计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(6): 1492−1505, 2026

基于对抗解耦与分布对齐的个体因果效应估计

曹付元[1,2] 李欣娜[1] 张佳璇[1]

[1](山西大学计算机与信息技术学院 太原 030006)

[2](计算智能与中文信息处理教育部重点实验室(山西大学) 太原 030006)

(cfy@sxu.edu.cn)

Individual Causal Effect Estimation via Adversarial Disentanglement and Distribution Alignment

Cao Fuyuan[1,2], Li Xinna[1], and Zhang Jiaxuan[1]

[1](*School of Computer and Information Technology*, *Shanxi University*, *Taiyuan* 030006)

[2](*Key Laboratory of Computational Intelligence and Chinese Information Processing* (*Shanxi University*), *Ministry of Education*, *Taiyuan* 030006)

Abstract The core challenge in estimating individual causal effects (ICE) from observational data lies in confounding bias and the unobservability of counterfactual outcomes. To eliminate confounding bias, existing methods typically balance all covariates as potential confounders. However, erroneously balancing instrumental variables and adjustment variables introduce estimation bias and increase variance. Although recent studies address this via covariate decomposition to separate confounders, their sample re-weighting approach for balance can lead to extreme weights while neglecting the sample imbalance between treatment and control groups, resulting in inaccurate ICE estimates. To address these limitations, we propose an individual causal effect estimation method based on adversarial learning, which optimizes representation learning and distribution balancing via an adversarial mechanism. Specifically: 1) Sample processing. The SMOTE algorithm preprocesses data to mitigate sample imbalance between treatment and control groups. 2) Representation learning. Two adversarial probability constraint modules are designed to achieve confounder separation via variable disentanglement. 3) Confounder balancing. A feature balancer and an adversarial discriminator dynamically align the spatial distributions of confounders across treatment and control groups, constructing an adversarial balancing network to avoid the extreme weight problem inherent in sample re-weighting. Experimental results demonstrate that our algorithm can precisely decompose confounders and significantly improve the accuracy of ICE estimation.

Key words individual causal effect; adversarial learning; disentangled representation; confounding balancing; sample imbalance

摘 要 从观测数据中估计个体因果效应面临的核心挑战在于混杂偏差与反事实结果的不可观测性。为消除混杂偏差,现有方法通常将全部协变量作为混杂因素进行平衡,但错误平衡工具变量和调整变量会引入估计偏差并扩大方差。尽管近期研究通过协变量分解来分离混杂因素,但其样本重加权的平衡方法

收稿日期: 2025-08-27; **修回日期**: 2026-02-04

基金项目: 国家自然科学基金项目(U24A20323, 62376145); 山西省科技创新人才团队专项(202204051002016); 太行山西省实验室技术攻关专项资助项目(THYF-JSZX-24010700)

This work was supported by the National Natural Science Foundation of China (U24A20323, 62376145), the Science and Technology Innovation Talent Team of Shanxi Province (202204051002016), and the Key Technology Research Project of Taihang Laboratory in Shanxi Province (THYF-JSZX-24010700).

会引起极端权重，同时忽视了处理组与对照组间的样本失衡问题，导致个体因果效应（individual causal effect, ICE）估计不准确。针对上述局限，提出了一种基于对抗学习的个体因果效应估计方法，该方法通过对抗机制优化表示学习与分布平衡。具体来说包括以下步骤：1）样本处理。采用 SMOTE 算法对数据进行预处理，以缓解处理组和对照组样本失衡问题。2）表示学习。设计 2 个对抗概率约束模块，通过变量解耦实现混杂因素分离。3）混杂平衡。通过特征平衡器与对抗判别器动态对齐混杂变量在处理组与对照组中的空间分布，构建对抗平衡网络，避免样本重加权中的极端权重问题。实验结果表明，所提方法可以精确地分解混杂因素并提高模型个体因果效应估计的准确性。

关键词 个体因果效应；对抗学习；分解表征；混杂平衡；样本量不平衡

中图法分类号 TP18

DOI: 10.7544/issn1000-1239.202550624 **CSTR:** 32373.14.issn1000-1239.202550624

个体因果效应（individual causal effect, ICE）估计作为因果推断领域的一项基本研究内容，旨在分析原因变量对结果变量的影响程度，其在个性化研究和精准决策等实际场景中发挥着重要的作用[1-5]。例如，在智能营销领域评估投放广告对不同消费者的购买能力的影响[6]；在医疗卫生领域评估某种新研发的药物对特定患者疾病治疗的影响[7]；在智慧教育领域评估不同教育策略对每个学生成绩的影响[8]。

为实现对 ICE 的可靠估计，研究者通常采用随机对照试验（randomized controlled trial, RCT）与观察性研究 2 种方式。但 RCT 往往成本高昂、耗时漫长，且可能面临伦理或可行性的限制。相比之下，丰富且易获取的观察性数据为因果效应估计提供了重要来源，这使得基于观察性研究的 ICE 估计受到了越来越多的关注。然而，由于观察性研究中的处理分配通常是非随机的，研究对象是否接受某种处理往往受到其自身特征的影响，从而产生混杂偏差[9-11]。例如，在医疗场景中，接受心脏药物治疗的患者通常比未接受治疗的患者年龄更大，这是因为医生的治疗决策往往依赖于患者的年龄。在这种情况下，年龄作为混杂变量同时影响处理分配和患者的康复程度，这种混杂偏差会进一步影响因果效应估计的准确性。

为了缓解混杂偏差，现有研究主要通过平衡处理组与对照组之间的分布差异来估计反事实结果。然而，大部分方法[12-13]通过将全部可观测协变量都视为混杂变量进行分布平衡，忽略了其中可能存在的非混杂变量。近期研究[14-15]表明，工具变量（仅影响处理分配）和调整变量（仅影响结果）的盲目平衡反而会引入估计偏差并扩大方差。尽管最近一些工作如 DeR-CFR[16]考虑了对协变量进行分解以分离混杂变量，但该方法在平衡混杂变量时采用的样本重加权策略可能会引发极端权重问题，导致 ICE 估计不稳定。同时，该方法在协变量平衡过程中忽视了样本数量失衡问题（如在医疗研究中，接受新疗法的患者数量远少于未接受治疗的患者），这种失衡会导致模型对多数组中的样本产生学习偏好而难以准确拟合少数群体的关键特征，从而影响模型对反事实结果预测的准确性，最终降低 ICE 的估计精度。

针对上述问题，本文提出了一种基于对抗解耦与分布对齐的 ICE 估计方法，该方法通过对抗机制以优化分解表示学习与混杂变量平衡。首先，为了缓解样本失衡对模型学习的负面影响，采用 SMOTE（synthetic minority over-sampling technique）算法对数据进行预处理，通过在少数类样本的特征空间中进行插值来生成合成样本，不仅有效平衡数据集，还能避免简单随机过采样导致的过拟合问题，同时增强模型对少数类特征的学习能力，为后续的因果效应估计提供更可靠的数据基础；然后，设计 2 个对抗性概率约束模块来约束组件之间的概率关系，使模型能够准确识别不同变量的潜在表示，实现混杂因素与工具变量和调整变量的精准分离，从而规避因盲目平衡工具变量与调整变量而引入的估计偏差；接着采用对抗平衡网络，通过特征平衡器与对抗判别器动态对齐处理组与对照组中混杂变量的分布，从而避免样本重加权方法可能引发的极端权重问题；最后，利用结果回归预测网络估计潜在结果，进而计算 ICE。实验结果表明，本文方法在 4 个合成数据集和 2 个半合成数据集上的性能均优于基准模型，验证了精确解耦混杂变量并实现分布平衡对于提升 ICE 估计效果的关键作用，显著提高了观察性研究中 ICE 估计的准确性。

本文贡献主要包括 3 个方面：

1）采用 SMOTE 算法生成少数类别的合成样本，

优化类别间样本分布,缓解样本失衡问题;

2)提出对抗性解耦表征学习方法,通过构建对抗性概率约束模块,有效分离工具变量、混杂变量和调整变量,从而规避因盲目平衡工具变量与调整变量而引入的估计偏差;

3)设计深度对抗平衡网络,通过特征平衡器与判别器博弈进行混杂变量平衡,实现分布对齐,避免了样本重加权方法可能引发的极端权重问题。

1 相关工作

现有面向观测数据的因果效应估计方法主要可分为传统统计学方法和深度学习方法两大类,下面将详细阐述这2类方法。

1.1 传统方法

在传统方法中,解决观察性研究中的混杂偏差问题通常采用倾向得分匹配(propensity score matching, PSM)、逆概率加权(inverse probability weighting, IPW)和双重稳健(doubly robust, DR)方法[17-20]。倾向得分匹配通过计算个体倾向得分并据此匹配处理组与对照组以实现协变量分布平衡,但其依赖线性模型假设,难以处理高维非线性数据;逆概率加权通过样本重加权调整分布差异,但对倾向得分估计误差敏感,极端权重可能放大方差。还有一些方法[21-24]通过熵平衡、残差平衡和稳定平衡来优化样本权重以提升平衡效果,但这些方法将所有可观测到的协变量都当成混杂变量进行平衡。最近,有研究[25-26]指出了在因果推理中进行混杂分离和选择的必要性,因为对某些非混杂因素的控制(如工具变量、调整变量)会产生额外的偏倚并放大方差。

1.2 深度学习方法

近年来,随着深度学习的发展,其强大的非线性建模能力与全样本信息整合优势,使得越来越多的研究开始引入深度学习模型来估计因果效应。Johansson等人[27]结合领域适应和表示学习的思路,提出平衡神经网络(balancing neural network, BNN)方法,构建了一种新的反事实推理框架。随后,Shalit等人[28]对BNN进行扩展,提出了TARNet(treatment agnostic representation network)与CFRNet(counterfactual regression network)两种方法,TARNet方法通过引入共享表示层来提升因果效应的估计精度,CFRNet方法则进一步融入积分概率度量(integral probability metric, IPM),通过最小化处理组与对照组在表示空间中的分布差异,如使用最大均值差异(maximum mean discrepancy, MMD)或Wasserstein距离来缓解混杂偏差。基于以上工作,Yao等人[29]提出了一种基于深度表示学习的局部相似性保留的ICE估计方法SITE(similarity preserved individual treatment effect),Hassanpour等人[30]提出了基于表示学习的上下文感知重要性采样重新加权的方法。此外,还有一些研究通过生成对抗网络(generative adversarial network, GAN)来估计ICE。GANITE(GAN for individual treatment effect)方法[31]利用GAN框架来生成反事实结果的代理,并通过这些代理训练ICE估计器,这种方法不仅能够利用已观察到的事实数据进行学习,还能在一定程度上考虑未观察到的反事实结果,从而提高因果效应估计的准确性。SCIGAN(estimating the effects of continuous interventions using GAN)方法[32]则进一步扩展了GAN框架,用于估计连续处理措施的因果效应,该方法通过构建分层鉴别器,能够同时估计多种不同连续处理措施的反事实结果,并在理论和实验上证明了其有效性。基于GAN的对抗思想,ABCEI(adversarial balancing-based representation learning for causal effect inference)方法[33]创新性地采用对抗训练来平衡潜在空间中的协变量分布,并引入互信息估计器以解决信息丢失问题。但这些方法均采用全局平衡策略而未区分协变量角色,可能引入工具变量的干扰。为此,部分研究尝试通过变量分解来区分不同角色的协变量。D^2VD(data-driven variable decomposition)方法[34]提出了一种数据驱动的变量分解算法,用于自动分离混杂变量和调整变量,但忽略了工具变量和混杂变量之间的区别,导致混杂变量分离不精确。DR-CFR(disentangled representations for counterfactual regression)方法[35]提出了反事实回归的非纠缠表示,但该算法不能保证清晰地分解混杂变量。DeR-CFR(decomposed representations for counterfactual regression)方法[16]则将协变量进一步分解为工具变量、混杂变量和调整变量,并利用正交约束增强变量分解的可解释性,但该方法引入个体级别的样本权重作为可学习参数来实现平衡,可能导致极端权重值,进而增大估计方差。同时,以上方法均未能同时考虑混杂偏差与样本不平衡问题,限制了在观测数据中的实际应用效果。

2 预备知识

在本节中,首先给出了观察性研究中处理效应估计的符号和假设,然后定义了工具变量(instrumental

variable)I、混杂变量(confounder variable)C和调整变量(adjustment variable)A，并分析了I，C，A的可识别性。该分析有助于剖析每个组件的概率属性，并为2.2节中提出的分解表示学习算法提供指导。

2.1 基本假设

在因果推断中，包含潜在结果模型(potential outcome model, POM)[36]与结构因果模型(structure causal model, SCM)[37]两大主流体系。POM通过引入反事实结果来形式化定义因果效应，为核心估计任务提供了清晰的数学语言；SCM则借助有向无环图刻画变量间的因果依赖关系，为系统理解多重因果关系提供了结构化视角。尽管两者在方法论表述上有所不同，但在因果效应的可识别性层面具有内在一致性。本文选择潜在结果框架作为ICE估计的主要建模基础，因其无需引入复杂的图论运算符即可清晰定义因果量，便于构建面向预测的深度学习模型；同时，为增强模型在复杂观测数据下的结构合理性与可解释性，本文引入因果图框架作为辅助工具，用于构建变量间的因果拓扑结构，如图1所示，特别是在识别混杂变量、工具变量和调整变量的角色方面提供理论支持。这种双框架融合策略为处理观察性数据中的复杂依赖关系提供了更为严谨的解决路径。

在本文中，仅探讨观测数据中二元处理情形下的ICE估计。令$\mathcal{X}\subseteq\mathbb{R}^d$表示$d$维协变量空间，$\mathcal{T}=\{0,1\}$表示二元处理空间，$\mathcal{Y}\subseteq\mathbb{R}$表示潜在结果空间。假设观测数据集包含$n$个独立同分布样本，其具体形式表示为$\mathcal{D}=\{\boldsymbol{x}_i,t_i,y_i\}_{i=1}^n$，数据服从分布$P(\boldsymbol{x},t,y^0,y^1)$。每个样本的协变量由$\boldsymbol{x}_i\in\mathcal{X}$表示，处理状态由二元变量$t_i\in\mathcal{T}$定义，其中$t_i=0$时表示第$i$个样本属于对照组，$t_i=1$时表示第$i$个样本属于处理组。每个样本对应2种潜在结果，其中$y_i^0$表示第$i$个样本未接受处理时的结果，$y_i^1$表示第$i$个样本接受处理后的结果。实际观测结果$y_i^{t_i}\in\mathcal{Y}$仅反映样本在其实际分配处理状态下的结果(即事实结果)，而未分配状态下的结果(反事实结果$y_i^{1-t_i}$)则无法直接观测。

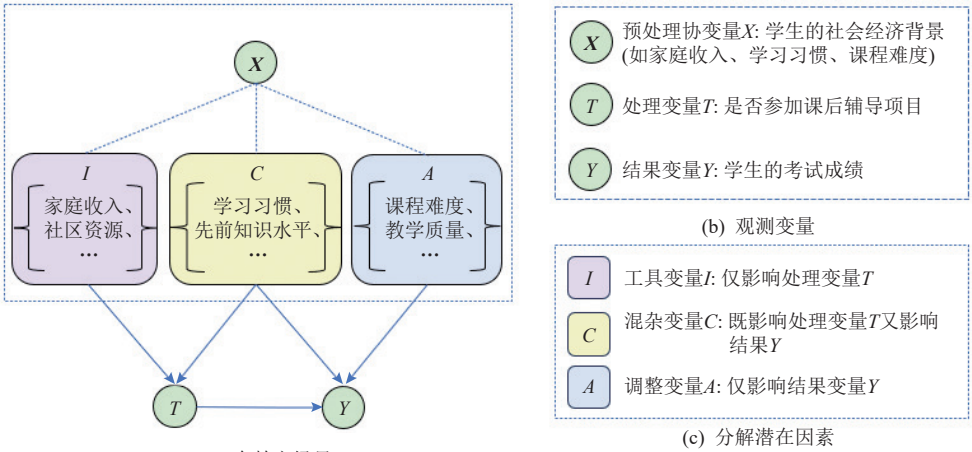


Fig. 1 Causality framework application in educational context

图1 因果关系框架在教育场景中的应用

样本i的ICE被定义为接受处理后与未接受处理时的潜在结果之间的差异：

$$ITE_i=y_i^1-y_i^0。\tag{1}$$

本文研究基于以下4个假设，这些假设共同构成了从观测数据中识别ICE的理论前提与充分必要条件。本文的研究工作立足于此理论框架，旨在提出一种更优的估计方法，以提升因果效应估计的可靠性与精度。各项假设具体内容如下：

假设 1. 稳定单元处理值假设(SUTVA)。每个个体的潜在结果不受其他个体所接受的处理值的影响，即样本间的潜在结果互不影响；同时对每个样本来说，每种处理只有1种特定的结果。

假设 2. 一致性(consistency)假设。对于一个处理t，其相应观测到的结果y是在处理t下的潜在结果y^t。

假设 3. 可忽略性(ignorability)假设。当t为处理变量，y^1和y^0分别为接受处理和未接受处理的潜在结果，且$\boldsymbol{x}$为给定协变量时，样本被分配的处理与样本的潜在结果独立，即满足$(y^1,y^0)\perp t|\boldsymbol{x}$。

假设 4. 正值性(positivity)假设。对于任意的协变量$\boldsymbol{x}\in\mathcal{X}$，样本被分配$t=1$的概率大于0，即满足$0<P(t=1|\boldsymbol{X}=\boldsymbol{x})<1$。

2.2 I，C，A的定义及其理论分析

如图 1 所示，基于因果图模型，将I, C, A形式化定义如下[16]。

定义 1. 将工具变量I、混杂变量C、调整变量A定义如下：

1)工具变量I。仅影响处理变量T，与结果变量Y无直接关联(除非通过T)，即工具变量I的路径满足$I \to T$且$I \nrightarrow Y$。

2)混杂变量C。同时影响处理变量T和结果变量Y，即混杂变量C的路径满足$C \to T$且$C \to Y$。

3)调整变量A。仅影响结果变量Y，与处理变量T独立，即调整变量A的路径满足$A \to Y$且$A \nrightarrow T$。

根据以上定义，进一步提出以下命题，以支持这 3 类潜在因素的分解及其表征学习。

命题 1. 若I, C, A满足定义 1，则：

1)混杂变量集合C能够阻塞从T到Y的所有后门路径；

2)在施行操作$do(t)$后的结果分布可以表示为$P(Y|X,do(t)) = P(Y|C,A,do(t))$。

命题 2. 变量集合$\{I,C,A\}$可以从联合分布$P(X, T,Y)$中通过以下条件识别：

1)调整变量A，

$X_j \in A \Leftrightarrow \{X_j \perp T\} \wedge \{\neg X_j \perp Y\}$。

2)工具变量I，

$X_j \in I \Leftrightarrow \{X_j \notin A\} \wedge \{\neg X_j \perp T\} \wedge \{X_j \perp Y|T,C\}$。

3)混杂变量C，$X_j \in C \Leftrightarrow \{X_j \notin A, X_j \notin I\}$。

其中X_j表示协变量集中第j个协变量，符号“$\perp$”表示独立，符号“$\neg\perp$”表示不独立。简言之，命题 2 表明：$\{A \perp T, \neg A \perp Y\}$，$\{I \perp Y|C \cup T, \neg I \perp T\}$且三者无交集。命题 2 展示了$\{I,C,A\}$基于概率关系的可识别性。

命题 2 通过定义 1 的内容可以很容易地理解，调整变量A与处理变量T独立，但与结果变量Y相关，形成了以Y为碰撞节点的对撞结构，即$T \to Y \leftarrow A$；工具变量I与处理变量T相关，且在给定T和混杂变量C的条件下与Y独立，其路径为$I \to T \to Y$，需要通过调整C阻断间接关联；混杂变量C同时关联处理变量T和结果变量Y，构成了后门路径$T \leftarrow C \to Y$，需要通过后门调整消除偏差。

基于上述理论，在实践中，本文通过 3 个神经网络分别学习分解的潜在表示$\{I(\boldsymbol{X}),C(\boldsymbol{X}),A(\boldsymbol{X})\}$，其中，命题 2 中的非独立性约束通过回归网络最大化相关表示的预测能力来实现，而独立性约束通过引入对抗的方式去实现。

3 所提方法

在本节中，将系统性地阐述基于对抗解耦与分布对齐的 ICE 估计方法。

3.1 整体框架

本文方法的整体模型结构如图 2 所示，主要由以下 4 个核心部分构成：

1)样本过采样模块。在模型训练前先采用 SMOTE 算法对训练集中的少数群体(如处理组)进行过采样，生成合成样本以平衡处理组和对照组的样本量，从而缓解样本失衡问题。

2)分解表征网络。将观测协变量$\boldsymbol{X}$解耦为工具变量$I(\boldsymbol{X})$、混杂变量$C(\boldsymbol{X})$和调整变量$A(\boldsymbol{X})$的潜在表示，并通过最大化$G_{I\to T}(\cdot)$，$G_{A\cup T\to Y}(\cdot)$两个回归网络的预测结果优化表征网络的解耦效果；同时构建对抗约束模块，通过概率约束确保$I(\boldsymbol{X})$与Y的条件独立，以及$A(\boldsymbol{X})$与T的独立性，进一步优化协变量的解耦表征。

3)对抗平衡网络。通过特征平衡器与判别器的对抗训练，动态对齐混杂变量$C(\boldsymbol{X})$在处理组与对照组的分布，以缓解混杂偏差的影响。

4)结果预测网络。通过 2 个潜在结果的回归预测网络去估计个体的潜在结果，从而进行 ICE 的估计。

3.2 核心架构设计

3.2.1 样本过采样

在观察性研究中，处理组与对照组的样本数量往往存在显著不均衡，这种分布差异容易导致模型训练产生偏差，进而影响因果效应估计的准确性。为解决这一问题，本文在模型训练前引入 SMOTE 算法。该算法通过在少数群体样本的特征空间中进行线性插值生成合成样本，既有效扩大了少数群体的样本规模，又避免了简单重复采样可能导致的过拟合问题。具体实现过程如下：

首先对训练集中的数据进行类别不平衡判定，计算处理组与对照组的样本比例，定义不平衡系数IR来量化样本失衡程度，具体公式为

$$IR = \frac{\max(N_0, N_1)}{\min(N_0, N_1)}, \tag{2}$$

其中$N_0 = \sum_{i=1}^{N}(1-t_i)$表示对照组样本量，$N_1 = \sum_{i=1}^{N} t_i$表示处理组样本量，当$IR$大于阈值时，判定为存在类别不平衡问题，在此情况下，为保障后续模型训练的公

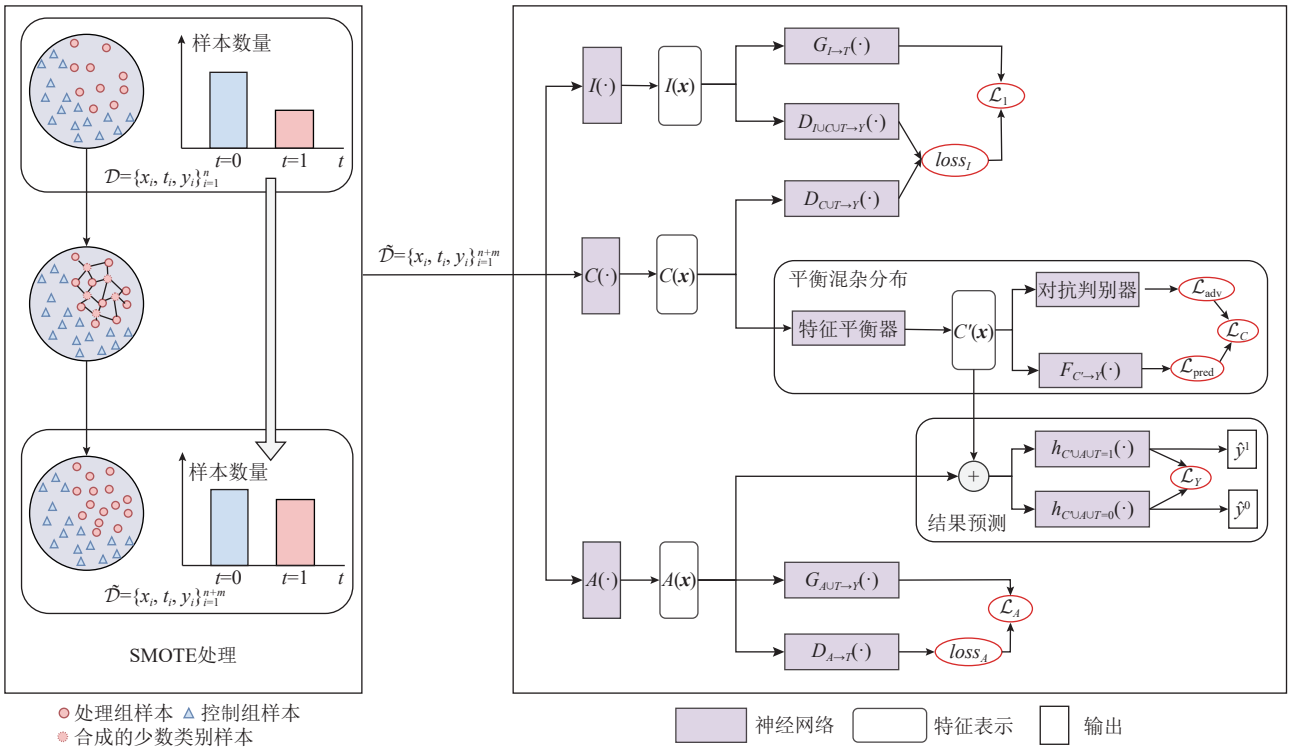


Fig. 2 Model structure diagram

图 2 模型结构图

正性，对少数组样本进行过采样。

对于少数类样本集合中的每个样本点，首先计算其在特征空间中的k近邻（默认设置$k=5$），然后随机选择一个近邻样本，通过线性插值生成新的合成样本，具体生成公式为

$$s_{\text{new}} = s + \xi(s_{\text{neighbor}} - s)\,, \tag{3}$$

其中s表示少数类样本集合中的原始样本点，s_{neighbor}是从其k近邻中随机选择的一个近邻样本点，ξ是在[0,1]范围内均匀分布的随机数。

在采样过程中，通过控制合成样本的生成数量，将处理组与对照组的样本比例调整至1/2~1，既有效缓解了样本分布的不均衡问题，又避免了因过度合成而引入噪声。此外，为保证后续评估的客观性，过采样操作只应用于训练集，测试集保持原始数据分布，以确保模型在评估阶段能够真实反映其在自然数据分布下的泛化性能。

3.2.2 分解协变量

在协变量分解过程中，通过3个神经网络来学习分解的潜在表示$\{I(\boldsymbol{X}), C(\boldsymbol{X}), A(\boldsymbol{X})\}$，其中工具变量与调整变量的非独立性约束通过2个回归网络$G_{I\to T}(\cdot)$，$G_{A\cup T\to Y}(\cdot)$最大化$I(\boldsymbol{X})$，$A(\boldsymbol{X})$表示的预测能力来实现，独立性约束通过对抗的思想来实现，具体做法如下。

1）调整变量A。从命题2中可知，调整变量A仅影响结果变量Y且调整变量A与处理变量T独立。为了实现$A(\boldsymbol{X})\perp T$，即满足$P(T|A(\boldsymbol{X}))=P(T)$，本文引入了一个辅助预测网络$D_{A\to T}(\cdot)$，根据$A(\boldsymbol{X})$预测$T$。该训练分为2个阶段：第1阶段，首先固定表示$A(\boldsymbol{X})$，通过最小化$loss_A$的损失来更新$D_{A\to T}(\cdot)$；第2阶段，固定$D_{A\to T}(\cdot)$网络，通过最大化$loss_A$的损失来更新$A(\cdot)$的表示，使其满足$A(\boldsymbol{X})\perp T$。对于$\neg A(\boldsymbol{X})\perp Y$，为了防止调整变量的信息被嵌入到其他变量中，引入了预测网络$G_{A\cup T\to Y}(\cdot)$，使其在更新$A(\cdot)$的过程中最小化$loss_{AY}$的损失。具体对应的损失函数为

$$loss_A = \sum_i l(D_{A\to T}(A(\boldsymbol{x}_i)), t_i)\,, \tag{4}$$

$$loss_{AY} = \sum_i l(G_{AUT\to Y}(A(\boldsymbol{x}_i), t_i), y_i)\,, \tag{5}$$

$$\mathcal{L}_A = loss_{AY} - loss_A\,, \tag{6}$$

其中$l(\cdot,\cdot)$表示连续结果的L_2损失函数或二元结果的对数损失函数。

2）工具变量I。从命题2中可知，工具变量I仅影响处理变量T并且工具变量I与结果变量Y无直接关联（除非通过T）。为了实现$I(\boldsymbol{X})\perp Y|\{T, C(\boldsymbol{X})\}$，即满足$P(Y|C(\boldsymbol{X})\cup I(\boldsymbol{X})\cup T)=P(Y|C(\boldsymbol{X})\cup T)$，本文引入了2个辅助预测网络$D_{C\cup T\to Y}(\cdot)$和$D_{I\cup C\cup T\to Y}(\cdot)$，该训练分为2

个阶段：第 1 阶段，首先固定$I(\cdot)$和$C(\cdot)$的表示，分别通过最小化$loss_{I1}$和$loss_{I2}$的损失来更新$D_{C\cup T\to Y}(\cdot)$和$D_{I\cup C\cup T\to Y}(\cdot)$；第 2 阶段，固定$D_{C\cup T\to Y}(\cdot)$和$D_{I\cup C\cup T\to Y}(\cdot)$网络，通过最小化 2 个预测量之间的差异$l_d(\cdot,\cdot)$，即通过最小化$loss_I$的损失来更新$I(\cdot)$和$C(\cdot)$的表示，使其满足$I(\boldsymbol{X})\perp Y|\{T,C(\boldsymbol{X})\}$。对于$\neg I(\boldsymbol{X})\perp T$，为了防止工具变量的信息被嵌入到其他变量中，引入了预测网络$G_{I\to T}(\cdot)$，使其在更新$I(\cdot)$的过程中最小化$loss_{IT}$的损失。具体对应的损失函数为

$$loss_{I1}=\sum_i l(D_{C\cup T\to Y}(C(\boldsymbol{x}_i),t_i),y_i)\,, \tag{7}$$

$$loss_{I2}=\sum_i l(D_{I\cup C\cup T\to Y}(I(\boldsymbol{x}_i),C(\boldsymbol{x}_i),t_i),y_i)\,, \tag{8}$$

$$loss_I=\sum_i l(D_{I\cup C\cup T\to Y}(I(\boldsymbol{x}_i),C(\boldsymbol{x}_i),t_i),D_{C\cup T\to Y}(C(\boldsymbol{x}_i),t_i))\,, \tag{9}$$

$$loss_{IT}=\sum_i l(G_{I\to T}(I(\boldsymbol{x}_i)),t_i)\,, \tag{10}$$

$$\mathcal{L}_I=loss_{IT}-loss_I\,, \tag{11}$$

其中$l(\cdot,\cdot)$表示连续结果的L_2损失函数或二元结果的对数损失函数。

3)混杂变量C。$C\notin A$且$C\notin I$，I，C，A这 3 个变量没有交集。为实现输入变量$\boldsymbol{X}$在表示模块$\{I(\boldsymbol{X}),C(\boldsymbol{X}),A(\boldsymbol{X})\}$间的显式分解，本文受 Wu 等人[16]工作的启发，引入深度正交正则化，通过施加正交性约束，强制特征表示模块$I(\boldsymbol{X}),C(\boldsymbol{X}),A(\boldsymbol{X})$不相交。具体而言，以表示网络$I(\boldsymbol{X})$为例(假设网络包含$l$层)，其各层权重矩阵为$\boldsymbol{W}_{I1},\boldsymbol{W}_{I2},\cdots,\boldsymbol{W}_{Il}$，首先通过连乘各层权重矩阵构造综合权重矩阵$\boldsymbol{W}_I=\boldsymbol{W}_{I1}\times\boldsymbol{W}_{I2}\times\cdots\times\boldsymbol{W}_{Il}\in\mathbb{R}^{d\times m}$(其中$d$为输入变量$\boldsymbol{X}$的维度，$m$为潜在表示$I(\boldsymbol{X})$的维度)，并逐行取平均得到变量贡献向量$\bar{\boldsymbol{W}}_I\in\mathbb{R}^d$，类似地，计算$\bar{\boldsymbol{W}}_C$，$\bar{\boldsymbol{W}}_A$。为强制不同表示网络关注输入变量的不同子集，设计正交损失函数以最小化不同模块对同一变量的依赖；同时进一步对贡献向量施加归一化约束，强制其总和接近 1，损失函数为

$$\mathcal{L}_O=\left[\bar{\boldsymbol{W}}_I^{\mathrm{T}}\cdot\bar{\boldsymbol{W}}_C^{\mathrm{T}}+\bar{\boldsymbol{W}}_C^{\mathrm{T}}\cdot\bar{\boldsymbol{W}}_A^{\mathrm{T}}+\bar{\boldsymbol{W}}_A^{\mathrm{T}}\cdot\bar{\boldsymbol{W}}_I^{\mathrm{T}}\right]+\sum_{\bar{\boldsymbol{W}}\in\{\bar{\boldsymbol{W}}_I,\bar{\boldsymbol{W}}_C,\bar{\boldsymbol{W}}_A\}}\left[\left(\sum_{k=1}^{d}\bar{W}[k]-1\right)^2\right], \tag{12}$$

其中$\bar{W}_I[k]$表示第k个输入协变量对表示$I(\boldsymbol{X})$的平均贡献，$\bar{W}_C[k]$，$\bar{W}_A[k]$以相同的方式定义。

3.2.3 平衡混杂变量

对于混杂变量C，因其同时影响处理变量T和结果变量Y，构成了后门路径$T\leftarrow C\rightarrow Y$，导致混杂变量可能会引发混杂偏差。为了消除这种偏差，需要通过平衡混杂变量来阻断后门路径，进而打破混杂变量和处理变量之间的依赖性。在本研究中，采用对抗平衡的方法来消除混杂变量C的影响，其核心目标是通过特征平衡器使处理组($T=1$)和对照组($T=0$)在平衡后的特征空间中分布不可区分，从而阻断混杂变量C对处理变量T的干扰。具体做法如下：

首先构建一个包含特征平衡器 G 和对抗判别器 D 的对抗学习框架。特征平衡器 G 的输入为原始混杂变量$C(\boldsymbol{X})$，输出为平衡后的特征$C'=G(C(\boldsymbol{X}))$，特征平衡器 G 通过非线性映射调整C的分布，使得处理组和对照组在C'上的分布趋于一致。对抗判别器 D 的输入为C'，输出为样本属于处理组的概率$D(C')$，其任务是从C'中尽可能准确地预测T，但在训练过程中通过梯度反转层反向传播时反转梯度方向，迫使 G 的优化目标与 D 相反，即最大化 D 的预测误差，从而消除C'，T的关联性，其对抗损失函数为

$$\mathcal{L}_{\text{adv}}=-E[T\ln D(C')+(1-T)\ln(1-D(C'))]\,。 \tag{13}$$

在平衡混杂变量C的表征时，仅切断混杂变量C与处理变量T的关联，同时保留混杂变量C与结果变量Y的关联性。为维持平衡后特征C'对结果变量Y的预测能力，引入了预测网络 F，其输入为C'，输出对Y的预测值$\hat{Y}=F(C')$，通过均方误差来约束C'与Y的关联性，具体损失函数为

$$\mathcal{L}_{\text{pred}}=E[(F(C')-Y)^2]\,。 \tag{14}$$

因此，对抗平衡框架的总损失函数被定义为

$$\mathcal{L}_C=\boldsymbol{\gamma}_1\mathcal{L}_{\text{adv}}+\boldsymbol{\gamma}_2\mathcal{L}_{\text{pred}}\,, \tag{15}$$

其中$\boldsymbol{\gamma}_1$，$\boldsymbol{\gamma}_2$为模型的超参数，$\mathcal{L}_{\text{adv}}$为对抗损失，$\mathcal{L}_{\text{pred}}$为预测损失。

3.2.4 结果预测网络

在分解网络和对抗平衡网络的基础上构建结果回归预测网络，用于估计处理变量T对结果变量Y的因果效应。

对于处理组$T=1$和对照组$T=0$，本文训练了 2 个回归网络$h_{C'\cup A\cup T=1}(\cdot)$，$h_{C'\cup A\cup T=0}(\cdot)$，分别基于$T=1$和$T=0$的样本观察结果。如图 2 模型，本研究仅基于$A(\boldsymbol{X})$和平衡后$C'(\boldsymbol{X})$的表示来训练这 2 个回归模型。具体损失函数为

$$\mathcal{L}_Y=\sum_i l(h_{C'\cup A\cup T\to Y}(C'(\boldsymbol{x}_i),A(\boldsymbol{x}_i),t_i),y_i)\,, \tag{16}$$

其中$l(\cdot,\cdot)$表示连续结果的L_2损失函数或二元结果的对数损失函数。

3.3 模型总损失

在 3.2 节中详细介绍了本文的研究方法，模型的

总损失函数为

$$\mathcal{L}=\mathcal{L}_Y+\alpha\mathcal{L}_A+\beta\mathcal{L}_I+\mu\mathcal{L}_O+\mathcal{L}_C+\lambda\|\boldsymbol{\Theta}\|_2, \quad (17)$$

其中$\lambda\|\boldsymbol{\Theta}\|_2$是对模型参数$\boldsymbol{\Theta}$的$L_2$正则化。

4 实　　验

为评估所提方法在分解表示学习及ICE估计2个方面的性能，本节在4个合成数据集与2个半合成数据集上实施了实验。

4.1 实验设置

1)数据集。本研究中所使用的数据集为Wu等人[16]提供的4个合成数据集Binary_8_8_8_3 000/10 000, Binary_16_16_16_3 000/10 000和2个因果效应估计研究中常用的公开数据集IHDP和Twins。

2)合成数据集。本文直接采用Wu等人[16]提供的二进制设置$\{T=0,1;Y=0,1\}$的合成数据集，数据集名为Binary_m_I_m_C_m_A_n(其中m_I, m_C, m_A, n分别为工具变量、混杂变量和调整变量的维度以及样本量的个数，$n\in\{3\,000,10\,000\}$, $m_I=m_C=m_A\in\{8,16\}$。以Binary_16_16_16_3 000数据集为例，数据生成过程如下：

$\boldsymbol{X}_I=(X_1,X_2,\cdots,X_{16})$, $\boldsymbol{X}_C=(X_{17},X_{18},\cdots,X_{32})$, $\boldsymbol{X}_A=(X_{33},X_{34},\cdots,X_{48})$，除了上述3类变量外，协变量$\boldsymbol{X}$还包括噪声变量$m_D=2$的额外维度。协变量独立地从标准正态分布$N(0,1)$中生成，样本大小为$n=3\,000$。用$\boldsymbol{X}_{IC}=(\boldsymbol{X}_I,\boldsymbol{X}_C)$来生成处理变量$T$, $T\sim B(1,P(\boldsymbol{x}_{IC}))$, $P(\boldsymbol{x}_{IC})=[1+\exp(-(\theta_t\boldsymbol{x}_{IC}+\epsilon))]^{-1}$, $\theta_t\sim U((8,16)^{m_I+m_C})$, $\epsilon\sim N(0,0.1^2)$；用$\boldsymbol{X}_{CA}=(\boldsymbol{X}_C,\boldsymbol{X}_A)$来生成结果变量$Y$，首先生成 $\mu_1=\theta_{y^1}(\boldsymbol{x}_{CA}^{\mathrm{T}}\boldsymbol{x}_{CA})$ 和$\mu_0=\theta_{y_0}\boldsymbol{x}_{CA}$，其中$\theta_{y^0},\theta_{y^1}\sim U((8,16)^{m_C+m_A})$，然后生成二元结果$y^0=\mathrm{sgn}(\max(0,\mu_1-\tilde{\mu}_1))$和$y^1=\mathrm{sgn}(\max(0,\mu_0-\tilde{\mu}_0))$，其中$\tilde{\mu}_1$和$\tilde{\mu}_0$表示中位数，最后，生成$y=y^1t+y^0(1-t)$。

3)IHDP数据集。Hill[38]基于婴儿健康与发展计划研究(IHDP)的原始数据构建了一个因果推断数据集，旨在评估“专家家访”对“早产儿未来认知测试评分”的影响。该数据集通过删除处理组中的非随机子集，构造了协变量分布不平衡的处理组和对照组。数据集共包含747个样本，涵盖25个维度的协变量(包括母婴特征、家庭环境等)。本研究将遵循Hill[38]的标准设置进行因果效应估计。

4)Twins数据集。原始Twins数据集来源于1989—1991年在美国出生的所有双胞胎[39]。当某个体为双胞胎中较重者时，处理变量标记为$t_i=1$，较轻者标记为$t_i=0$。此外，数据集包含30个与父母、妊娠及出生相关的变量，研究结果为儿童1年后的死亡率。数据集聚焦于体重低于2 kg且无特征缺失的同性别双胞胎，最终数据集包含11 400条记录。为了模拟观察性研究，在数据集中使用特征信息选择性地观察双胞胎中的一个(产生混杂偏差)，如下所示，$t_i|\boldsymbol{x}_i\sim Bern(sigmoid(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{x}_i+n))$，其中$n\sim N(0,0.1)$, $\boldsymbol{\Phi}^{\mathrm{T}}\sim U((-0.1,0.1)^{1\times30})$。

5)基线方法。在本文中，将所提方法与以下基线模型进行了比较。① CFR-MMD[28]：基于反事实回归的表示学习方法，使用MMD作为分布距离度量，强制处理组和对照组的表示分布对齐；②CFR-WASS[28]：改用Wasserstein距离进行分布对齐(相比MMD对低维分布更敏感)，通过分布对齐减少混杂偏差；③ TARNet[28]：通过共享表示层学习处理组与对照组的共同特征以提升潜在结果的预测精度；④ SITE[29]：通过添加位置相关深度度量和中间点距离最小化约束，在保持局部相似性信息的基础上平衡协变量分布；⑤ CFR-ISW[30]：在反事实回归框架中引入重要性抽样权重(importance sampling weights, ISW)，通过调整样本权重重新加权损失函数，使处理组和对照组的数据分布更接近；⑥ DR-CFR[35]：通过解耦表示学习分离与处理变量和结果变量相关的特征以减少混杂偏差；⑦ DeR-CFR[16]：在DR-CFR基础上，进一步细化表示分解，并引入动态样本权重减少混杂偏差；⑧ CrossNet[40]：交叉利用处理组与对照组样本训练潜在结果预测模型，并通过分布一致性约束学习满足无混淆性的充分表示，提升因果效应的估计精度。

6)评估指标。在实证评估中，本文针对每个数据集分别进行样本内(within-sample)与样本外(out of-sample)评估。样本内评估基于训练数据集开展，反映模型对已知数据的拟合能力；样本外评估则针对测试数据集，检验模型对新样本的泛化能力。

在合成数据集、IHDP和Twins数据集上，本研究采用异质效应估计精度(precision in estimation of heterogeneous effect, PEHE)[38]作为个体水平的性能指标：

$$\varepsilon_{\mathrm{PEHE}}=\frac{1}{N}\sum_{i=1}^{N}((\hat{y}_i^1-\hat{y}_i^0)-(y_i^1-y_i^0))^2, \quad (18)$$

其中$\hat{y}_i^1$和$\hat{y}_i^0$分别为模型对个体i在处理($T=1$)和未处理($T=0$)时潜在结果的预测值，y_i^1和y_i^0为个体i的真实潜在结果。$\varepsilon_{\mathrm{PEHE}}$表示模型预测的ICE与真实ICE

之间的误差，$\varepsilon_{\text{PEHE}}$值越低，模型预测效果越好。

7)超参数设置。在本方法中，选择 ELU 作为非线性激活函数，采用 Adam 优化器以最小化模型的目标函数。在训练过程中，为每个模块分配一个自适应权重，并将所有样本视为一个整批，最大迭代次数为 100。表 1 给出了本文所使用的超参数设置。

Table 1 Model Hyper-Parameters Settings

表 1 模型超参数设置

超参数	范围
调整变量分解参数α	{0.001,0.01,0.1,0.5,1,5}
工具变量分解参数β	{0.001,0.01,0.1,1,5,10,100}
混杂变量平衡参数γ_1,γ_2	{0.001,0.1,0.5,1,3,5}
正交正则化参数μ	{0.001,0.01,0.1,1,5,10,100}
整体正则化参数λ	{0.001,0.01,0.1,1,3,5}

4.2 在合成数据集上的实验

为系统评估本文方法在分解表示学习中的性能，本节在 4 个具有明确变量类型标注的合成数据集上开展了实验验证。所用数据集包括 2 种配置：低维配置数据集(Binary_8_8_8_3 000/10 000)包含 8 个工具变量(维度 1~8)、8 个混杂变量(维度 9~16)以及 8 个调整变量(维度 17~24)；高维配置数据集(Binary_16_16_16_3 000/10 000)包含 16 个工具变量(维度 1~16)、16 个混杂变量(维度 16~32)以及 16 个调整变量(维度 33~48)。每种配置均包含 3 000 和 10 000 这 2 种样本量，以全面评估方法在不同数据复杂度下的性能表现。

实验通过式(12)中的贡献度矩阵$\{\bar{\boldsymbol{W}}_I,\bar{\boldsymbol{W}}_C,\bar{\boldsymbol{W}}_A\}$量化协变量中各变量对工具变量$I$、混杂变量$C$和调整变量$A$的分解表示强度。图 3 可视化结果清晰展示了各协变量被准确分类至对应变量类型的情况，直观验证了本文方法在从低维到高维的不同数据集上均能有效分离 3 类基本因素并保持其语义真实性，且分解精度随样本量增加呈上升趋势。

在此基础上，本文进一步计算精确率、召回率与 $F1$ 分数等指标，对分解准确性进行定量评估并在图 4 中予以可视化展示。结果表明，本文方法在 4 个数据集上均取得优异且稳定的总体性能：工具变量的识别呈现较高的召回率，混杂变量的判定保持了较高的精确率，而调整变量则在各项指标上表现最为均衡。图 4 的定量结果与图 3 的可视化分析相互印证，共同证实了本文方法在变量分离方面的有效性与鲁棒性。

同时，本节实验将本文方法与 4.1 节中的基线方法在 4 个合成数据集上进行整体性能比较。表 2 结果显示，本文方法在所有合成数据集上的$\varepsilon_{\text{PEHE}}$指标均优于对比基线，展现出最优且稳定的估计性能。具体而言，未显式进行分布平衡的 TARNet 方法性能相对有限，凸显了处理混杂偏差的必要性；基于分布对齐的 CFR-MMD, CFR-WASS 等方法虽在一定程度上提升了估计精度，但由于未显式解耦混杂变量，其性能提升受到制约；CrossNet 方法通过交叉利用 2 组样本训练反事实预测器并施加分布一致性约束，虽较传统 CFR 方法有所改进，但其仍采用单一共享表示，未能彻底剔除与工具变量相关的噪声，因而在性能上仍存在局限。相比之下，本文方法通过显式解耦协变量为工具变量、混杂变量和调整变量，并结合对抗平衡机制，能够更有效地控制混杂偏差，同时避免无效变量的干扰，从而在反事实预测和个体因果效应估计任务中取得显著优势。这一结果充分验证

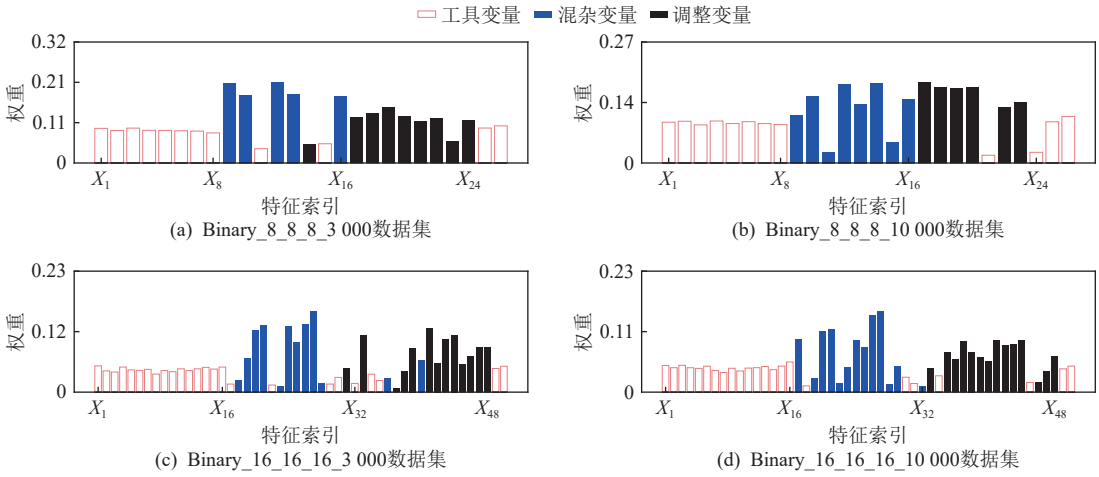
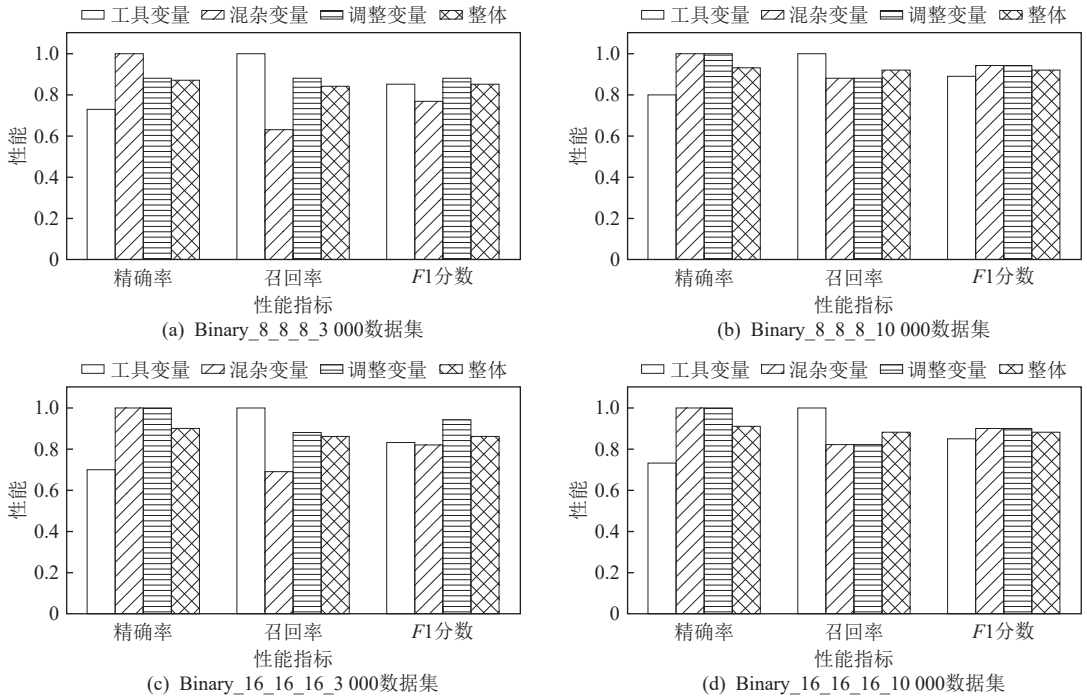


Fig. 3 Visualization results of covariate contributions to disentangled representations of$\{I,C,A\}$ on synthetic datasets

图 3 合成数据集上协变量对$\{I,C,A\}$分解表示贡献的可视化结果



(a) Binary_8_8_8_3 000数据集　(b) Binary_8_8_8_10 000数据集

(c) Binary_16_16_16_3 000数据集　(d) Binary_16_16_16_10 000数据集

Fig. 4　Quantitative evaluation results of $\{I,C,A\}$ variable identification performance on synthetic datasets

图 4　合成数据集上对$\{I,C,A\}$变量识别性能的定量评估结果

Table 2　Results on Synthetic Datasets

表 2　在合成数据集上的结果

方法	样本内$\varepsilon_{\text{PEHE}}$			
	Binary_8_8_8_3 000	Binary_8_8_8_10 000	Binary_16_16_16_3 000	Binary_16_16_16_10 000
CFR-MMD	0.384±0.004	0.276±0.004	0.491±0.005	0.399±0.005
CFR-WASS	0.378±0.004	0.277±0.004	0.513±0.007	0.408±0.005
TARNet	0.620±0.008	0.577±0.007	0.680±0.008	0.617±0.023
SITE	0.550±0.007	0.497±0.006	0.585±0.005	0.608±0.006
CFR-ISW	0.383±0.005	0.279±0.004	0.538±0.003	0.441±0.005
DR-CFR	0.377±0.002	0.288±0.005	0.544±0.004	0.427±0.015
DeR-CFR	0.325±0.002	0.234±0.003	0.404±0.003	0.307±0.002
CrossNet	0.601±0.012	0.559±0.009	0.667±0.014	0.607±0.008
本文	**0.315±0.002**	**0.229±0.002**	**0.397±0.003**	**0.301±0.003**
方法	样本外$\varepsilon_{\text{PEHE}}$			
	Binary_8_8_8_3 000	Binary_8_8_8_10 000	Binary_16_16_16_3 000	Binary_16_16_16_10 000
CFR-MMD	0.465±0.006	0.327±0.006	0.574±0.007	0.463±0.006
CFR-WASS	0.469±0.011	0.320±0.006	0.553±0.006	0.469±0.005
TARNet	0.632±0.007	0.597±0.007	0.684±0.008	0.672±0.009
SITE	0.561±0.005	0.506±0.006	0.588±0.007	0.612±0.009
CFR-ISW	0.461±0.006	0.334±0.006	0.553±0.006	0.501±0.005
DR-CFR	0.469±0.011	0.333±0.006	0.551±0.008	0.486±0.011
DeR-CFR	0.409±0.009	0.286±0.007	0.485±0.006	0.376±0.006
CrossNet	0.613±0.010	0.609±0.008	0.679±0.011	0.653±0.006
本文	**0.401±0.003**	**0.279±0.006**	**0.475±0.005**	**0.368±0.004**

注: “±”前数值表示平均值, “±”后数值表示误差。黑体数值表示最优值。

了变量分解与分布对齐相结合的策略在提升因果推断精度方面的有效性和必要性。

4.3 在真实数据集上的实验

为验证本文方法在真实场景下的泛化能力，本节在广泛使用的因果推断基准数据集 IHDP 和 Twins 上进行了实验，表 3 和表 4 分别展示了本文方法与 4.1 节中的基线方法在 2 个数据集上的整体性能对比。结果表明，本文方法在 2 个半合成数据集上均取得了具有竞争力的性能表现。

Table 3 $\varepsilon_{\text{PEHE}}$ on IHDP Dataset

表 3 在 IHDP 数据集上的$\varepsilon_{\text{PEHE}}$

方法	样本内	样本外
CFR-MMD	0.792±0.015	0.869±0.028
CFR-WASS	0.769±0.015	0.827±0.020
TARNet	0.800±0.016	0.950±0.031
SITE	0.720±0.018	0.791±0.022
CFR-ISW	0.700±0.028	0.735±0.108
DR-CFR	0.657±0.028	0.789±0.091
DeR-CFR	0.519±0.032	0.529±0.068
CrossNet	1.054±0.001	1.093±0.120
本文	**0.503±0.019**	**0.525±0.045**

注：黑体数值表示最优值。“±”前数值表示平均值，“±”后数值表示误差。

Table 4 $\varepsilon_{\text{PEHE}}$ on Twins Dataset

表 4 在 Twins 数据集上的$\varepsilon_{\text{PEHE}}$

方法	样本内	样本外
CFR-MMD	0.279±0.009	0.284±0.005
CFR-WASS	0.278±0.006	0.281±0.005
TARNet	0.280±0.006	0.286±0.004
SITE	0.279±0.004	0.283±0.006
CFR-ISW	0.279±0.008	0.283±0.006
DR-CFR	0.277±0.001	0.282±0.005
DeR-CFR	0.277±0.001	0.280±0.005
CrossNet	0.319±0.001	0.321±0.004
本文	**0.209±0.005**	**0.223±0.004**

注：黑体数值表示最优值。“±”前数值表示平均值，“±”后数值表示误差。

相较于仅学习共享表示的 TARNet 方法，以及采用不同正则化约束进行协变量平衡的 CFR-MMD, CFR-WASS, SITE, CFR-ISW 方法和引入交叉训练机制的 CrossNet 方法，本文方法在 IHDP 数据集上不仅实现了更为显著的$\varepsilon_{\text{PEHE}}$提升，同时在性能表现上也优于采用分解表示策略的 DR-CFR 和 DeR-CFR 方法。该性能提升主要源于 2 方面：其一，在表示学习阶段显式解耦协变量，有效降低了非混杂变量对 ICE 估计的干扰；其二，在数据处理阶段采用 SMOTE 算法合成处理组样本，显著缓解了 IHDP 数据集中处理组样本量远小于对照组（处理组 139 例、对照组 608 例）而导致的样本不平衡问题。此外，在包含高维混杂因子的 Twins 数据集上，本文方法同样保持了竞争力，展现出良好的鲁棒性。这些结果共同验证了：通过显式解耦协变量的表示学习策略，并结合 SMOTE 样本平衡技术，能够有效缓解混杂偏差与样本数量不一致问题，进而提升反事实结果及 ICE 的估计准确性。

4.4 消融实验

为了验证本文方法中每个组件的有效性，本节进行消融实验研究。首先，为评估本文方法中 2 个对抗约束模块的必要性及其对协变量解耦的有效性，依次移除对应的对抗约束模块，并在专门设计的合成数据集 Binary_16_16_16_3 000（用于验证协变量解耦）上进行了对比实验。图 5 和图 6 分别展示了移除不同模块后，协变量中各变量对解耦表示 I, C, A 的贡献可视化结果。

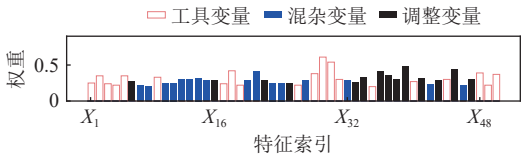


Fig. 5 Visualization results for the removal of $loss_I$

图 5 移除$loss_I$的可视化结果

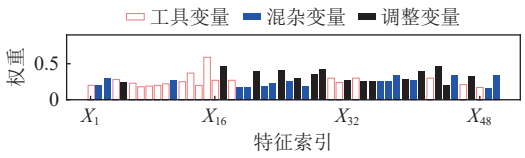


Fig. 6 Visualization results for the removal of $loss_A$

图 6 移除$loss_A$的可视化结果

为进一步量化分析 2 个对抗约束模块对变量分解性能的影响，图 7 和图 8 分别展示了移除“工具变量 I”与“调整变量 A”的对抗约束模块后，在变量识别任务上的定量结果。如图 7 所示，移除工具变量 I 的对抗约束模块后，导致 I 与 C 的识别性能显著下降，而 A 的识别相对不受影响；反之，如图 8 所示，移除调整变量 A 的对抗约束模块后，导致 A 与 C 的分解严重混淆，而 I 的识别保持良好。上述结果表明，2 个对抗约束模块通过明确的功能分工实现了有效互补，共同为提升工具变量 I、调整变量 A 与混杂变量 C 的整体分离性能做出了关键贡献。

同时，本文也在 IHDP 数据集上进行了消融实验，以评估本文方法中各模块的贡献，结果如表 5 所示。

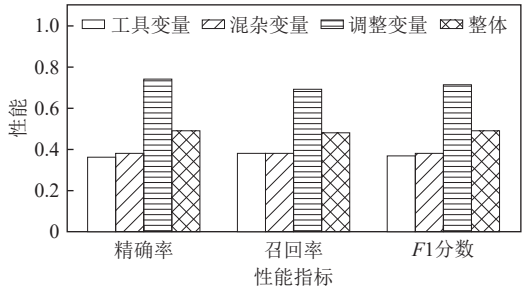


Fig. 7 Quantitative evaluation results for the removal of $loss_I$

图 7 移除$loss_I$的定量评估结果

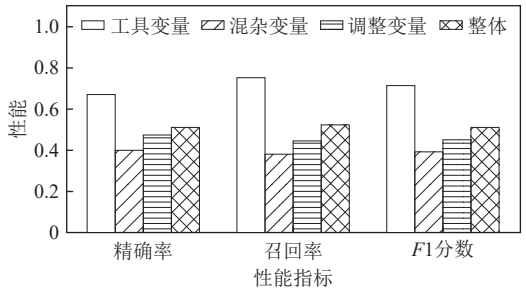


Fig. 8 Quantitative evaluation results for the removal of $loss_A$

图 8 移除$loss_A$的定量评估结果

实验结果表明，移除方法中任意一个模块均会影响混杂变量的分解表示学习和 ICE 的估计性能。

4.5 超参数敏感性分析

本节以合成数据集 Binary_16_16_16_10 000 为例，对模型中涉及的超参数$\{\alpha,\beta,\gamma_1,\gamma_2,\mu,\lambda\}$开展敏感性分析，通过依次调整单一参数，同步评估其对$\varepsilon_{\text{PEHE}}$与结果预测准确率的影响，以探究各超参数取值对模型性能的影响规律，如图 9 所示。分析结果表明，超参数α与λ对模型性能具有显著影响。α作为调整变量分解参数，其取值直接影响调整变量A的分解效果，取值过小会导致调整变量分解不充分，过大易引入噪声，进而影响因果效应的估计精度；λ作为整体模型参数的L_2正则化系数，主要用于控制模型复杂度，避免过拟合，实验表明适当的λ能有效提升模型泛化能力，但过大的λ会导致模型欠拟合，如图 9(f)所示，当λ=1 时模型效果最好。相比之下，超参数β，γ_1，γ_2在实验取值范围内对最终预测精度的影响相对较小，但会对工具变量分离与混杂变量平衡起到必要作用；超参数μ通过正交约束促进变量解耦，进一步提升模型稳定性。综上，精确的变量分解与适度的正则化是提升模型性能的关键。

此外，为进一步探究模型在不同数据分布下的适应性与超参数选择规律，本文也在 IHDP 与 Twins

Table 5 $\varepsilon_{\text{PEHE}}$ Obtained in Ablation Study on IHDP Dataset

表 5 IHDP 数据集上消融实验得到的$\varepsilon_{\text{PEHE}}$

方法	样本内	样本外
-without-SMOTE	0.604±0.042	0.643±0.135
-without-$loss_A$	0.625±0.026	0.668±0.106
-without-$loss_I$	0.588±0.030	0.604±0.075
-without-$\mathcal{L}_C$	0.632±0.039	0.641±0.089
-without-$\mathcal{L}_O$	0.527±0.033	0.555±0.098
本文	**0.503±0.019**	**0.525±0.045**

注：黑体数值表示最优值。“±”前数值表示平均值，“±”后数值表示误差。

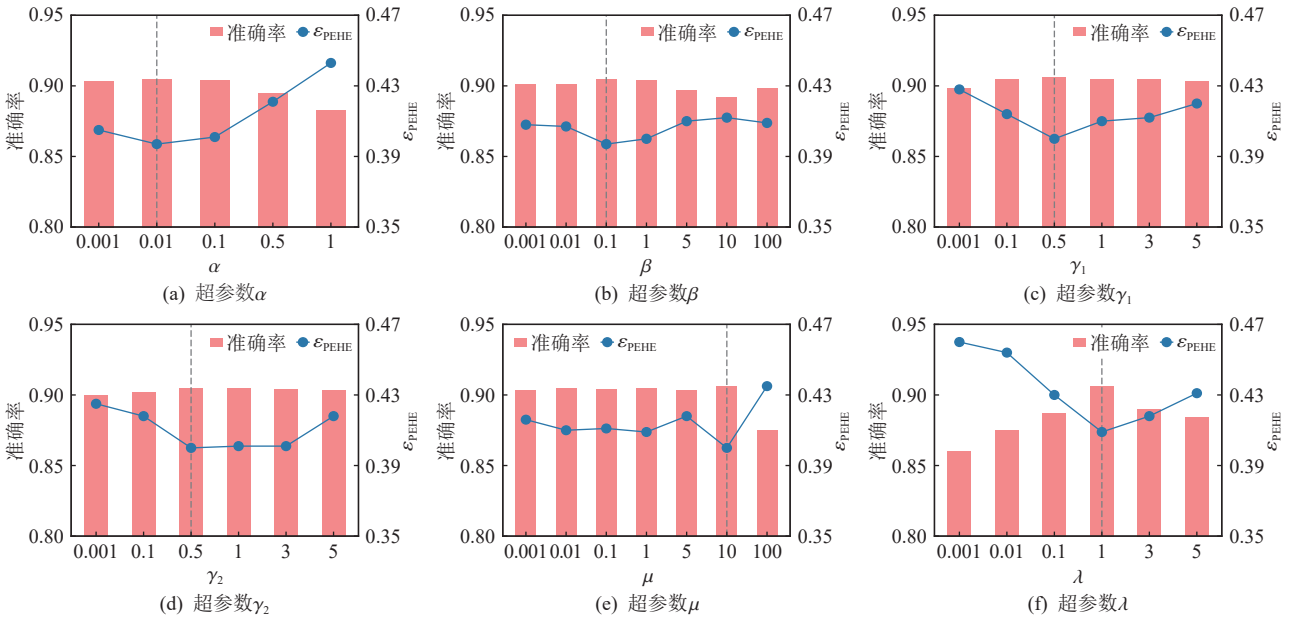


Fig. 9 Hyperparameter sensitivity analysis

图 9 超参数敏感性分析

这2个基准数据集上进行了系统的参数调优,所得最优超参数组合如表6所示。该结果不仅印证了前述敏感性分析中关于α与λ关键作用的结论,也体现出不同数据结构对超参数取值的差异性要求,为模型在实际场景中的应用提供了可靠的参数配置依据。

Table 6 Optimal Hyperparameters

表6 最优超参数

数据集	取值
IHDP	{0.5, 100, 1, 1, 10, 0.01}
Twins	{0.01, 0.001, 0.5, 0.5, 5, 5}
合成数据集	{0.01, 0.1, 0.5, 0.5, 10, 1}

注:"{}"中分别是α, β, γ_1, γ_2, μ, λ的取值。

5 结 论

针对现有ICE估计方法的局限性,本研究提出了一种基于对抗解耦与分布对齐的ICE估计方法,该方法通过引入对抗机制优化分解表示学习与混杂变量平衡。与以往方法相比,本研究同时解决了混杂变量分离与组间样本数量不平衡问题,并利用对抗平衡策略避免了极端样本权重问题。具体而言:为缓解样本失衡,采用SMOTE算法对观测数据进行预处理,通过合成少数类样本实现不同组别间的样本数量平衡;为精确解耦混杂变量、工具变量与调整变量,在协变量分解过程中设计对抗约束模块,以约束组件间的概率关系来实现3类变量的有效分离;为缓解极端样本权重问题,采用对抗平衡网络对齐处理组与对照组的混杂变量分布。最后,基于解耦后的调整变量与平衡后的混杂变量进行ICE估计。在4个合成数据集与2个半合成数据集上的实验表明:本文方法通过精准解耦混杂变量、有效平衡组间分布以及缓解样本失衡,显著提升了观察性研究中ICE的估计精度。尽管如此,本文方法仍存在进一步优化的空间。目前方法主要针对组间分布不平衡,对组内样本代表性不足的问题尚未深入探讨,且当前方法仅适用于二元处理场景。然而,现实世界中的因果推断问题往往涉及更复杂的处理设置,例如多值处理、连续处理等。因此,将本文方法扩展为能同时应对组内与组间分布不平衡问题的多元处理效应估计框架,进一步提升现实场景中ICE估计的准确性与场景适应性,是未来值得深入研究的重要方向。

作者贡献声明:李欣娜负责提出算法模型,制定实验方案并完成实验,撰写和修改论文;曹付元和张佳璇提出算法思想和指导意见,并修改论文。

参 考 文 献

[1] Fernández-Loría C, Provost F. Causal decision making and causal effect estimation are not the same… and why it matters[J]. INFORMS Journal on Data Science, 2022, 1(1): 4−16

[2] Liu Dugang, Tang Xing, Gao Han, et al. Explicit feature interaction-aware uplift network for online marketing[C]//Proc of the 29th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2023: 4507−4515

[3] Kazemi A, Ester M. Adversarially balanced representation for continuous treatment effect estimation[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 13085−13093

[4] Li Jianing, Xiong Ruibin, Lan Yanyan, et al. Overview of the frontier progress of causal machine learning[J]. Journal of Computer Research and Development, 2023, 60(1): 59−84 (in Chinese)
(李家宁, 熊睿彬, 兰艳艳, 等. 因果机器学习的前沿进展综述[J]. 计算机研究与发展, 2023, 60(1): 59−84)

[5] Zhang Jiaxuan, Eldele E, Cao Fuyuan, et al. Counterfactual contrastive learning with normalizing flows for robust treatment effect estimation[C/OL]//Proc of the 42nd Int Conf on Machine Learning. New York: ACM, 2025[2025-07-19]. https://openreview.net/pdf?id=VmXgkCmLFD

[6] Li Yijun, Leung C H, Sun Xiangqian, et al. The causal impact of credit lines on spending distributions[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 180−187

[7] Jaki T, Chang Chi, Kuhlemeier A, et al. Predicting individual treatment effects: Challenges and opportunities for machine learning and artificial intelligence[J]. KI-Künstliche Intelligenz, 2025, 39(1): 27−32

[8] Martínez-Sánchez Á, Arranz G, Lozano-Durán A. Decomposing causality into its synergistic, unique, and redundant components[J]. Nature Communications, 2024, 15: 9296

[9] Yao Liuyi, Chu Zhixuan, Li Sheng, et al. A survey on causal inference[J]. ACM Transactions on Knowledge Discovery from Data, 2021, 15(5): 1−46

[10] Cheng G, Hardt M, Mendler-Dünner C. Causal inference out of control: Estimating performativity without treatment randomization [C]//Proc of the 41st Int Conf on Machine Learning. New York: ACM, 2024: 8077−8103

[11] Kong I, Park Y, Jung J, et al. Covariate balancing using the integral probability metric for causal inference[C]//Proc of the 40th Int Conf on Machine Learning. New York: ACM, 2023: 17430−17461

[12] Zhao Ziyu, Kuang Kun, Li Bo, et al. Differentiated matching for individual and average treatment effect estimation[J]. Data Mining and Knowledge Discovery, 2023, 37(1): 205−227

[13] Johansson F D, Shalit U, Kallus N, et al. Generalization bounds and representation learning for estimation of potential outcomes and causal effects[J]. Journal of Machine Learning Research, 2022, 23(166): 1−50

[14] Shi C, Blei D M, Veitch V. Adapting neural networks for the estimation of treatment effects[C]//Proc of the 33rd Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2019: 2503–2513

[15] Kuang Kun, Cui Peng, Li Bo, et al. Treatment effect estimation with data-driven variable decomposition[C]//Proc of the 31st AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 140−146

[16] Wu Anpeng, Yuan Junkun, Kuang Kun, et al. Learning decomposed representations for treatment effect estimation[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(5): 4989−5001

[17] 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

[18] Li Sheng, Vlassis N, Kawale J, et al. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns[C]//Proc of the 25th Int Joint Conf on Artificial Intelligence. San Francisco: Margan Kaufmann, 2016: 3768–3774

[19] Li Xiaohui, Cao C C, Shi Yuhan, et al. A survey of data-driven and knowledge-aware explainable AI[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(1): 29−49

[20] 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

[21] Athey S, Imbens G W, Wager S. Approximate residual balancing: Debiased inference of average treatment effects in high dimensions[J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2018, 80(4): 597−623

[22] Zubizarreta J R. Stable weights that balance covariates for estimation with incomplete outcome data[J]. Journal of the American Statistical Association, 2015, 110(511): 910−922

[23] Li Jiuyong, Ma Saisai, Le T, et al. Causal decision trees[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 29(2): 257−271

[24] He Haibo, Garcia E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263−1284

[25] VanderWeele T J. Principles of confounder selection[J]. European Journal of Epidemiology, 2019, 34(3): 211−219

[26] Kuang Kui, Cui Peng, Li Bo, et al. Treatment effect estimation via differentiated confounder balancing and regression[J]. ACM Transactions on Knowledge Discovery from Data, 2019, 14(1): 1−25

[27] Johansson F, Shalit U, Sontag D. Learning representations for counterfactual inference[C]//Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 3020−3029

[28] Shalit U, Johansson F D, Sontag D. Estimating individual treatment effect: Generalization bounds and algorithms[C]//Proc of the 34th Int Conf on Machine Learning. New York: ACM, 2017: 3076−3085

[29] Yao Liuyi, Li Sheng, Li Yaliang, et al. Representation learning for treatment effect estimation from observational data[C]//Proc of the 32nd Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2018: 8397−8408

[30] Hassanpour N, Greiner R. CounterFactual regression with importance sampling weights[C]//Proc of the 28th Int Joint Conf on Artificial Intelligence. San Francisco: Margan Kaufmann, 2019: 5880−5887

[31] Yoon J, Jordon J, Van Der Schaar M. GANITE: Estimation of individualized treatment effects using generative adversarial nets[C/OL]//Proc of the 6th Int Conf on Learning Representations. 2018[2024-09-26]. https://openreview.net/pdf?id=ByKWUeWA-

[32] Bica I, Jordon J, van der Schaar M. Estimating the effects of continuous-valued interventions using generative adversarial networks[C]//Proc of the 34th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 16434−16445

[33] Du Xin, Sun Lei, Duivesteijn W, et al. Adversarial balancing-based representation learning for causal effect inference with observational data[J]. Data Mining and Knowledge Discovery, 2021, 35(4): 1713−1738

[34] Kuang Kun, Cui Peng, Zou Hao, et al. Data-driven variable decomposition for treatment effect estimation[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(5): 2120−2134

[35] Hassanpour N, Greiner R. Learning disentangled representations for counterfactual regression[C/OL]//Proc of the 7th Int Conf on Learning Representations. 2019[2025-06-28]. https://openreview.net/pdf?id=HkxBJT4YvB

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

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

[38] Hill J L. Bayesian nonparametric modeling for causal inference[J]. Journal of Computational and Graphical Statistics, 2011, 20(1): 217−240

[39] Almond D, Chay K Y, Lee D S. The costs of low birth weight[J]. The Quarterly Journal of Economics, 2005, 120(3): 1031−1083

[40] Shi Pengfei, Zhong Wei, Zhang Xinyu, et al. Estimating conditional average treatment effects via sufficient representation learning[C]// Proc of the 33rd Int Joint Conf on Artificial Intelligence. Cambridge, MA: MIT, 2024: 4894−4901



Cao Fuyuan, born in 1974. PhD, professor. Senior member of CCF. His main research interests include machine learning, causal inference, and data mining.

曹付元, 1974 年生。博士，教授。CCF 高级会员。主要研究方向为机器学习、因果推断、数据挖掘。



Li Xinna, born in 1999. Master candidate. Her main research interest includes causal inference.

李欣娜, 1999 年生。硕士研究生。主要研究方向为因果推断。



Zhang Jiaxuan, born in 1998. PhD candidate. Her main research interest includes causal inference.

张佳璇, 1998 年生。博士研究生。主要研究方向为因果推断。