图片混合效果．

Fig. 6 Example of counterfactual image hybridization [62]

图 6 反事实图像混合示例[62]

基于反事实的模型解释方法相对于归因解释的优势在于其直接提供了改变当前模型预测结果的操作手段．然而一些文献指出，反事实解释提出的建议并不会考虑实际实施的代价，甚至可能是无法操作的．文献[63]研究了反事实解释偏离数据分布的问题，提出基于马氏距离和局部异常因子的代价函数约束反事实解释的可行度，将寻找可行反事实解释的问题转化为混合整数线性优化的求解问题．文献[64]在此基础上基于因果图分析了在多个特征上反事实操作的顺序问题，因果图可由因果发现技术获得．文献[65]研究了在特征为二值情景下的反事实解释的可行性问题，证明寻找最优反事实策略是 NP 难的，因此提出一种高效的随机算法进行近似求解．文献[66]研究了特征之间存在因果关联时如何提供可行反事实解释的问题，在假设因果图结构已知的情况下，用高斯过程建模结构方程的不确定性，提出个体和亚群体级别的 2 类可行性反事实解释，使用梯度优化的方式求解．

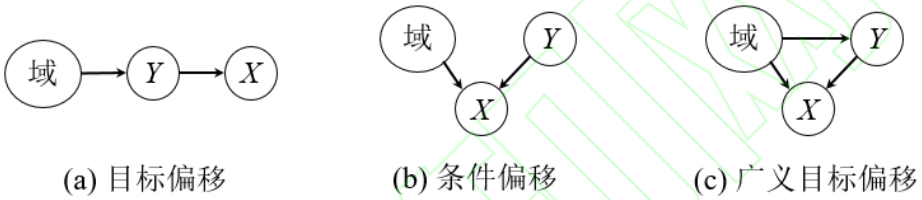
2.2 可迁移性问题

机器学习研究通常会在一个给定的训练数据集上训练模型，然后在同数据分布的验证集或测试集上进行测试，这种情况下模型的表现称为分布内泛化（in-distribution generalization）．在一般的应用场景中，机器学习模型会部署在特定数据环境中，并使用该环境中产生的数据进行模型训练，其性能表现可以用分布内泛化能力来度量．然而在一些场景中，目标环境中的标注数据难以获取，因此更多的训练数据只能由相似的替代环境提供．例如训练自动驾驶的智能体时由于风险过高不能直接在真实道路上行驶收集数据，而只能以模拟系统中所获取的数据为主进行训练．这种场景下的机器学习任务又称为域适应（domain adaptation），属于迁移学习（transfer learning）的范畴，即将源域（source domain）中所学到知识迁移至目标域（target domain）．这里的域（domain）和环境（environment）的含义相同，可以由产生数据的不同概率分布来描述，下文将沿用文献中各自的习惯称呼，不再对这 2 个概念进行区分．

在可迁移性问题中，因果理论的主要价值在于提供了清晰的描述语言和分析工具，使研究者能够更准确地判断可迁移和不可迁移的成分，有助于设计针对不同场景的解决方案．因果推断中关注的效应估计问题本质上是在研究改变特定环境作用机制而保持其他机制不变的影响，这与迁移学习中域的改变的假设相符，即目标域和源域相比继承了部分不变的机制可以直接迁移，而剩余部分改变的机制则需要进行适应．因此在因果理论的指导下，迁移学习中的关键问题就是建模并识别变与不变的机制．目前因果迁移学习一般假设输入 X 与输出 Y 之间有直接因果关系，重点关注无混杂因素情况下变量的因果方向和不变机制，以下对相关工作展开介绍．

文献[67]是早期研究因果理论对机器学习指导作

用的经典工作，主要使用结构方程模型研究了输入变量 X 与输出变量 Y 之间的因果方向对可迁移性的影响：如果有 $X \to Y$，那么输入分布 $P(X)$ 与条件分布 $P(Y \mid X)$ 可视为独立的机制，目标域数据所提供的输入 $P'(X)$ 信息对 $P'(Y \mid X)$ 的预测不会产生直接作用，而输出 $P'(Y)$ 却因包含了 $P'(Y \mid X)$ 的信息而有助于预测；如果有 $Y \to X$，则输入分布 $P(Y)$ 与条件分布 $P(X \mid Y)$ 可视为独立的机制，结论将与前面情况完全相反，这种情况称为反因果（anti-causal）．正向因果情景中仅 $P(X)$ 发生改变而 $P(Y \mid X)$ 不变的情况常被称为协变量偏移（covariate shift, CovS）．文献[68]针对实际情形中更常见的反因果迁移问题进行了进一步的建模，如图 7 所示：如果只有 $P(Y)$ 发生了改变则称为目标偏移（target shift, TarS）；如果只有 $P(X \mid Y)$ 发生了改变则称为条件偏移（conditional shift, ConS）；如果两者都发生了改变则称为广义目标偏移（generalized target shift, GeTarS）．这些工作为因果理论指导迁移学习奠定了基础．



(a) 目标偏移 (b) 条件偏移 (c) 广义目标偏移

Fig. 7 Causal graphs of 3 types of anti-causal transfer problems [68]

图 7 3 类反因果迁移问题的因果图[68]

Table 2 Application of Causal Methods on Transferability Problems

表 2 因果方法在可解迁移性问题上的应用

分类	典型思路和方法
仅考虑输入输出与域变量间的因果图	求解在协变量偏移[69]、目标偏移[70]、条件偏移[71]、广义目标偏移[68,72]情况下的建模方法
考虑含其他复杂变量的因果图	引入先验因果图[73-76]或从数据中进行因果发现[77]

后续许多工作沿用这一框架展开，在不同的先验因果图结构下求解迁移学习问题．文献[78]探讨了在有多个源域提供数据的情况下如何求解各类反因果迁移问题．文献[69]提出协变量偏移情况下对 $P(Y \mid X)$ 不变的假设过强，认为只需假设存在特征集合 S 使得 $P(Y \mid S)$ 跨环境不变即可，并设计搜索算法寻找 S．文献[70]针对目标偏移问题已有方法无法处理高维数据、连续数据和大规模数据等问题，提出一种新的标签变换方法求解，将源域的标签 Y 变换之后再重新训练或微调获得 $P(Y \mid X)$ 模型．文献[71]研究条件偏移情况，基于变分自编码器结构学习 X 的隐变量表达，并引入对抗训练使语义表达与域表达解耦合，语义表达即可用于迁移．文献[72]指出在广义目标偏移的情况下使用文献[68]中的局部尺度变换方法可能无法满足需求，进而设计算法通过寻找条件可迁移成分（conditional transferable components）进行求解．

一些迁移学习的工作也考虑从其他角度引入因果理论和技术．文献[73]在因果图建模的基础上额外建模了结构方程，基于非线性独立成分分析构造目标域的伪样本对训练数据进行扩充．文献[74]利用因果

图在一个虚拟的“密室逃生”任务上建模不同层次的因果结构，以将所学知识迁移到未见过的相似场景．文献[75]研究了一种特殊的模仿学习迁移任务，即演示者与学习者接收不同的传感器输入，如自动驾驶智能体上路时无法观测到学习时的指示灯信号，使用 SCM 分析可变与不变的部分以指导学习．文献[76]针对小样本学习（few-shot learning）这一特殊的域适应任务，认为预训练知识是特征和标签的混杂因素，采用后门调整消除其影响．文献[77]将域适应问题转化为增广的因果图上的推断问题，在多个源域的数据上进行结构发现，然后使用条件生成对抗网络建模．

迁移学习问题与因果密切相关，对于跨环境不变机制的挖掘和利用始终是核心问题之一．由于问题场景的不同会导致因果机制可变也可不变，无法统一下定论，需要具体问题具体分析，因果机器学习在这一问题上仍有宽阔的发展空间．

2.3 鲁棒性问题

迁移学习允许模型获得目标环境的少量数据以进行适应，然而在一些高风险场景中，可能需要机器学习模型在完全陌生的环境中也能正常工作，如医疗、法律、金融及交通等．以自动驾驶为例，即使有大量的真实道路行驶数据，自动驾驶智能体仍会面临各种突发情况，这些情况可能无法被预见但仍需要被正确处理．这类任务无法提供目标环境下的训练数据，此时模型的表现称为分布外泛化（out-of-distribution generalization）．如果模型具有良好的分布外泛化能力，则称其具有鲁棒性（robustness）．

这类问题在未引入因果术语的情况下就已经展开了广泛的研究．如分布鲁棒性研究[79-81]考虑当数据分布改变在一定幅度之内时如何学习得到鲁棒的模型，常见思路是对训练样本做加权处理；对抗鲁棒性研究[8,82-83]考虑当样本受到小幅度扰动时模型不应当改变输出结果，常见思路是将对抗攻击样本加入训练．这类研究常常忽略变量间的因果结构，面临的主要问题是很难决定数据分布或者样本的扰动幅度大小和度量准则，这就使得研究中所做的假设很难符合真实场景，极大地限制了在实际中的应用．因果理论的引入为建模变量间的结构提供了可能，同时其蕴含的“机制不变性”原理为鲁棒性问题提供了更合理的假设，因为真实数据往往是从遵循物理规律不变的现实世界中采集获得．例如针对输入为 X 、输出为 Y 的预测问题，不考虑结构的分布鲁棒性方法会假设未知环境 $P'(X,Y)$ 应当与真实环境 $P(X,Y)$ 的差异较小，如限制联合分布的 KL 散度小于一定阈值；而考虑结构的因果方法则通常会假设机制不变，如当 Y 是 X 的因时假设 $P'(X\mid Y)=P(X\mid Y)$ 等，在因果关系成立的情况下后者通常是更合理的．

一些从伪相关特征入手研究鲁棒性问题的工作虽然未时使用因果术语，实际上已经引入了因果结构的假设．这些工作针对的往往是已知的伪相关特征，如图像分类任务中的背景、文本同义句判断 SNLI 数据集中的单条文本[84]、重复问题检测 QuaraQP 数据集中的样本频率[85]等．在实际场景中针对这些伪相关特征进行偏差去除（debias），以避免其分布发生变化时影响模型表现．这类工作的隐含假设的是伪相关特征与目标预测变量没有因果关系．一种直接的解决方法是调整训练数据的权重，使得伪相关特征不再与预测变量相关[85]． 还有 1 类方法会单独训练 1 个仅使用伪相关特征预测的模型，然后将其与主模型融合在一起再次训练，完成后仅保留主模型[86-87]．然而由于实际应用中通常很难预先确定伪相关特征，这类工作在解决鲁棒性问题上具有明显的局限性．

因果理论的引入对于解决鲁棒性问题提供了新的思路，主要的优势在于对变量结构的建模和更合理的假设．这类方法包括反事实数据增强、因果效应校准和不变性学习．反事实数据增强考虑从数据入手消除伪相关关系，因果效应校准通过调整偏差特征的作用来减轻偏差，不变性学习通过改变建模方式学习不变的因果机制，以下分别展开介绍．

Table 3 Application of Causal Methods on Robustness Problems

表 3 因果方法在鲁棒性问题上的应用

分类		典型思路和方法
反事实数据增强	伪相关特征反事实	构造额外训练数据，在保持预测结果不变的前提下微调数据[88-93]
	因果特征反事实	构造额外训练数据，更改关键因果特征并修改预测结果[92-95]
因果效应校准	基于后门调整	根据对问题的认识指出混杂因素，对其估计后消除影响[96-99]
	基于中介分析	根据对问题的认识指出中介变量，对其估计后消除影响[97,100-102]
不变性学习	稳定学习	将每个特征视为处理变量，通过样本加权消除混杂，识别因果特征[103-107]
	不变因果预测	基于多环境训练数据，利用假设检验确定因果特征集合[108-110]
	不变风险最小化	基于多环境训练数据，在模型优化目标中添加跨环境不变约束，学习因果特征[111- 113]

2.3.1 反事实数据增强

反事实数据增强（counterfactual data augmentation）的核心思想是针对真实的因果关系额外构造反事实数据加入训练，以消除非因果变量与预测变量间的相关性，这里的因果关系通常是由人的先验认知给出的.“反事实”指的是对样本做改动，通过改变关键的因果特征使得预测结果改变.文献[114]给出了这类方法的有效性分析，下面对这类方法的相关工作进行简要介绍.

在自然语言处理领域主要关注文本分类任务中的数据增强.文献[88]针对文本数据中的性别-职业偏差，将性别相关词语替换成相反性别的对应词语作为数据增强.文献[89]同样针对性别偏差，认为直接替换性别词加入数据会造成统计属性的异常，因此建议改为随机替换原有数据，并额外提出一种姓名干预的方法将与性别相关的姓名词一同替换.文献[90]指出性别词替换的方法并不适用于某些性别与语法关联紧密的语言，如西班牙语和希伯来语，因此提出 1 套新的方法针对这类场景，在对性别词进行干预后重新推断新的词形和句法标签，在整条文本上进行调整.文献[91]针对文本情感分类任务中未知的伪相关特征问题，通过人工编辑文本使得改动幅度不大且情感类别反转，修改得到的文本作为训练数据扩充.文献[94]同样针对文本情感分类任务，通过匹配含义相近但标签相反的文本来寻找关键因果词，然后在原始文本上将因果词替换为其反义词，同时反转标签构成反事实数据.

在计算机视觉领域主要关注视觉问答和图像分类等任务中的数据增强.文献[92]在 VQA 任务中使用生成对抗网络合成图像进行数据增强，针对语义实体进行相关或无关物体的移除，从而去除模型中的一部分伪相关关系.文献[95]提出 2 种针对 VQA 任务的数据增强方法，即遮挡图像中的关键区域，或者遮挡问题文本中的关键词，2 种方法都不依赖人工标注.文献[115]使用 SCM 对 VQA 任务进行建模，通过推断外生变量的分布来构建改变图片或者改变问题的反事实数据.文献[116]提出在使用反事实数据时并不直接加入训练，而是与原数据配对构造对比损失，可以取得更好的效果.文献[117]提出在使用反事实数据增强方法时同时采用梯度监督正则项，可以进一步提高分布外泛化性能.文献[118-119]针对视觉-语言导航（vision-and-language navigation）任务，分别使用对抗训练寻找最难路径以及修改图像特征改变智能体行为的方式构造反事实样本.文献[120]借助因果独立机制概念，人为将图像生成过程分离为背景、形状和纹理的单独作用机制，从而构造出反事实图片的生成模型，将生成的伪造图片加入训练可提升图像分类模型鲁棒性.文献[93]在图像物体分类任务中利用人工标注的边界框信息，通过修改边界框内外的图像，分别构造类别改变和不变的 2 类反事实样本.文献[121]利用图像生成模型习得的隐状态特征表

达，使用主成分分析方法识别关键因果成分，通过干预隐状态和风格迁移的方式分别构建反事实图像，提升了图像分类模型的鲁棒性.

反事实数据增强作为一种与模型无关的技术，除了直接应用于去除伪相关特征外，本身也是一种解决训练数据不足的有效手段. 这种情况下也可以看作是过少的训练数据更容易带来各种伪相关特征问题. 文献[122]针对命名实体识别任务中数据标注代价高的问题，使用替换实体的方法进行数据增强，并从 SCM 的角度阐述了方法的合理性. 文献[123]研究基于语言的图片编辑任务中的数据稀缺问题，将语言指令关键词随机替换为同类别词进行数据增强. 文献[124]关注强化学习任务中的一类局部机制可以解耦合的场景，如打台球任务中短时间内台球只会两两碰撞，提出一种局部因果模型，通过替换可解耦的局部状态实现数据增强.

2.3.2 因果效应校准

针对机器学习的鲁棒性问题，有一类工作会根据人的先验知识，对容易带来偏差的特征的作用进行调整，使其符合真实的因果效应，从而实现跨环境预测的稳定性. 本文将这类研究统称为因果效应校准. 典型的思路是根据问题的特点提出对应的因果图假设，然后针对混杂因素变量使用后门调整进行校准，或者针对中介变量使用中介分析进行校准等. 以下对各个工作分别进行简要介绍.

文献[96]研究法庭意见文本生成任务，由于原告通常会在很可能被支持的情况下提起诉讼，因此主张是否受法庭支持成为了原告声明和法庭意见之间的混杂因素，使用后门调整处理后减少了支持主张的意见文本，更符合真实判决结果. 文献[97]在视觉对话任务中根据因果图结构提出 2 种校准策略（如图 8 所示）：一是切断对话历史对于未来对话文本的直接效应而仅保留经由问题文本的中介效应，二是建模未观测混杂因素并使用后门调整消除其带来的伪相关作用. 文献[98]在视觉常识推理等任务中，认为图像中的物体标签是混杂因素，使用后门调整处理后获得更准确的图像特征表达. 文献[99]在弱监督语义分割任务中，只利用图像标签作为监督信号，并认为图像标签是混淆因素，通过后门调整改善分割质量. 文献[100]研究场景图（scene graph）生成任务，由于训练数据中缺少针对图像中物体位置关系的精确描述，如本因该描述成“站在……上面”和“躺在……上面”，却使用了“在……上面”这种缺乏信息的描述，提出物体标签是图片特征对位置描述关系的中介变量，这一中介效应应当被削弱，因此在预测描述关系时使用全直接效应 TDE 来代替全效应 TE. 文献[101]指出分类问题中长尾分布的尾部预测不准的部分原因是在优化算法中使用了动量，而动量是输入和输出变量间的混杂因素，且动量在头部的投影是输入到输出间的中介变量，因此同时采用后门调整和 TDE 方法进行校准. 文献[102]研究视觉问答任务中问题文本引起的语言偏差问题，使用全间接作用 TIE 代替原有预测，避免问题文本对回答产生直接作用，获得更高的预测准确率.

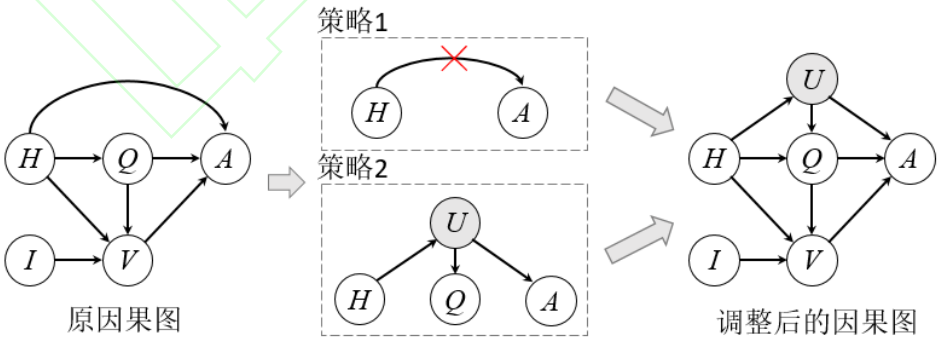


Fig. 8 Causal graph and 2 calibration strategies in visual dialogue tasks [97]

图 8 视觉对话任务的因果图和 2 种校准策略[97]

除了后门调整和中介分析以外，也有工作采用其他方法实现因果效应校准. 文献[125]在模仿学习任务中，由于专家对环境的观测与智能体的观测并不一致，因此定义了一种部分可观测的 SCM 进行建模和求解. 文献[126]研究了一类运行时混杂（runtime confounding）的问题，即模型在训练时可以访问所有特征，而在测试时却有部分特征无法获取，采用双稳健估计 DRE 算法解决该问题. 文献[127]针对图像分

类中的组合泛化问题，认为标签和属性相互独立且图像由两者生成，采用反向因果建模求解．文献[128]研究词向量中性别偏差的问题，认为单纯使性别无关词向量垂直于性别定义词向量不足以解决问题，因为性别无关词仍可能被聚类为同一簇从而提供偏差信息，因此使用半兄弟回归（half-sibling regression, HSR）[129]消除两者之间的混杂因素．文献[130]同样针对词向量的性别偏差问题，提出一种反事实生成的方法，将词向量解耦成性别相关和无关的 2 部分，通过反转性别标签得到性别相反的词向量，与原词向量取平均后得到中性的词向量．文献[131]使用 HSR 技术为词向量降噪，使内容词和功能词有更准确的含义表达．文献[132]在视觉问答任务中使用前门准则修正图像和问题对回答的因果作用，使模型中的注意力机制更好地捕获真实因果关系．

2.3.3 不变性学习

机器学习中的鲁棒性问题与现实物理世界中的因果不变性机制有着紧密的联系．由于实际应用考虑的往往是宏观的物理过程，任何因果机制都难以保证始终一成不变，因此考虑无任何约束的鲁棒性问题意义不大且没有必要，重要的是满足常见环境下的需求．要达到这一目标，建模常见环境中不变的因果机制就成了实现模型鲁棒性的必然需求．本文将这类研究统称为不变性学习．不同于反事实数据增强和因果效应校准等方法需要对伪相关特征有一定的认识，不变性学习可以对伪相关特征未知的情景进行处理．常见思路包括稳定学习（stable learning）、不变因果预测（invariant casual prediction, ICP）和不变风险最小化（invariant risk minimization, IRM），以下分别展开介绍．

稳定学习[103]指的是要求模型在不同的环境中具有稳定的性能表现，既要有较高的平均表现，也要有较低的方差．稳定学习假设预测目标仅由 1 组因果特征决定，其预测作用具有不变性，而其他特征为伪相关特征．稳定学习一般利用单个环境的数据，通过样本加权的方式消除伪相关特征的影响，从而使因果特征被保留下来．文献[104]提出因果正则化逻辑回归（causally regularized logistic regression, CRLR），对每个样本学习 1 个权重，在优化经验风险的同时需使得以每个特征为处理变量（treatment variable）的协变量分布尽可能一致，所学得的关键特征更符合人的判断标准．文献[103]明确提出稳定学习的概念，并提出深度全局平衡回归（deep global balancing regression, DGBR）算法，使用自编码器（auto-encoder）将特征降至低维空间，然后采用与 CRLR 相同的思路求解，根据学到的权重检查每个协变量条件下处理变量是否与预测结果变量独立，不独立的即为稳定特征．文献[105]指出 DGBR 算法可能存在模型设定偏误（misspecification）问题，提出去相关加权回归（decorrelated weighting regression）算法，引入特征非线性变换解决这一问题．文献[106]同样针对设定偏误问题，指出输入特征之间存在的共线性特点会放大模型设定偏误带来的误差，因此提出一种样本加权去相关算子（sample reweighted decorrelation operator）消除共线性．文献[107]针对 CRLR 和 DGBR 只能针对线性框架的缺陷，提出基于随机傅立叶特征的非线性特征去相关算法 StableNet，可以更有效地应用于图像等复杂数据类型．目前基于样本加权的稳定学习方法对于数据有一个较强的假设，即对于可能存在的因果特征和伪相关特征的组合均需要存在对应的训练样本，这在实际场景中可能难以满足．因此在该假设不成立时如何应对仍是有待研究的课题．

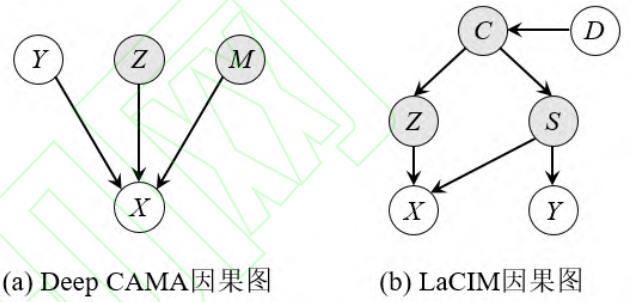
不变因果预测 ICP[108]方法的思路是借助多个环境的数据来确定跨环境不变的特征．文献[108]首次在线性框架下提出了 ICP，基于假设检验的方式确定不变因果特征集合，同时还可以给出置信区间．文献[109]将这一方法拓展至非线性框架，并且可以适用于连续环境．文献[110]将 ICP 方法进一步拓展至时间序列数据上．基于 ICP 的方法均要求因果变量是输入特征的子集，因此一般不适用于高维复杂数据如图像和文本等，然而其思想对这类问题的解决提供了很好的启发作用．

不变风险最小化 IRM[111]方法延续了 ICP 的思想，同样是借助多个环境的数据来学习跨环境鲁棒的模型，但不再从输入特征集合中选择因果特征，而是使用模型抽取特征．IRM 认为机器学习模型可以拆分为特征抽取器和预测器 2 个部分，即输入样本首先通过特征抽取器得到分布式表达，然后预测器将该表达映射为目标输出结果．IRM 假设因果特征应当使得预测器保持跨环境不变性，这种不变性约束被转化为损失函数中的正则项，即在各个环境数据上损失函数对于预测器参数的梯度尽可能为零．许多后续研究沿用或者借鉴了 IRM 的方法．文献[112]考虑强化学习在多环境中的泛化问题，假设下一状态仅与当前状态构成因果关联并构建因果图，使用 IRM 学习状态摘要表达，然后对接下游任务．文献[133]考虑模型隐私保护问题，证明了因果学习得到的模型相对于相关学习得

到的模型的分布外泛化误差更小，且能够抵抗隐私攻击，方法使用 IRM 实现．文献[113]针对 IRM 所需要的多环境数据的构造问题，提出在没有显式环境类别标注时可以引入辅助的环境推断任务，直接由单一数据集构建多环境子集．

除以上常见方法外，也有研究工作基于不变性学习探索了其他方案．文献[134]针对反因果任务使用因果图建模（如图 9(a)所示），提出 Deep CAMA 方法解决模型鲁棒性问题．除输入变量 X 和输出变量 Y 以外还引入了其他未观测变量，分为可干预的变量 M 及不可干预的变量 Z ．对分解的因果图各部分使用神经网络建模，利用原始训练数据和干预过的数据进行证据下限优化（evidence lower bound optimization）．其中干预数据指的是对变量 M 的 do 操作，具体的干预数值可以通过推断获得，而原始训练数据被视为 $do(M=0)$ 操作下的观测．文献[135]同样针对鲁棒性问题使用因果图建模（如图 9(b)所示），提出潜在因果不变模型（latent causal invariance model, LaCIM），认为不同的域 D 决定了混杂因素 C，进而生成构成输入 X 的 2 组特征 Z 和 S，其中 S 决定了输出结果 Y，而 Z 与 Y 无因果关系．除 D 可变以外，其他因果机制视为不变．该方法利用变分自编码器（variational auto-encoder）将变量 X 和 Y 编码至隐空间，并视为由代表 S 和 Z 的 2 部分组成，两者共同通过解码器重构 X ，同时令 S 通过单独的解码器重构输出变量 Y ．在测试阶段，给定输入 X 后输出 Y 可以由因果图上的推断过程获得．文献[136]认为图像数据由内容 C 和风格 S 共同决定，对于分类任务而言不论风格怎样改变，内容对类别的作用机制是固定不变的，即 $P(Y\mid C,do(S=s))=P(Y\mid C,do(S=s'))$．因此利用大量无监督图像数据，通过旋转、裁剪、灰度调整等风格干预操作构造成对图像，使特征抽取模型所习得的特征表达在成对的图像对实例判别（instance discrimination）任务有相似的预测作用．在无监督预训练之后得到的特征抽取模型可用于下游任务的学习，能够有效提升分布外泛化能力．文献[137]认为在分类任务中相同物体的不同表现应当具有不变的特征表达，在模型优化目标中添加额外的正则项，要求同类别的随机选取的 2 个样本具有较高的匹配程度，其中匹配程度的度量方式通过对比学习（contrastive learning）习得．文献[138]在多实例学习任务中，认为实例集合的标签取决于集合中的某些关键实例，将其称为因果实例（causal instance），且认为利用因果实例判别标签的过程在协变量偏移场景下具有不变性．因此采用 RCM 建模并利用回归估计识别因果实例，然后通过与因果实例进行比对来确定集合判别结果．

基于不变性学习解决模型鲁棒性问题是一种在机器学习中引入因果的自然方式，同时也有较好的发展前景．目前已有工作只是在该领域的初步尝试，需要针对不同任务和数据设定不同的假设，并分别设计求解方案，缺乏统一的方法论的指导，仍有待进一步研究探索．



(a) Deep CAMA因果图　　(b) LaCIM因果图

Fig. 9 Causal graph of invariance-learning methods [134-135]

图 9 不变性学习方法的因果图[134-135]

2.4 公平性问题

机器学习中的公平性（fairness）指的是，对于特定的敏感特征如性别、年龄、种族等，不同的取值不应该影响某些任务中机器学习模型的预测结果，如贷款发放、法律判决、招生招聘等．公平性对于机器学习在社会决策中的应用是十分重要的考虑因素，与因果有密切的关系，直观上体现为敏感特征不应成为预测结果的因变量．模型中存在的不公平常常由伪相关特征问题导致，因此公平性也可以视为针对敏感特征的鲁棒性，但有着自己独有的术语和研究体系．下面首先介绍一下公平性的基本概念，然后介绍因果理论在公平性问题中的应用．

公平性的定义和度量指标目前十分多样化，并没有完全统一确定，不同的定义所反映的问题也有所不同，甚至可能是相互不兼容的[139]．为便于表述，记敏感特征为 A，其他观测特征为 $\boldsymbol{X}$，真实输出结果为 Y，模型为 f，模型预测结果为 $\hat{Y}=f(A,\boldsymbol{X})$（本节所用符号与前文无关）．早期公平性问题的相关工

作并没有考虑因果，最简单直白的方式是在决策时避免使用敏感特征[140]，即 $f(A,\boldsymbol{X})=f(\boldsymbol{X})$．然而这一方案显然是不够的，因为其他特征中也可能会包含敏感特征的信息．因此一般会考虑个体级别的公平性或者群体级别的公平性的度量，并设计方法实现．个体公平性（individual fairness）通常会限制相似的个体之间应该有相似的预测结果[141]，难点在于相似性指标的设计；群体公平性（group fairness）会定义不同的群体并设置度量指标使得各个群体之间差异尽可能小，一种思路是人群平等（demographic parity）[142]，希望在不同敏感特征取值的群体中预测结果的分布一致，即 $P(\hat{Y}\mid A=0)=P(\hat{Y}\mid A=1)$；另一种思路是机会均等（equality of opportunity）[143]，希望在那些本该有机会的人群所获得的机会不受敏感特征的影响，即 $P(\hat{Y}\mid A=0,Y=1)=P(\hat{Y}\mid A=1,Y=1)$；还有一种思路是条件公平（conditional fairness）[144]，希望在任意公平变量 F 条件下不同敏感特征群体的结果一致，即 $P(\hat{Y}\mid A=0,F)=P(\hat{Y}\mid A=1,F)$．这些定义并不考虑特征内部的依赖关系，对模型的决策机制也没有区分性，在更细致的公平性分析中难以满足要求．因果理论的引入为公平性研究起到了极大地推动作用，许多概念必须借助因果的语言才能表达．

Table 4 Application of Causal Methods on Fairness Problems

表 4 因果方法在公平性问题上的应用

分类	典型思路和方法
反事实公平性度量	提出基于反事实的个体公平性指标[145-152]
公平模型构建	利用先验因果图指导模型的公平化构建[153-155]

较早引入因果的公平性研究工作是反事实公平性（counterfactual fairness）[145]．这里的反事实指的是，仅仅改变个体的敏感特征而保持其他特征不变，包括未观测的特征．反事实公平性指的是对任何个体的反事实操作都不应当影响其预测结果，即 $P(\hat{Y}_{A=a}\mid \boldsymbol{X}=\boldsymbol{x},A=a)=P(\hat{Y}_{A=a'}\mid \boldsymbol{X}=\boldsymbol{x},A=a)$．这种定义避免了个体公平性相似性指标设计困难的问题，同时相对于群体公平性又有更高的要求．具体实现公平性的方法通常是利用数据推断未观测变量作为数据增强，或者避免使用敏感特征及其在因果图上的后继节点作为模型输入．反事实公平性的一个重要研究内容是特定路径（path-specific）上的反事实公平性，即考虑在因果图上从敏感特征到预测结果的不同路径，造成直接影响的路径会引发公平性，而间接影响的路径则未必引发不公．文献[146]针对这一问题提出“未解决的歧视”（unresolved discrimination）概念，指出任何基于观测的标准均无法判断模型是否表现出未解决的歧视问题，通过施加干预不变性的约束可以解决这一问题．文献[147]同样针对特定路径的反事实公平性，将问题转化为约束优化问题，使用逆概率加权 IPW 方法求解．文献[148]针对前面 2 项工作容易丢失个性化信息的问题，提出一种基于隐变量的方法修正在不公平路径上敏感变量的后继节点的观测．文献[149]延续前者的研究内容，将个体级别的讨论拓展到子群体的级别，并提出方法解决这一框架下的可识别性问题．文献[150]研究了反事实公平性中的可识别性问题，即反事实结果是否能够通过观测数据获得唯一解．该工作指出当且仅当敏感特征后继和预测结果的祖先存在交集时不可识别，这种情况下虽无法确定唯一解，但可以计算上下界．文献[151]研究了文本数据中国家、职业、姓名等敏感特征对情感预测的反事实公平性问题．文献[152]在文本分类任务中提出反事实符号公平性（counterfactual token fairness）新概念，即针对敏感词的反事实公平性，提出敏感词替换和反事实逻辑匹配的方法解决该问题．

除了反事实公平性，一些工作也从其他角度引入了因果技术．文献[153]研究了多模型级联构成的决策系统的公平性问题，如果单个模型存在不公平问题则会导致整体不公平，但逐个模型进行处理的效率太低，因此将整体系统建模为 SCM，视单个模型的调整为因果图上的软干预，从而实现全局的高效求解．文献[154]考虑多步决策中存在数据缺失时的公平性问题，使用因果图建模这一问题，并提出一种去中心化的方法避免公平性算法依赖那些缺失后无法恢

复的信息. 文献[155]考虑从构建数据集的角度实现公平性，利用生成对抗网络建模在敏感特征受到干预时的生成机制，通过控制干预下的样本生成过程，消除数据中的不公平因素. 文献[156]基于 RCM 提出 2 种新的群体公平性指标：平均因果效应公平（FACE）和处理组平均因果效应公平（FACT），相当于反事实公平性在群体上的平均度量，使用倾向性得分方法 IPTW 进行因果效应估计. 文献[157]尝试基于中介分析将全局变化量（total variation, TV）拆解成多个细粒度的度量，包括反事实直接效应、反事实间接效应和伪效应. 文献[158]延续前者的工作，将方法拓展到 TV 以外的其他度量指标，并给出了这类问题的求解方法. 文献[159]针对达到公平性需要重新训练模型的问题，提出一种基于样本加权的反事实分布修正方法，可以避免重新训练的开销. 文献[160]考虑动态系统中的公平性度量问题，将动态系统建模为 SCM，则公平性度量就成了因果效应估计问题，使用双稳健估计 DRE 方法进行处理.

目前针对机器学习公平性问题的研究已经与因果密切相关，包括描述语言、建模方法和求解手段都在一定程度上依赖因果研究的相关成果，预计未来因果理论在这一方向将持续起到不可替代的作用.

2.5 反事实评估问题

反事实评估（counterfactual evaluation）指的是机器学习模型的优化目标本身是反事实的，这通常出现在使用有偏差的标注数据训练得到无偏模型的情景，例如基于点击数据的检索和推荐系统学习任务. 由于任务本身需要反事实术语进行表述，因果理论对这类问题的建模和研究起到了关键性的作用.

Table 5 Application of Causal Methods on Counterfactual Evaluation Problems

表 5 因果方法在反事实评估问题上的应用

分类	典型思路和方法
推荐系统非随机缺失问题求解	利用倾向性得分或者反事实风险最小化方法修正策略效用[161-172]
检索系统位置偏差问题求解	利用倾向性得分方法修正相关性[173-179]

以推荐系统为例，这类任务的目的是根据用户的意图和喜好向用户展示相关性更高的物品，如文档、商品及广告等. 由于难以获得物品真实相关性的人工标注，实际应用中通常会使用用户的点击（click）数据指导模型学习. 然而系统每次只能向用户展示部分物品，这就使得未展示物品无法被估计相关性，从而对系统策略的评估带来偏差. 考虑假设所有物品都被展示的情况下的点击率即属于反事实评估问题. 由于未展示物品是由系统策略决定而非完全随机，其点击数据的缺失也是非随机的，因此也被称为非随机缺失（missing-not-at-random, MNAR）问题. 文献[180]首次在广告推荐系统中使用因果图建模这类问题（如图 10 所示），并指出这种情况下的系统评估是反事实的.

这种情况需要估计物品是否被观测这一变量对是否被点击这一变量的因果效应，可以用 RCM 建模，使用逆倾向性得分（inverse propensity scoring, IPS）方法修正偏差. 这里的倾向性得分指的是物品被观测到的概率，用得分的倒数作为权重为训练样本加权，即可消除偏差. 记用户特征为 $\boldsymbol{X}$，物品特征为 $\boldsymbol{Y}$，向用户展示物品的策略为 h，倾向性得分为 P，是否被点击为 δ（本节所用符号与前文无关），则对于展示策略效用的无偏 IPS 估计为

$$\hat{R}(h)=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta^{(i)}\cdot h(\boldsymbol{y}^{(i)}\mid\boldsymbol{x}^{(i)})}{p^{(i)}}. \tag{7}$$

文献[161]称这类任务为从 Bandit 反馈日志中批量学习（batch learning from logged bandit feedback, BLBF），在 IPS 的基础上额外采用权重裁剪和方差正则，提出一种新方法称为反事实风险最小化（counterfactual risk minimization, CRM）：

$$\hat{h}_{\mathrm{CRM}}=\arg\min_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}-\delta^{(i)}\cdot\min\left\{M,\frac{h(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)})}{p^{(i)}}\right\}+\lambda\sqrt{\frac{Var(u_h)}{n}}\right\}. \tag{8}$$

其中 M 为权重裁剪参数，λ 为方差正则权重，u_h 为带权重裁剪的 IPS 估计值.

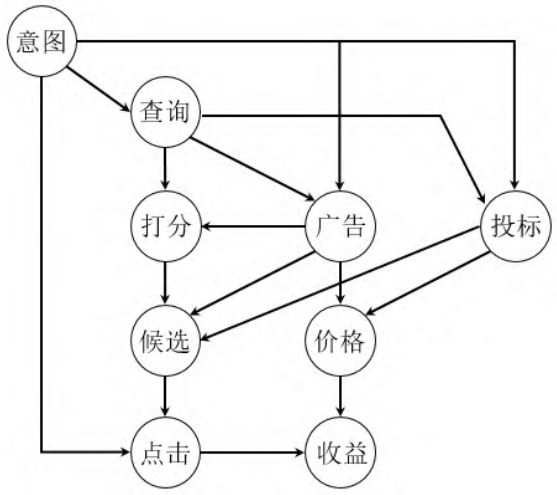


Fig. 10 Causal graph in recommendation systems [180]

图 10 广告推荐系统的因果图[180]

大量工作延续 IPS 和 CRM 方法展开研究. 文献[162]指出 CRM 中存在倾向性过拟合问题，提出自归一化估计方法解决. 文献[163]指出 CRM 中的方差估计需要遍历整个训练集导致计算开销大，提出一种变分散度最小化方法解决. 文献[164]从贝叶斯视角重新分析 CRM，提出一种更易实现的新正则化方法. 文献[165]针对系统行为空间极大的情况提出分布鲁棒的 CRM 算法. 文献[166]将 IPS 推广至更广泛的评价指标，并针对推荐系统任务提出倾向性得分的估计方法. 文献[167]将 IPS 拓展至隐反馈问题，即只有点击记录而没有未点击记录的情况. 文献[168] 额外考虑推荐系统的使用会改变未来用户行为的问题，基于 IPS 提出一种因果嵌入表达方法. 文献[169]提出在 IPS 中使用估计的倾向性得分要比真实的倾向性的分获得更低的方差. 文献[170]考虑推荐的集合被捆绑为一个整体同时选择推荐或者不推荐，因此原有 IPS 作为针对单个物品的方法在此并不适用，提出一种变分样本加权的方法来解决. 文献[171]指出由于展示物品只是整体的一部分，因此整体系统的展示机制不可识别，提出一种对抗学习的方案改进 IPS. 文献[172]考虑 BLBF 问题中反馈信息存在序结构的情况，由于 CRM 方法无法利用结构信息，因此提出一种基于域适应的算法进行求解. BLBF 这一建模框架也被用于推荐系统以外任务，如文献[181]在语义解析任务中引入人工反馈信号，同样使用 IPS 方法进行反事实评估.

检索系统中也会面临类似的 MNAR 问题，这类应用需要对展示物品进行排序，与用户需求相关性更高的物品应当排在更靠前的位置. 这种情况下用户选择物品的点击行为会受物品列表的展示位置影响，位置越靠后则越不容易被用户观测到，进而使得点击率也偏低，因此这一问题也被称为位置偏差（position bias）问题. 文献[173]指出了这一问题，并使用 IPS 方法进行处理，其中倾向性得分指的是物品在当前位置被观测到的概率. 一个关键的问题是如何估计倾向性得分，一般需要在线上系统中单独收集数据进行估计. 文献[174]指出直接采用随机策略估计倾向性得分会影响用户体验，提出一种期望最大化算法避免随机策略；文献[175]指出倾向性得分估计和消除位置偏差的任务互为对偶任务，提出对偶学习方法同时学习 2 个模型. 一些工作也考虑对 IPS 进行改进，文献[176]对 IPS 方法进行扩展，能够适配一般的加性排序指标和非线性模型；文献[177]针对 IPS 稳定性问题，提出使用采样代替加权的做法，使训练更稳定；文献[178]将 IPS 由点击模型推广到级联模型. 文献[179]对比了 IPS 和在线学习方法，指出 IPS 在偏差和噪声较小的情况下优于在线学习方法.

反事实评估问题在检索和推荐系统中的技术已经相对成熟和固化，许多文献除了沿用 IPS 和 CRM 的概念以外未必会使用额外的因果术语进行表述，但这并不影响因果理论在其中的根基作用. 未来如果出现其他需要使用反事实评估的场景，也可以继续通过因果分析与已有技术快速建立联系.

2.6 其他问题

因果机器学习的研究工作种类十分丰富，除了在可解释性、可迁移性、鲁棒性、公平性和反事实评估这些主要问题上的研究以外，还有部分其他方面的研究. 以下选择其中值得关注的部分工作进行简要介绍.

因果理论在一些需要建模变量间结构信息的情况下十分有效. 文献[182-184]研究多臂老虎机问题（multi-armed bandit）中变量存在因果结构的情况，称为结构老虎机（structured bandit）问题，指出忽略因果结构可能导致次优的解，并设计各类方法求解. 文献[185]指出模仿学习中忽略因果关系会导致错误识别问题，即更多的学习数据反而导致性能下降. 因此将变量组织成图结构，随机连接一些节点，依据性能表现学习背后的真实因果图. 文献[186]研究

机器学习任务在特征和输出之间存在因果图的情况下，可以在预测的同时进行因果发现，作为一种正则化手段可以使回归任务更准确. 文献[187]研究特征之间存在结构关系的解耦表达任务，在模型中设计SCM层结构，借助对特征的额外标注学习特征表达，得到的解耦模型可实现干预或反事实下的生成.

因果理论中的反事实思想和技术为多个领域的问题提供了求解思路. 文献[188]在不完全信息博弈问题中，使用反事实的思想设计了反事实遗憾最小化（counterfactual regret minimization）算法，已成为求解该问题的重要方法[189-192]，其中反事实指的是将当前策略替换为最优策略会带来多大改进. 文献[193]研究强化学习中多类别分布下的 SCM 可识别性问题，通过选择风险最高的反事实轨迹，提供离线策略评估方案，为专家提供诊断建议. 文献[194]研究离线强化学习问题，将原有的在自身策略下的探索改为基于日志的反事实探索，获得性能提升. 文献[195]研究使用 Actor-Critic 方法训练场景图生成模型，提出在 Critic 模型中使用反事实结果作为基线可以提升生成效果. 文献[196]研究对话生成任务，借助已经生成的回复文本来构建反事实回复文本，获得更好的生成质量. 文献[197]在文本分类任务的注意力监督方法中，使用反事实推理替代人工标注，得到了优于人工标注的结果. 文献[198]在弱监督视觉语言举证（vision-language grounding）任务中提出一种反事实对比学习方法提升了举证效果.

因果机器学习本身也提出了更高层级的问题，即干预和反事实结果预测问题，这需要机器学习和因果推断 2 个领域的协作才能完成. 文献[199-200]分别基于生成对抗网络和变分自编码器实现干预和反事实下的图像生成能力. 文献[201-202]分别在文本生成领域提出和求解反事实故事重写任务. 文献[203]在文本生成任务中根据不同的属性要求针对已有文本生成不同的反事实文本. 文献[204]尝试解决多智能体任务中针对环境改变的反事实提问. 文献[205]在 3D 物理引擎世界中预测改变初态后的反事实未来发展.

3 总结与展望

本文介绍了因果相关的概念、模型和方法，并着重对因果机器学习在各类问题上的前沿研究工作展开详细介绍，包括可解释性问题、可迁移性问题、鲁棒性问题、公平性问题和反事实评估问题等. 从现有的应用方式来看，因果理论对于机器学习的帮助在不同的问题上具有不同的表现，包括建模数据内部结构、表达不变性假设、引入反事实概念和提供效应估计手段等，这在缺少因果术语和方法的时代是难以实现的. 有了因果理论的帮助，机器学习甚至可以探讨过去无法讨论的问题，如干预和反事实操作下的预测问题.

对于可解释性、公平性和反事实评估问题，因果理论和方法已成为描述和求解问题所不可缺少的一部分，且应用方式也渐趋成熟. 这是由于对特征的重要程度的估计、对模型公平性的度量和对反事实策略效用的评估均属于因果效应估计的范畴，问题本身需要使用因果的术语才能得到清晰且完整的表达，因果推断的相关方法自然也可以用于问题的求解. 可以预见，未来这些问题将继续作为因果理论和方法的重要应用场景，伴随因果推断技术的发展，向着更加准确和高效的目标前进.

对于可迁移性和鲁棒性问题，目前所采用的因果相关方法大多还处于较浅的层次，有待深入挖掘探索. 在这些问题上，因果推断的相关技术不易直接得到应用，这是由于这类问题的目标不再是单纯估计因果效应或者发现因果结构，而是需要识别跨环境不变的机制. 这对于因果而言是一项全新的任务，需要研究新的方法来求解. 在机器学习尤其是深度学习中，这项任务的主要难点在于数据的高维复杂性. 对于图像和文本等数据而言，其显式特征高度耦合，难以从中提取出有效的因果变量，阻碍了效应估计和结构发现等后续分析手段. 目前所采用的反因果迁移、反事实数据增强和因果效应校准等手段大多只能针对可观测的已知变量进行处理，适用范围受到很大限制. 相对的，不变性学习有能力处理未知的伪相关特征并识别因果特征，具有良好的发展前景. 然而目前的不变性学习方法也存在局限性，主要在于对数据做了较强的因果结构假设，一方面数据可能无法满足假设而又缺少验证假设的手段，另一方面需要为满足不同假设的数据设计不同的方法而缺乏通用性. 因此，未来在这些方向上都值得开展研究. 一种思路是继续针对具体任务做出不同的因果结构假设，并设计对应的学习算法，这就需要构建成体系的解决方案并配备验证假设的手段；另一种思路是从数据本身出发，推断和发现潜在的因果结构，这就需要研究全新的方法来突破由数据的高维复杂性带来的障碍.

从因果机器学习的研究进展来看，机器学习领域的因果革命将大有可为. 不可否认，当前正处于因果

革命的起步阶段，由于现实问题存在极高的复杂性，这一革命的历程也将曲折而艰辛，需要更多的研究和支持. 希望更多的研究者能够加入到因果机器学习的研究中来，共同创造和见证因果革命的新时代.

参 考 文 献

[1] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444

[2] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 1026-1034

[3] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis [C/OL] //Proc of the 7th Int Conf on Learning Representations. 2019[2021-11-03]. https://openreview.net/pdf?id=B1xsqj09Fm

[4] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 1877-1901

[5] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search [J]. Nature, 2016, 529(7587): 484-489

[6] Senior A W, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning [J]. Nature, 2020, 577(7792): 706-710

[7] Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program [J]. AI Magazine, 2019, 40(2): 44-58

[8] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [C/OL] //Proc of the 2nd Int Conf on Learning Representations. 2014[2021-11-03]. https://arxiv.org/abs/1312.6199

[9] Barocas S, Hardt M, Narayanan A. Fairness in machine learning [EB/OL]. 2017[2021-11-13]. https://fairmlbook.org/pdf/fairmlbook.pdf

[10] Pearl J, Mackenzie D. The Book of Why: The New Science of Cause and Effect [M]. New York: Basic Books, 2018

[11] Pearl J. Theoretical impediments to machine learning with seven sparks from the causal revolution [J]. arXiv preprint, arXiv:1801.04016, 2018

[12] Miao Wang, Liu Chunchen, Geng Zhi. Statistical approaches for causal inference [J]. Scientia Sinica Mathematica, 2018, 48(12):1753-1778(in Chinese)
(苗旺, 刘春辰, 耿直. 因果推断的统计方法[J]. 中国科学: 数学, 2018, 48(12): 1753-1778)

[13] Guo Ruocheng, Cheng Lu, Li Jundong, et al. A survey of learning causality with data: Problems and methods [J]. ACM Computing Surveys (CSUR), 2020, 53(4): 1-37

[14] Yao Liuyi, Chu Zhixuan, Li Sheng, et al. A survey on causal inference [J]. arXiv preprint, arXiv:2002.02770, 2020

[15] Schölkopf B. Causality for machine learning [J]. arXiv preprint, arXiv:1911.10500, 2019

[16] Schölkopf B, Locatello F, Bauer S, et al. Toward causal representation learning [J]. Proceedings of the IEEE, 2021, 109(5): 612-634

[17] Splawa-Neyman J, Dabrowska D M, Speed T P. On the application of probability theory to agricultural experiments. Essay on principles. Section 9 [J]. Statistical Science, 1990, 5(4): 465-472

[18] Rubin D B. Estimating causal effects of treatments in randomized and nonrandomized studies [J]. Journal of Educational Psychology, 1974, 66(5): 688-701

[19] Pearl J. Causality [M]. Cambridge, UK: Cambridge University Press, 2009

[20] Granger C W J. Investigating causal relations by econometric models and cross-spectral methods [J]. Econometrica, 1969, 37(3): 424-438

[21] Rubin D B. Randomization analysis of experimental data: The Fisher randomization test comment [J]. Journal of the American Statistical Association, 1980, 75(371): 591-593

[22] Rosenbaum P R, Rubin D B. The central role of the propensity score in observational studies for causal effects [J]. Biometrika, 1983, 70(1): 41-55

[23] Hirano K, Imbens G W, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score [J]. Econometrica, 2003, 71(4): 1161-1189

[24] Robins J M, Rotnitzky A, Zhao Lueping. Estimation of regression coefficients when some regressors are not always observed [J]. Journal of the American Statistical Association, 1994, 89(427): 846-866

[25] Dudík M, Langford J, Li Lihong. Doubly robust policy evaluation and learning [C] //Proc of the 28th Int Conf on Machine Learning. Madison, WI: Omnipress, 2011: 1097-1104

[26] Kuang Kun, Cui Peng, Li Bo, et al. Estimating treatment effect in the wild via differentiated confounder balancing [C] //Proc of the 23rd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2017: 265-274

[27] Imbens G W, Rubin D B. Causal Inference in Statistics, Social, and Biomedical Sciences [M]. Cambridge, UK: Cambridge University Press, 2015

[28] Yao Liuyi, Chu Zhixuan, Li Sheng, et al. A survey on causal inference [J]. arXiv preprint, arXiv:2002.02770, 2020

[29] Pearl J. Causal diagrams for empirical research [J]. Biometrika, 1995, 82(4): 669-688

[30] Spirtes P, Glymour C. An algorithm for fast recovery of sparse causal graphs [J]. Social Science Computer Review, 1991, 9(1): 62-72

[31] Verma T, Pearl J. Equivalence and synthesis of causal models [C] //Proc of the 6th Annual Conf on Uncertainty in Artificial Intelligence. Amsterdam: Elsevier, 1990: 255-270

[32] Spirtes P, Glymour C N, Scheines R, et al. Causation, Prediction, and Search [M]. Cambridge, MA: MIT Press, 2000

[33] Schwarz G. Estimating the dimension of a model [J]. The Annals of Statistics, 1978, 6(2): 461-464

[34] Chickering D M. Optimal structure identification with greedy search [J]. Journal of Machine Learning Research, 2002, 3(Nov): 507-554

[35] Shimizu S, Hoyer P O, Hyvärinen A, et al. A linear non-Gaussian acyclic model for causal discovery [J]. Journal of Machine Learning Research, 2006, 7(10): 2003−2030

[36] Zhang Kun, Hyvärinen A. On the identifiability of the post-nonlinear causal model [C] //Proc of the 25th Conf on Uncertainty in Artificial Intelligence. Arlington, VA: AUAI Press, 2009: 647-655

[37] Pearl J. Direct and indirect effects [C] //Proc of the 17th Conf on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers Inc, 2001: 411-420

[38] VanderWeele T. Explanation in Causal Inference: Methods for Mediation and Interaction [M]. Oxford, UK: Oxford University Press, 2015

[39] Chen Kerui, Meng Xiaofeng. Interpretation and understanding in machine learning [J]. Journal of Computer Research and Development, 2020, 57(9): 1971-1986(in Chinese)
(陈珂锐, 孟小峰. 机器学习的可解释性[J]. 计算机研究与发展, 2020, 57(9): 1971-1986)

[40] Ribeiro M T, Singh S, Guestrin C. " Why should I trust you?" Explaining the predictions of any classifier [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135-1144

[41] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 618-626

[42] Sundararajan M, Taly A, Yan Qiqi. Axiomatic attribution for deep networks [C] //Proc of the 34th Int Conf on Machine Learning. Cambridge MA: JMLR, 2017: 3319-3328

[43] Lundberg S M, Lee S I. A unified approach to interpreting model predictions [C] //Proc of the 31st Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2017: 4765-4774

[44] Alvarez-Melis D, Jaakkola T. A causal framework for explaining the predictions of black-box sequence-to-sequence models [C] //Proc of the 2017 Conf on Empirical Methods in Natural Language Processing. Stroudsburg,PA: ACL, 2017: 412-421

[45] Schwab P, Karlen W. CXPlain: Causal explanations for model interpretation under uncertainty [C] //Proc of the 33rd Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2019: 10220-10230

[46] Chattopadhyay A, Manupriya P, Sarkar A, et al. Neural network attributions: A causal perspective [C] //Proc of the 36th Int Conf on Machine Learning. Cambridge MA: JMLR, 2019: 981-990

[47] Frye C, Rowat C, Feige I. Asymmetric Shapley values: Incorporating causal knowledge into model-agnostic explainability [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 1229-1239

[48] Heskes T, Sijben E, Bucur I G, et al. Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 4778-4789

[49] Goyal Y, Wu Ziyan, Ernst J, et al. Counterfactual visual explanations [C] //Proc of the 36th Int Conf on Machine Learning. Cambridge MA: JMLR, 2019: 2376-2384

[50] Wang P, Vasconcelos N. SCOUT: Self-aware discriminant counterfactual explanations [C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 8981-8990

[51] Hendricks L A, Hu Ronghang, Darrell T, et al. Generating counterfactual explanations with natural language [J]. arXiv preprint, arXiv:1806.09809, 2018

[52] Chang Chunhao, Creager E, Goldenberg A, et al. Explaining image classifiers by counterfactual generation [C/OL] //Proc of the 7th Int Conf on Learning Representations, 2019[2021-11-03]. https://openreview.net/pdf?id=B1MXz20cYQ

[53] Kanehira A, Takemoto K, Inayoshi S, et al. Multimodal explanations by predicting counterfactuality in videos [C] //Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8594-8602

[54] Akula A R, Wang Shuai, Zhu Songchun. CoCoX: Generating conceptual and counterfactual explanations via fault-lines [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 2594-2601

[55] Madumal P, Miller T, Sonenberg L, et al. Explainable reinforcement learning through a causal lens [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 2493-2500

[56] Mothilal R K, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations [C] //Proc of the 2020 Conf on Fairness, Accountability, and Transparency. New York: ACM, 2020: 607-617

[57] Albini E, Rago A, Baroni P, et al. Relation-based counterfactual explanations for Bayesian network classifiers [C] //Proc of the 29th Int Joint Conf on Artificial Intelligence, Red Hook, NY: Curran Associates Inc, 2020: 451-457

[58] Kenny E M, Keane M T. On generating plausible counterfactual and semi-factual explanations for deep learning [C] //Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 11575-11585

[59] Abrate C, Bonchi F. Counterfactual graphs for explainable classification of brain networks [J]. arXiv preprint, arXiv:2106.08640, 2021

[60] Yang Fan, Alva S S, Chen J, et al. Model-based counterfactual synthesizer for interpretation [J]. arXiv preprint, arXiv:2106.08971, 2021

[61] Parmentier A, Vidal T. Optimal counterfactual explanations in tree

ensembles [J]. arXiv preprint, arXiv:2106.06631, 2021

[62] Besserve M, Mehrjou A, Sun R, et al. Counterfactuals uncover the modular structure of deep generative models [C/OL] //Proc of the 8th Int Conf on Learning Representations. 2020[2021-11-03]. https://openreview.net/pdf?id=SJxDDpEKvH

[63] Kanamori K, Takagi T, Kobayashi K, et al. DACE: Distribution-aware counterfactual explanation by mixed-integer linear optimization [C] //Proc of the 19th Int Joint Conf on Artificial Intelligence. Red Hook, NY: Curran Associates Inc, 2020: 2855-2862

[64] Kanamori K, Takagi T, Kobayashi K, et al. Ordered counterfactual explanation by mixed-integer linear optimization [C] //Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 11564-11574

[65] Tsirtsis S, Gomez-Rodriguez M. Decisions, counterfactual explanations and strategic behavior [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 16749-16760

[66] Karimi A H, von Kügelgen B J, Schölkopf B, et al. Algorithmic recourse under imperfect causal knowledge: A probabilistic approach [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 265-277

[67] Schölkopf B, Janzing D, Peters J, et al. On causal and anticausal learning [C] //Proc of the 29th Int Conf on Machine Learning. Madison, WI: Omnipress, 2012: 459-466

[68] Zhang Kun, Schölkopf B, Muandet K, et al. Domain adaptation under target and conditional shift [C] //Proc of the 30th Int Conf on Machine Learning. Cambridge MA: JMLR, 2013: 819-827

[69] Rojas-Carulla M, Schölkopf B, Turner R, et al. Invariant models for causal transfer learning [J]. The Journal of Machine Learning Research, 2018, 19(1): 1309-1342

[70] Guo Jiaxian, Gong Mingming, Liu Tongliang, et al. LTF: A label transformation framework for correcting target shift [C] //Proc of the 37th Int Conf on Machine Learning. Cambridge MA: JMLR, 2020: 3843-3853

[71] Cai Ruichu, Li Zijian, Wei Pengfei, et al. Learning disentangled semantic representation for domain adaptation [C] //Proc of the 28th Int Joint Conf on Artificial Intelligence, Red Hook, NY: Curran Associates Inc, 2019: 2060-2066

[72] Gong Mingming, Zhang Kun, Liu Tongliang, et al. Domain adaptation with conditional transferable components [C] //Proc of the 33rd Int Conf on Machine Learning. Cambridge MA: JMLR, 2016: 2839-2848

[73] Teshima T, Sato I, Sugiyama M. Few-shot domain adaptation by causal mechanism transfer [C] //Proc of the 37th Int Conf on Machine Learning. Cambridge MA: JMLR, 2020: 9458-9469

[74] Edmonds M, Ma Xiaojian, Qi Siyuan, et al. Theory-based causal transfer: Integrating instance-level induction and abstract-level structure learning [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 1283-1291

[75] Etesami J, Geiger P. Causal transfer for imitation learning and decision making under sensor-shift [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 10118-10125

[76] Yue Zhongqi, Zhang Hanwang, Sun Qianru, et al. Interventional few-shot learning [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 2734-2746

[77] Zhang Kun, Gong Mingming, Stojanov P, et al. Domain adaptation as a problem of inference on graphical models [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 4965-4976

[78] Zhang Kun, Gong Mingming, Schölkopf B. Multi-source domain adaptation: A causal view [C] //Proc of the 29th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2015: 3150-3157

[79] Bagnell J A. Robust supervised learning [C] //Proc of the 20th National Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2005: 714-719

[80] Hu Weihua, Niu Gang, Sato I, et al. Does distributionally robust supervised learning give robust classifiers? [C] //Proc of the 35th Int Conf on Machine Learning. Cambridge MA: JMLR, 2018: 2029-2037

[81] Rahimian H, Mehrotra S. Distributionally robust optimization: A review [J]. arXiv preprint, arXiv:1908.05659, 2019

[82] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [C/OL] //Proc of the 5th Int Conf on Learning Representations. 2017[2021-11-14]. https://openreview.net/pdf?id=B1xsqj09Fm

[83] Xu Han, Ma Yao, Liu Haochen, et al. Adversarial attacks and defenses in images, graphs and text: A review [J]. International Journal of Automation and Computing, 2020, 17(2): 151-178

[84] Gururangan S, Swayamdipta S, Levy O, et al. Annotation artifacts in natural language inference data [C] //Proc of the 16th Conf of the North American Chapter of the ACL: Human Language Technologies, Vol 2. Stroudsburg, PA: ACL, 2018: 107-112

[85] Zhang Guanhua, Bai Bing, Liang Jian, et al. Selection bias explorations and debias methods for natural language sentence matching datasets [C] //Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 4418-4429

[86] Clark C, Yatskar M, Zettlemoyer L. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases [C] //Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: ACL, 2019: 4060-4073

[87] Cadene R, Dancette C, Cord M, et al. Rubi: Reducing unimodal biases for visual question answering [C] //Proc of the 33rd Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2019: 841-852

[88] Lu Kaiji, Mardziel P, Wu Fangjing, et al. Gender bias in neural natural

language processing [G] //LNCS 12300: Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday. Berlin: Springer, 2020: 189-202

[89] Maudslay R H, Gonen H, Cotterell R, et al. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution [C] //Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: ACL, 2019: 5270-5278

[90] Zmigrod R, Mielke S J, Wallach H, et al. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology [C] //Proc of the 57th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2019: 1651-1661

[91] Kaushik D, Hovy E, Lipton Z. Learning the difference that makes a difference with counterfactually-augmented data [C/OL] //Proc of the 8th Int Conf on Learning Representations. 2020[2021-11-14]. https://openreview.net/pdf?id=Sklgs0NFvr

[92] Agarwal V, Shetty R, Fritz M. Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing [C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 9690-9698

[93] Chang Chunhao, Adam G A, Goldenberg A. Towards robust classification model by counterfactual and invariant data generation [C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 15212-15221

[94] Wang Zhao, Culotta A. Robustness to spurious correlations in text classification via automatically generated counterfactuals [C] //Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 14024-14031

[95] Chen Long, Yan Xin, Xiao Jun, et al. Counterfactual samples synthesizing for robust visual question answering [C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10800-10809

[96] Wu Yiquan, Kuang Kun, Zhang Yating, et al. De-biased court's view generation with causality [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 763-780

[97] Qi Jia, Niu Yulei, Huang Jianqiang, et al. Two causal principles for improving visual dialog [C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10860-10869

[98] Wang Tan, Huang Jiangqiang, Zhang Hanwang, et al. Visual commonsense R-CNN [C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10760-10770

[99] Zhang Dong, Zhang Hanwang, Tang Jinhui, et al. Causal intervention for weakly-supervised semantic segmentation [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 655-666

[100] Tang Kaihua, Niu Yulei, Huang Jianqiang, et al. Unbiased scene graph generation from biased training [C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 3716-3725

[101] Tang Kaihua, Huang Jianqiang, Zhang Hanwang. Long-tailed classification by keeping the good and removing the bad momentum causal effect [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 1513-1524

[102] Niu Yulei, Tang Kaihua, Zhang Hanwang, et al. Counterfactual VQA: A cause-effect look at language bias [C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 12700-12710

[103] Kuang Kun, Cui Peng, Athey S, et al. Stable prediction across unknown environments [C]//Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 1617-1626

[104] Shen Zheyan, Cui Peng, Kuang Kun, et al. Causally regularized learning with agnostic data selection bias [C] //Proc of the 26th ACM Int Conf on Multimedia. New York: ACM, 2018: 411-419

[105] Kuang Kun, Xiong Ruoxuan, Cui Peng, et al. Stable prediction with model misspecification and agnostic distribution shift [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 4485-4492

[106] Shen Zheyan, Cui Peng, Zhang Tong, et al. Stable learning via sample reweighting [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 5692-5699

[107] Zhang Xingxuan, Cui Peng, Xu Renzhe, et al. Deep stable learning for out-of-distribution generalization [J]. arXiv preprint, arXiv:2104.07876, 2021

[108] Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: Identification and confidence intervals [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2016, 78(5): 947-1012

[109] Christina H D, Nicolai M, Jonas P. Invariant causal prediction for nonlinear models [J/OL]. Journal of Causal Inference, 2018, 6(2):20170016[2021-11-15]. https://www.degruyter.com/document/doi/10.1515/jci-2017-0016/pdf

[110] Pfister N, Bühlmann P, Peters J. Invariant causal prediction for sequential data [J]. Journal of the American Statistical Association, 2019, 114(527): 1264-1276

[111] Arjovsky M, Bottou L, Gulrajani I, et al. Invariant risk minimization [J]. arXiv preprint, arXiv:1907.02893, 2019

[112] Zhang A, Lyle C, Sodhani S, et al. Invariant causal prediction for block mdps [C] //Proc of the 37th Int Conf on Machine Learning. Cambridge MA: JMLR, 2020: 11214-11224

[113] Creager E, Jacobsen J H, Zemel R. Environment inference for invariant learning [C] //Proc of the 38th Int Conf on Machine Learning. Cambridge

MA: JMLR, 2021: 2189-2200

[114] Kaushik D, Setlur A, Hovy E H, et al. Explaining the efficacy of counterfactually augmented data [C/OL] //Proc of the 9th Int Conf on Learning Representations. 2021[2021-11-14]. https://openreview.net/pdf?id=HHiiQKWsOcV

[115] Abbasnejad E, Teney D, Parvaneh A, et al. Counterfactual vision and language learning [C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10044-10054

[116] Liang Zujie, Jiang Weitao, Hu Haifeng, et al. Learning to contrast the counterfactual samples for robust visual question answering [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 3285-3292

[117] Teney D, Abbasnedjad E, van den Hengel A. Learning what makes a difference from counterfactual examples and gradient supervision [C] //Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 580-599.

[118] Fu T J, Wang X E, Peterson M F, et al. Counterfactual vision-and-language navigation via adversarial path sampler [C] //Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 71-86

[119] Parvaneh A, Abbasnejad E, Teney D, et al. Counterfactual vision-and-language navigation: Unravelling the unseen [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 5296-5307

[120] Sauer A, Geiger A. Counterfactual generative networks [C/OL] //Proc of the 9th Int Conf on Learning Representations. 2021[2021-11-14]. https://openreview.net/pdf?id=BXewfAYMmJw

[121] Mao Chengzhi, Cha A, Gupta A, et al. Generative interventions for causal learning [C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 3947-3956

[122] Zeng Xiangji, Li Yunliang, Zhai Yuchen, et al. Counterfactual generator: A weakly-supervised method for named entity recognition [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 7270-7280

[123] Fu T J, Wang Xin, Grafton S, et al. Iterative language-based image editing via self-supervised counterfactual reasoning [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 4413-4422

[124] Pitis S, Creager E, Garg A. Counterfactual data augmentation using locally factored dynamics [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 3976-3990

[125] Zhang Junzhe, Kumor D, Bareinboim E. Causal imitation learning with unobserved confounders [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 12263-12274

[126] Coston A, Kennedy E, Chouldechova A. Counterfactual predictions under runtime confounding [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 4150-4162

[127] Atzmon Y, Kreuk F, Shalit U, et al. A causal view of compositional zero-shot recognition [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 1462-1473

[128] Yang Zekun, Feng Juan. A causal inference method for reducing gender bias in word embedding relations [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 9434-9441

[129] Schölkopf B, Hogg D W, Wang Dun, et al. Modeling confounding by half-sibling regression [J]. Proceedings of the National Academy of Sciences, 2016, 113(27): 7391-7398

[130] Shin S, Song K, Jang J H, et al. Neutralizing gender bias in word embedding with latent disentanglement and counterfactual generation [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing: Findings. Stroudsburg, PA: ACL, 2020: 3126-3140

[131] Yang Zekun, Liu Tianlin. Causally denoise word embeddings using half-sibling regression [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 9426-9433

[132] Yang Xu, Zhang Hanwang, Qi Guojin, et al. Causal attention for vision-language tasks [C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 9847-9857

[133] Tople S, Sharma A, Nori A. Alleviating privacy attacks via causal learning [C] //Proc of the 37th Int Conf on Machine Learning. Cambridge MA: JMLR, 2020: 9537-9547

[134] Zhang Cheng, Zhang Kun, Li Yingzhen. A causal view on robustness of neural networks [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 289-301

[135] Sun Xinwei, Wu Botong, Liu Chang, et al. Latent causal invariant model [J]. arXiv preprint, arXiv:2011.02203, 2020

[136] Mitrovic J, McWilliams B, Walker J C, et al. Representation learning via invariant causal mechanisms [C/OL] //Proc of the 9th Int Conf on Learning Representations. 2021[2021-11-14]. https://openreview.net/pdf?id=9p2ekP904Rs

[137] Mahajan D, Tople S, Sharma A. Domain generalization using causal matching [J]. arXiv preprint, arXiv:2006.07500, 2020

[138] Zhang Weijia, Liu Lin, Li Jiuyong. Robust multi-instance learning with stable instances [C] //Proc of the 24th European Conf on Artificial Intelligence. Ohmsha: IOS, 2020: 1682-1689

[139] Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores [J]. arXiv preprint, arXiv:1609.05807, 2016.

[140] Grgic-Hlaca N, Zafar M B, Gummadi K P, et al. The case for process fairness in learning: Feature selection for fair decision making [C/OL] //Symp on Machine Learning and the Law at the 30th Conf on Neural Information Processing Systems. 2016[2021-11-17].

http://www.mlandthelaw.org/papers/grgic.pdf [该研讨会为 NIPS 会议的一部分，但有单独的文集，详见 https://nips.cc/Conferences/2016/Schedule?showEvent=6258]

[141] Dwork C, Hardt M, Pitassi T, et al. Fairness through awareness [C] //Proc of the 3rd Innovations in Theoretical Computer Science Conf. New York: ACM, 2012: 214-226

[142] Calders T, Kamiran F, Pechenizkiy M. Building classifiers with independency constraints [C] //Proc of the 9th IEEE In Conf on Data Mining Workshops. Piscataway, NJ: IEEE, 2009: 13-18

[143] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning [C] //Proc of the 30th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2016: 3315-3323

[144] Xu Renzhe, Cui Peng, Kuang Kun, et al. Algorithmic decision making with conditional fairness [C] //Proc of the 26th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2020: 2125-2135

[145] Kusner M J, Loftus J, Russell C, et al. Counterfactual fairness [C] //Proc of the 31st Int Conf on Neural Information Processing Systems. New York: ACM, 2017: 4066-4076

[146] Kilbertus N, Rojas-Carulla M, Parascandolo G, et al. Avoiding discrimination through causal reasoning [C] //Proc of the 31st Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2017: 656-666

[147] Nabi R, Shpitser I. Fair inference on outcomes [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018:1931-1940

[148] Chiappa S. Path-specific counterfactual fairness [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 7801-7808

[149] Wu Yongkai, Zhang Lu, Wu Xintao, et al. PC-fairness: A unified framework for measuring causality-based fairness [C] //Proc of the 33rd Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2019: 3404-3414

[150] Wu Yongkai, Zhang Lu, Wu Xintao. Counterfactual fairness: Unidentification, bound and algorithm [C] //Proc of the 28th Int Joint Conf on Artificial Intelligence, Red Hook, NY: Curran Associates Inc, 2019: 1438-1444

[151] Huang P S, Zhang Huan, Jiang R, et al. Reducing sentiment bias in language models via counterfactual evaluation [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing: Findings. Stroudsburg, PA: ACL, 2020: 65-83

[152] Garg S, Perot V, Limtiaco N, et al. Counterfactual fairness in text classification through robustness [C] //Proc of the 33rd AAAI/ACM Conf on AI, Ethics, and Society. Menlo Park, CA: AAAI, 2019: 219-226

[153] Hu Yaowei, Wu Yongkai, Zhang Lu, et al. Fair multiple decision making through soft interventions [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 17965-17975

[154] Goel N, Amayuelas A, Deshpande A, et al. The importance of modeling data missingness in algorithmic fairness: A causal perspective [C] //Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021:7564-7573

[155] Xu Depeng, Wu Yongkai, Yuan Shuhan, et al. Achieving causal fairness through generative adversarial networks [C] //Proc of the 28th Int Joint Conf on Artificial Intelligence. Red Hook, NY: Curran Associates Inc, 2019: 1452-1458

[156] Khademi A, Lee S, Foley D, et al. Fairness in algorithmic decision making: An excursion through the lens of causality [C] //Proc of the 28th World Wide Web Conf. New York: ACM, 2019: 2907-2914

[157] Zhang Junzhe, Bareinboim E. Fairness in decision-making—The causal explanation formula [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018:2037-2045

[158] Zhang Junzhe, Bareinboim E. Equality of opportunity in classification: A causal approach [C] //Proc of the 32nd Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2018: 3671-3681

[159] Wang Hao, Ustun B, Calmon F. Repairing without retraining: Avoiding disparate impact with counterfactual distributions [C] //Proc of the 36th Int Conf on Machine Learning. Cambridge MA: JMLR, 2019: 6618-6627

[160] Creager E, Madras D, Pitassi T, et al. Causal modeling for fairness in dynamical systems [C] //Proc of the 37th Int Conf on Machine Learning. Cambridge MA: JMLR, 2020: 2185-2195

[161] Swaminathan A, Joachims T. Batch learning from logged bandit feedback through counterfactual risk minimization [J]. The Journal of Machine Learning Research, 2015, 16(1): 1731-1755

[162] Swaminathan A, Joachims T. The self-normalized estimator for counterfactual learning [C] //Proc of the 29th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2015: 3231-3239

[163] Wu Hang, Wang May. Variance regularized counterfactual risk minimization via variational divergence minimization [C] //Proc of the 35th Int Conf on Machine Learning. Cambridge MA: JMLR, 2018: 5353-5362

[164] London B, Sandler T. Bayesian counterfactual risk minimization [C] //Proc of the 36th Int Conf on Machine Learning. Cambridge MA: JMLR, 2019: 4125-4133

[165] Faury L, Tanielian U, Dohmatob E, et al. Distributionally robust counterfactual risk minimization [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 3850-3857

[166] Schnabel T, Swaminathan A, Singh A, et al. Recommendations as treatments: Debiasing learning and evaluation [C] //Proc of the 33rd Int Conf on Machine Learning. Cambridge MA: JMLR, 2016: 1670-1679

[167] Yang Longqi, Cui Yin, Xuan Yuan, et al. Unbiased offline recommender

evaluation for missing-not-at-random implicit feedback [C] //Proc of the 12th ACM Conf on Recommender Systems. New York: ACM, 2018: 279-287

[168] Bonner S, Vasile F. Causal embeddings for recommendation [C] //Proc of the 12th ACM Conf on Recommender Systems. New York: ACM, 2018: 104-112

[169] Narita Y, Yasui S, Yata K. Efficient counterfactual learning from bandit feedback [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 4634-4641

[170] Zou Hao, Cui Peng, Li Bo, et al. Counterfactual prediction for bundle treatment [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 19705-19715

[171] Xu Da, Ruan Chuanwei, Korpeoglu E, et al. Adversarial counterfactual learning and evaluation for recommender system [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 13515-13526

[172] Lopez R, Li Chenchen, Yan Xiang, et al. Cost-effective incentive allocation via structured counterfactual inference [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 4997-5004

[173] Joachims T, Swaminathan A, Schnabel T. Unbiased learning-to-rank with biased feedback [C] //Proc of the 10th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2017: 781-789

[174] Wang Xuanhui, Golbandi N, Bendersky M, et al. Position bias estimation for unbiased learning to rank in personal search [C] //Proc of the 11th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2018: 610-618

[175] Ai Qingyao, Bi Keping, Luo Cheng, et al. Unbiased learning to rank with unbiased propensity estimation [C] //Proc of the 41st Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2018: 385-394

[176] Agarwal A, Takatsu K, Zaitsev I, et al. A general framework for counterfactual learning-to-rank [C] //Proc of the 42nd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2019: 5-14

[177] Jagerman R, de Rijke M. Accelerated convergence for counterfactual learning to rank [C] //Proc of the 43rd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2020: 469-478

[178] Vardasbi A, de Rijke M, Markov I. Cascade model-based propensity estimation for counterfactual learning to rank [C] //Proc of the 43rd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2020: 2089-2092

[179] Jagerman R, Oosterhuis H, de Rijke M. To model or to intervene: A comparison of counterfactual and online learning to rank from user interactions [C] //Proc of the 42nd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2019: 15-24

[180] Bottou L, Peters J, Quiñonero-Candela J, et al. Counterfactual reasoning and learning systems: The example of computational advertising [J]. The Journal of Machine Learning Research, 2013, 14(1): 3207-3260

[181] Lawrence C, Riezler S. Improving a neural semantic parser by counterfactual learning from human bandit feedback [C] //Proc of the 56th Annual Meeting of the ACL, Vol 1. Stroudsburg, PA: ACL, 2018: 1820-1830

[182] Bareinboim E, Forney A, Pearl J. Bandits with unobserved confounders: A causal approach [C] //Proc of the 29th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2015: 1342-1350

[183] Lee S, Bareinboim E. Structural causal bandits: Where to intervene? [C] //Proc of the 32nd Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2018: 2568-2578

[184] Lee S, Bareinboim E. Structural causal bandits with non-manipulable variables [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 4164-4172

[185] Haan P, Jayaraman D, Levine S. Causal confusion in imitation learning [C] //Proc of the 33rd Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2019: 11698-11709

[186] Kyono T, Zhang Yao, van der Schaar M. CASTLE: Regularization via auxiliary causal graph discovery [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 1501-1512

[187] Yang Mengyue, Liu Frurui, Chen Zhitang, et al. CausalVAE: Disentangled representation learning via neural structural causal models [C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 9593-9602

[188] Zinkevich M, Johanson M, Bowling M, et al. Regret minimization in games with incomplete information [C] //Proc of the 20th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2007: 1729-1736

[189] Brown N, Lerer A, Gross S, et al. Deep counterfactual regret minimization [C] //Proc of the 36th Int Conf on Machine Learning. Cambridge MA: JMLR, 2019: 793-802

[190] Farina G, Kroer C, Brown N, et al. Stable-predictive optimistic counterfactual regret minimization [C] //Proc of the 36th Int Conf on Machine Learning. Cambridge MA: JMLR, 2019: 1853-1862

[191] Brown N, Sandholm T. Solving imperfect-information games via discounted regret minimization [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 1829-1836

[192] Li Hui, Hu Kailiang, Zhang Shaohua, et al. Double neural counterfactual regret minimization [C/OL] //Proc of the 8th Int Conf on Learning Representations. 2020[2021-11-14]. https://openreview.net/pdf?id=ByedzkrKvH

[193] Oberst M, Sontag D. Counterfactual off-policy evaluation with Gumbel-max structural causal models [C] //Proc of the 36th Int Conf on

Machine Learning. Cambridge MA: JMLR, 2019: 4881-4890

[194] Buesing L, Weber T, Zwols Y, et al. Woulda, coulda, shoulda: Counterfactually-guided policy search [C/OL] //Proc of the 9th Int Conf on Learning Representations. 2019[2021-11-14]. https://openreview.net/pdf?id=BJG0voC9YQ

[195] Chen Long, Zhang Hanwang, Xiao Jun, et al. Counterfactual critic multi-agent training for scene graph generation [C] //Proc of the 2019 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4613-4623

[196] Zhu Qingfu, Zhang Weinan, Liu Ting, et al. Counterfactual off-policy training for neural dialogue generation [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 3438-3448

[197] Choi S, Park H, Yeo J, et al. Less is more: Attention supervision with counterfactuals for text classification [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 6695-6704

[198] Zhang Zhu, Zhao Zhou, Lin Zhejie, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding [C] //Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2020: 655-666

[199] Kocaoglu M, Snyder C, Dimakis A G, et al. CausalGAN: Learning causal implicit generative models with adversarial training [C] //Proc of the 6th Int Conf on Learning Representations, 2018[2021-11-03]. https://openreview.net/pdf?id=BJE-4xW0W

[200] Kim H, Shin S, Jang J H, et al. Counterfactual fairness with disentangled causal effect variational autoencoder [C] //Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 8128-8136

[201] Qin Lianhui, Bosselut A, Holtzman A, et al. Counterfactual story reasoning and generation [C] //Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: ACL, 2019: 5046-5056

[202] Hao Changying, Pang Liang, Lan Yanyan, et al. Sketch and customize: A counterfactual story generator [C] //Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 12955-12962.

[203] Madaan N, Padhi I, Panwar N, et al. Generate your counterfactuals: Towards controlled counterfactual generation for text [C] //Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 13516-13524

[204] Peysakhovich A, Kroer C, Lerer A. Robust multi-agent counterfactual prediction [C] //Proc of the 33rd Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2019: 3083-3093

[205] Baradel F, Neverova N, Mille J, et al. CoPhy: Counterfactual learning of physical dynamics [C/OL] //Proc of the 8th Int Conf on Learning Representations. 2020[2021-11-14]. https://openreview.net/pdf?id=SkeyppEFvS

Li Jianing, born in 1992. PhD candidate. His research interests include machine learning and text generation.

李家宁，1992 年生．博士研究生．主要研究方向为机器学习和文本生成．

Xiong Ruibin, born in 1996. Graduate student. His research interests include machine learning and natural language processing.

熊睿彬，1996 年生．硕士．主要研究方向为机器学习和自然语言处理．

Lan Yanyan, born in 1982. PhD, professor. Senior member of CCF. Her research interests include machine learning, information retrieval, and natural language processing.

兰艳艳，1982 年生．博士，教授．CCF 高级会员．主要研究方向为机器学习、信息检索和自然语言处理．

Pang Liang, born in 1990. PhD, associate researcher. Member of CCF. His research interests include natural language generation and information retrieval.

庞亮，1990 年生．博士，副研究员．CCF 会员．主要研究方向为自然语言生成和信息检索．

Guo Jiafeng, born in 1980. PhD, professor. Member of CCF. His research interests include data mining and information retrieval.

郭嘉丰，1980 年生．博士，研究员．CCF 会员．主要研究方向为数据挖掘和信息检索．

Cheng Xueqi, born in 1971. PhD, professor. Fellow of CCF. His research interests include network science and social computing, web search and mining, internet information security, distributed system and mass simulation platform.

程学旗，1971 年生．博士，研究员．CCF 会士．主要研究方向为网络科学与社会计算、互联网搜索与挖掘、互联网信息安全、分布式系统和大规模仿真平台．

作者贡献声明

李家宁和熊睿彬合作完成文献调研、内容整理和文章写作，对本文具有同等贡献．

兰艳艳对本文选题、组织结构和文章写作提供了关键性的指导意见．

庞亮对本文组织结构和部分内容提供了重要的指导意见．

郭嘉丰和程学旗对本文的选题提供了重要的指导意见．