基于增量观测数据的持续因果推断

Zhixuan Chu蚂蚁集团中国杭州
chuzhixuan.czx@alibaba-inc.com

Ruopeng Li蚂蚁集团中国杭州
ruopeng.lrp@antgroup.com

Stephen Rathbun佐治亚大学美国雅典
rathbun@uga.edu

Sheng Li弗吉尼亚大学美国夏洛茨维尔
shengli@virginia.edu

arXiv:2303.01775v1 [cs.LG] 3 Mar 2023

摘要—大数据时代见证了来自移动设备与社交网络、在线广告、网络挖掘、医疗保健、教育、公共政策、营销活动等领域的观察性数据日益丰富，这推动了因果效应估计的发展。尽管学术界在应对反事实结果缺失和选择偏差等挑战方面取得了显著进展，但相关研究仅关注特定来源且平稳的观测数据，这在大多数工业应用中并不现实。本文探讨了一种新的工业问题：如何利用逐步获取的观测数据估计因果效应，并据此提出三项新的评估标准，包括可扩展性、适应性和可访问性。我们提出一种持续因果效应表示学习方法，用于估计来自非平稳数据分布、逐步获取的观测数据中的因果效应。该方法并不需要访问所有已见观测数据，而只存储从先前数据中学习到的有限特征表示子集。通过结合选择性平衡表示学习、特征表示蒸馏和特征变换，该方法能够对新数据进行持续因果效应估计，同时不损害对原始数据的估计能力。大量实验表明，持续因果效应估计意义重大，且验证了所提方法的有效性。

索引词—因果推断、持续学习、增量学习、观察性数据

I. 引言

深入理解观察性数据中的因果关系，对于经济学、医疗保健、公共政策、网络挖掘、在线广告和营销活动等众多领域至关重要 [1], [32]。尽管在克服观察性数据因果效应估计中的挑战方面已取得重大进展，例如反事实结局缺失以及处理组和对照组之间的选择偏差，但现有方法仅关注特定来源的平稳观察性数据。这些学习策略假设所有观察性数据在训练阶段都已可用，且仅来自一个来源 [4], [5], [18], [20]。

随着工业应用领域的快速发展，这一假设在实践中并不成立。这里以支付宝为例，它是全球最大的移动支付平台之一，为数十亿用户提供金融服务。每天都会产生并收集海量数据，其中包含大量与隐私相关的信息，这些数据来自不同城市或国家。总而言之，可以归纳为以下两点。第一点基于观察性数据的特征：这些数据会从非平稳数据分布中逐步产生。例如，某项营销活动的电子金融记录数量每天都在增长，或者该活动的电子金融记录可能来自不同城市，甚至不同国家。这一特征意味着，人们无法在同一时间点从单一来源获取所有观察性数据。第二点基于对可访问性的现实考量。例如，当有新的观察性数据可用时，如果我们想用新数据改进先前基于原始数据训练的模型，原始训练数据可能会因各种原因而无法再访问，例如，旧数据可能未被记录、属于专有数据、涉及金融数据敏感性、存储成本过高，或受到个人信息隐私约束 [34]。在各种学术和工业应用中，这种可访问性方面的实际问题普遍存在。归根结底，在大数据时代，我们在基于观察性数据进行因果推断时面临新的挑战：如何针对逐步可用的观察性数据实现**可扩展性**，如何在处理组和对照组不平衡问题之外，应对额外的领域自适应问题并实现**适应性**，以及如何应对海量数据的**可访问性**问题 [2]。

然而，现有的因果效应估计方法无法应对上述新挑战，即可扩展性、适应性和可访问性。尽管可以调整现有的因果推断方法以应对这些新挑战，但这些调整后的方法仍然存在不可避免的缺陷。以下介绍三种直接的调整策略：(1) 如果我们直接将基于原始数据预先训练的模型应用于新的观察性数据，由于不同数据源之间存在领域偏移问题，模型在新任务上的性能将非常差；(2) 如果我们利用新获取的数据重新训练先前学习的模型，以适应数据分布的变化，旧知识将被新知识完全或部分覆盖，从而导致模型在旧任务上的性能严重下降。这就是众所周知的 灾难性遗忘问题 [7], [22]；(3) 为了克服灾难性遗忘问题，我们可以依赖旧数据的存储，将旧数据和新数据结合起来，然后从头开始重新训练模型。然而，这种策略在记忆方面效率低下且耗时——

而且，长期存储数据还会带来版权或隐私问题等实际 concerns [26]。我们在第四节中的实证评估表明，将这三种策略中的任何一种与现有的因果效应估计方法结合使用都存在缺陷。

为解决上述问题，我们提出了一种**持**续因果**效**应**表**征**学**习方法（CERL），用于利用增量获取的观察性数据估计因果效应。我们无需访问所有先前的观察性数据，而只存储从先前数据中学习到的有限子集特征表示。通过结合选择性平衡表示学习、特征表示蒸馏和特征变换，我们的方法保留从先前数据中学到的知识，并利用新数据更新这些知识，从而能够针对增量新增数据持续进行因果效应估计，同时不损害对先前数据的估计能力。

总而言之，我们的主要贡献包括：（1）我们首次针对增量获取的观察性数据提出持续终身因果效应估计问题，并提出了三个相应的评估标准，即可扩展性、适应性和可访问性；（2）我们提出了一种基于深度表示学习和持续学习的持续终身因果效应估计新框架；（3）大量实验表明，现有方法在面对增量获取的观察性数据时存在不足，同时也证明了我们模型的优越性。

II. 背景与问题陈述

假设观察性数据包含从 d 个不同领域收集的 n 个体，且第 d 个数据集 D_d 包含从第 d 个领域收集的数据 $\{(x,\ y,\ t)|x \in X,\ y \in Y,\ t \in T\}$，该领域包含 n_d 个体。令 X 表示所有观测变量，Y 表示观察性数据中的结果，T 为二元变量。令 $D_{1:d} = \{D_1, D_2, ..., D_d\}$ 表示分别从 d 个不同领域收集的 d 个数据集的组合集合。对于 d 个数据集 $\{D_1, D_2, ..., D_d\}$，它们具有共同的观测变量，但由于数据集来自不同领域，因此每个数据集中关于 X、Y 和 T 的分布各不相同。观察性数据中的每个个体都接受了两种治疗之一。令 t_i 表示个体 i 的处理分配；$i = 1,$ ..... n。对于二元治疗，$t_i = 1$ 表示处理组，$t_i = 0$ 表示对照组。个体 i 在对其施加治疗 t 时的结局记为 y_t^i。对于观察性数据，只能观测到一个潜在结果。观测到的结果称为事实结果，其余未观测到的潜在结果称为反事实结果。

本文采用潜在结果框架来估计处理效应 [25], [29]。个体 i 的个体治疗效应 (ITE) 是接受治疗和接受对照时的潜在结果之差，定义为 $\text{ITE}_i = y_1^i - y_0^i$。平均治疗效应 (ATE) 是接受治疗和接受对照时的平均潜在结果之差，定义为 $\text{ATE} = \frac{1}{n}\sum_{i=1}^{n}(y_1^i - y_0^i)$。

潜在结果框架的成功建立在以下假设之上 [13]，这些假设确保处理效应可以被识别。**稳定个体处理值假设 (SUTVA)**：任何个体的潜在结果不会因分配给其他个体的治疗而改变；并且对于每个个体，每个治疗水平都不存在会导致不同潜在结果的不同形式或版本。**一致性**：治疗 t 的潜在结果等于实际接受的治疗为 t 时观测到的结果。**正值性**：对于 x的任意取值，处理分配不是确定性的，即对于所有 t 和x，$P(T = t|X = x) > 0$。**可忽略性**：给定协变量后，处理分配与潜在结果相互独立，即 $(y_1,\ y_0) \perp\!\!\!\perp t|x$。

我们的目标是开发一种新颖的持续因果推断框架，在无法访问先前数据 $D_{1:(d-1)}$的情况下，估计新获取的数据 D_d 和先前数据 $D_{1:(d-1)}$ 的因果效应。

III. 所提出的框架

工业应用中观察性数据的可用性有望促进因果效应估计模型的发展。我们以支付宝为例，它是全球最大的移动支付平台之一，为数十亿用户提供金融服务。每天都会产生大量包含诸多隐私相关信息的数据，并从不同城市或国家收集这些数据。因此，如何从非平稳数据域中逐步获取的观察性数据实现持续学习，成为因果效应估计领域的一个新方向。除了关注如何处理处理组和对照组之间的选择偏差问题之外，我们还需要全面考虑模型的三个方面，即针对逐步获取的观察性数据的**可扩展性**、针对各种数据源的**适应性**，以及针对海量数据的**可访问性**。

我们提出持续因果效应表征学习方法（CERL），用于利用逐步获取的观察性数据估计因果效应。CERL 基于用于处理效应估计的选择性平衡表示学习，并融入特征表示蒸馏，以保留从先前观察性数据中学到的知识。此外，为了在无法访问原始数据的情况下使更新后的模型适应原始数据和新数据，并解决处理组和对照组之间的选择偏差，我们提出了一种表示变换函数，将部分原始特征表示映射到新的特征表示空间，使全局特征表示空间在处理组和对照组之间实现平衡。因此，CERL 无需借助原始数据，即可对新数据进行持续因果效应估计，同时保留对先前数据的估计能力。

A. 模型架构

为了估计增量获取的观察性数据，CERL 框架主要由两个组

成部分：(1) 基线因果效应学习模型仅用于首批可用的观察性数据，因此无需考虑不同数据源之间的领域偏移问题。这部分等同于传统的因果效应估计问题；(2) 持续因果效应学习模型用于依次获取的观察性数据，需要处理更复杂的问题，例如知识迁移、灾难性遗忘、全局表示平衡和内存约束。

1) 基线因果效应学习模型：我们首先描述初始观察性数据集的基线因果效应学习模型，然后引入后续数据集。对于初始数据集中的因果效应估计，可以将其转化为传统的因果效应估计问题。受深度表示学习在反事实估计中取得的实证成功启发 [3], [28], ，我们提出学习处理组和对照组中样本的选择性平衡特征表示，然后基于学习到的表示空间推断潜在结果。

学习选择性平衡表示。首先，我们采用一种深度特征选择模型，使得变量选择能够在一个深度神经网络中实现，即 g_{w_1} ： $X \to R$，其中 X 表示原始协变量空间， R 表示表示空间， w_1 表示函数 g中的可学习参数。我们的模型采用弹性网络正则化项 [35]

$$L_{w_1} = \|w_1\|_2^2 + \|w_1\|_1. \tag{1}$$

贯穿全连接表示层的弹性网络会为重要特征分配更大的权重。该策略能够有效过滤无关变量，并突出重要变量。

由于处理组和对照组之间以及不同连续数据源之间存在选择偏差，混杂因素的幅度可能存在显著差异。为有效消除处理组和对照组之间以及不同数据源之间幅度显著差异所造成的不平衡，我们提议在最后的表示层使用余弦归一化。在多层神经网络中，传统上我们会计算先前层的输出向量与输入权重向量之间的点积，然后将点积结果输入激活函数。点积的结果没有界限。余弦归一化在神经网络中使用余弦相似度代替简单的点积，从而将预激活值限制在−1 与 1 之间。维度较高时，结果可能会更小。因此，方差可以控制在非常窄的范围 [21]内。余弦归一化定义如下

$$r = \sigma(r_{norm}) = \sigma\big(\cos(w, x)\big) = \sigma(\frac{w \cdot x}{|w|\,|x|}), \tag{2}$$

其中， r_{norm} 是归一化后的预激活值， w 是输入权重向量， x 是输入向量， σ 是非线性激活函数。

受 [28]，启发，我们在学习表征空间时采用积分概率度量 (IPM)，以平衡处理组和对照组。IPM 衡量处理组和对照组的表示分布之间的差异，因此我们希望最小化 IPM，使这两个分布更加相似。令 $P(g(x)|t = 1)$ 和 $Q(g(x)|t = 0)$ 分别表示处理组和对照组的表示向量的经验分布。我们采用定义在 1-Lipschitz 函数族上的 IPM，因此 IPM 即为 Wasserstein距离 [28], [30]。具体而言，基于 Wasserstein距离的 IPM 项定义为

$$\text{Wass}(P, Q) = \inf_{k \in \mathcal{K}} \int_{g(x)} \|k(g(x)) - g(x)\| P(g(x)) d(g(x)), \tag{3}$$

其中， $\mathcal{K} = \{k|Q(k(g(x))) = P(g(x))\}$ 定义了推前函数集，用于将治疗分布 P 的表示分布变换为对照 Q 的表示分布，且 $g(x) \in \{g(x)_i\}_{i:t_i=1}$。

推断潜在结果。我们的目标是学习一个函数h_{θ_1}：$R \times T \to Y$，将表示向量和处理分配映射到相应的观测结果，并且该函数可由深度神经网络参数化。为避免失去 T对 R的影响， $h_{\theta_1}(g_{w_1}(x)\ t)$被划分为两个独立任务，分别对应处理组和对照组。每个个体仅在与其观测到的治疗相对应的任务中更新。令$\hat{}y_i = h_{\theta_1}(g_{w_1}(x)\ t)$表示个体 i 在事实处理 t_i下推断出的观测结局。我们通过最小化事实结果预测的均方误差

$$L_Y = \frac{1}{n_1} \sum_{i=1}^{n_1} (\hat{y}_i - y_i)^2. \tag{4}$$

综合以上内容，我们的基线因果效应学习模型的目标函数为

$$L = L_Y + \alpha Wass(P, Q) + \lambda L_{w_1}, \tag{5}$$

其中， α 和 λ 表示用于控制目标函数中 $Wass$（P、Q）、L_{w_1}和 L_Y之间权衡的超参数。

*2)*模型学习的可持续性：到目前为止，我们已经利用来自单一来源的观察性数据构建了用于因果效应估计的基线模型。为了在学习新数据时避免灾难性遗忘，我们提议保留一部分低维特征表示，而不是保留所有原始协变量。我们还可以根据内存约束调整所保留特征表示的数量。

基线模型训练完成后，我们将一部分特征表示 $R_1 = \{g_{w_1}(x)|x \in D_1\}$ 及其对应的 $\{Y, T\} \in D_1$ 存储为记忆 M_1。我们可以通过聚类算法 [23], [31] 减少所存储表示向量的数量，以满足预先指定的内存约束。与随机子采样相比，聚类算法能够从分布中构建具有代表性的样本集合，并且只需更少的样本即可达到较高的近似质量。我们

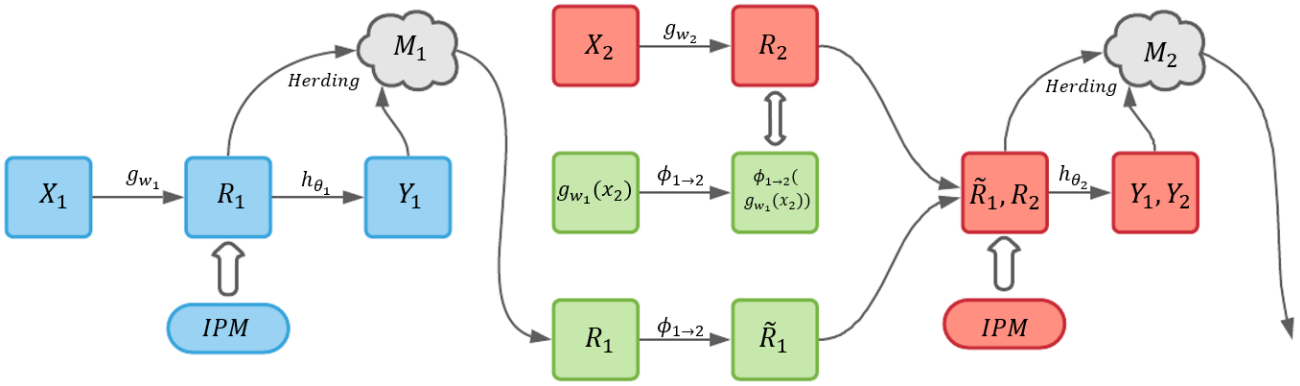

图 1. 蓝色部分是针对第一批观察性数据的基线因果效应学习模型。完成基线模型训练后，通过聚类算法将一部分特征表示 R_1 存入 M_1。绿色部分用于将 R_1 映射为经过变换的特征表示 $\tilde{R}_1$，使其与新的特征表示空间 R_2兼容。红色部分则基于特征蒸馏和平衡全局特征表示学习，针对 $\tilde{R}_1$ 和 R_2进行持续因果效应估计。

分别对处理组和对照组运行聚类算法，从处理组和对照组中存储数量相同的特征表示。此时，我们只存储记忆集 M_1 和模型 g_{w_1}，不存储原始数据D_1。

3) 持续因果效应学习：我们已存储记忆 M_1 和基线模型。为了对逐步获取的观察性数据持续估计因果效应，我们结合特征表示蒸馏与特征表示变换，在平衡的全局特征表示空间中，基于所有已见数据估计因果效应。CERL 的框架如图 1 所示。

特征表示蒸馏。对于从第二个领域收集到的下一个可用数据集 $D_2 = \{(x,\ y,\ t)|x \in X,\ y \in Y,\ t \in T\}$，我们采用相同的选择性表示学习方法 $g_{w_2}: X \to R_2$，对新参数 w_2进行弹性网络正则化（L_{w_2}）。由于我们希望模型能够估计先前数据和新数据的因果效应，因此希望新模型能从先前模型中继承一些知识。在持续学习中，通常采用知识蒸馏 [11],[19] 来缓解灾难性遗忘，其做法是鼓励原网络和新网络的输出相似，从而将知识从一个网络迁移到另一个网络。然而，对于持续因果效应估计问题，我们更关注特征表示，这些表示需要在治疗组与对照组之间，以及不同数据领域之间保持平衡。受[6], [12], [14], 启发，我们提出特征表示蒸馏，通过欧几里得距离促使基于基线模型的表示向量 $\{g_{w_1}(x)|x \in D_2\}$与基于新模型的表示向量 $\{g_{w_2}(x)|x \in D_2\}$ 相似。这种特征蒸馏有助于防止学习到的表示在新的特征表示空间中发生过度漂移。由于我们对特征表示进行了余弦归一化，并且 $\|A-B\|^2 = (A-B)\cdot(A-B) = \|A\|^2+\|B\|^2-2A^{\mathsf{T}}B = 2\big(1-cos(A,\ B)\big)$，

特征表示蒸馏定义如下：

$$L_{FD}(x) = 1 - cos\big(g_{w_1}(x), g_{w_2}(x)\big), \text{where } x \in D_2. \quad (6)$$

特征表示变换。我们已将先前的特征表示 R_1 存储在 M_1 中，并从新获取的数据中提取了新的特征表示 R_2。R_1和 R_2 位于不同的特征表示空间中，彼此不兼容，因为它们是由不同的模型学习得到的。此外，由于我们无法再访问先前的数据，因此无法通过新模型 g_{w_2}学习先前数据的特征表示。因此，为了平衡治疗组和对照组中包含先前表示与新表示的全局特征表示空间，需要一个特征变换函数，将先前的特征表示 R_1 转换为经过变换的特征表示 $\tilde{R}_1$，使其与新的特征表示空间 R_2兼容。我们将特征变换函数定义为 $\phi_{1\to2}: R_1 \to R_1$。我们还将由旧模型学习得到的新数据特征表示 D_2，即 $g_{w_1}(x)$，作为输入，以获得新数据经过变换的特征表示，即 $\phi_{1\to2}(g_{w_1}(x))$。为了使变换后的空间与新的特征表示空间兼容，我们通过使 $\phi_{1\to2}(g_{w_1}(x))$与 $g_{w_2}(x)$相似来训练变换函数$\phi_{1\to2}$，其中$x \in D_2$。损失函数定义为

$$L_{FT}(x) = 1 - cos\big(\phi_{1\to2}(g_{w_1}(x)), g_{w_2}(x)\big), \quad (7)$$

用于训练函数 $\phi_{1\to2}$，以在不同特征空间之间转换特征表示。然后，我们可以获得转换后的旧特征表示 $\tilde{R}_1 = \phi_{1\to2}(R_1)$，其与 R_2处于同一空间。

平衡全局特征表示空间。我们得到了一个全局特征表示空间，其中包含存储的旧数据经变换后的特征表示，以及新获取的数据的新特征表示。我们采用与基线模型相同的积分概率度量，以确保

在全局特征表示空间中，处理组和对照组的表示分布保持平衡。此外，我们定义一个潜在结果函数h_{θ_2}：（$\tilde{R}_1$，R_2）$\times T \to Y$。令 $\hat{y}_i^M = h_{\theta_2}(\phi_{1\to 2}(r_i)\ t)$，其中 $r_i \in M_1$，且 $\hat{y}_j^D = h_{\theta_2}(g_{w_2}(x_j)\ t)$，其中 $x_j \in D_2$ 表示推断的观测结果。我们的目标是最小化对全局特征表示（包括经变换的旧特征表示和新特征表示）预测事实结果时的均方误差

$$L_G = \frac{1}{\tilde{n}_1}\sum_{i=1}^{\tilde{n}_1}(\hat{y}_i^M - y_i^M)^2 + \frac{1}{n_2}\sum_{j=1}^{n_2}(\hat{y}_j^D - y_j^D)^2, \quad (8)$$

其中，$\tilde{n}_1$ 表示通过聚类算法存储在 M_1 中的样本数量，$y_i^M \in M_1$和 $y_j^D \in D_2$。

综上，我们的持续因果效应学习模型的目标函数为

$$L = L_G + \alpha Wass(P,Q) + \lambda L_{w_2} + \beta L_{FD} + \delta L_{FT}, \quad (9)$$

其中，α、λ、β和 δ 表示控制最终目标函数中 $Wass$（P,Q）、L_{w_2}、L_{FD}、L_{FT}和 L_G之间权衡的超参数。

B. CERL 概述

在上述章节中，我们介绍了基线因果效应学习模型和持续因果效应学习模型。训练完针对第二批数据的持续因果效应学习模型后，我们可以提取 $R_2 = \{g_{w_2}(x) | x \in D_2\}$ 以及 $R_1 = \{\phi_{1\to 2}(r) | r \in M_1\}$。我们将新的记忆集定义为 $M_2 = \{R_2, Y_2, T_2\} \cup \phi_{1\to 2}(M_1)$，其中 $\phi_{1\to 2}(M_1)$ 包含 R_1 以及存储在 M_1中的相应 $\{Y,T\}$。同样，为满足预先设定的内存约束，可以通过执行聚类算法来缩减 M_2，从处理组和对照组中存储相同数量的特征表示。

我们只存储新记忆集 M_2 和新模型 g_{w_2}，用于训练后续模型并平衡全局特征表示空间。无需再存储原始数据 D_1 和 D_2。

对于后续可用的观察性数据，我们也遵循相同的流程。当获得新的观察性数据D_d时，我们可以基于持续因果效应学习模型训练 $h_{\theta_d}(g_{w_d})$和 $\phi_{d-1\to d}$：$R_{d-1} \to R_{d-1}$。此外，新记忆集定义为：$M_d = \{R_d, Y_d, T_d\} \cup \phi_{d-1\to d}(M_{d-1})$。至此，我们的模型 $h_{\theta_d}(g_{w_d})$可以估计所有已见观察性数据的因果效应，而不受数据来源影响，且无需访问先前数据。如算法1所示，我们总结了 CERL 的流程。

算法1： 持续因果效应表示学习

数据： 给定 d逐步获取的观察性数据，从 D_1 到 D_d
如果 $\{x,y,t\} \in D_1$，**那么**
  *** 训练基线因果效应模型 $h_{\theta_1}(g_{w_1})$ ***
  $w_1,\ \theta_1 =$ 优化$(L_Y + \alpha Wass(P,\ Q) + \lambda L_{w_1})$
  $R_1 = \{g_{w_1}(x) | x \in D_1\}$
  $M_1 =$ 聚类选择$\{R_1, Y_1, T_1\}$
else
  对于 $\{x,y,t\} \in D_2, ..., D_d$**执行**
    *** 训练持续因果效应模型 $h_{\theta_d}(g_{w_d})$ ***
    $w_d, \theta_d, \phi_{d-1\to d} =$ 优化$(L_G + \tilde{\alpha} Wass(P,\ Q) + \lambda L_{w_2} + \beta L_{FD} + \delta L_{FT})$
    $R_{d-1} = \phi_{d-1\to d}(R_{d-1})$
    $R_d = \{g_{w_d}(x) | x \in D_d\}$
    $M_d =$ 聚类选择$(\{R_d, Y_d, T_d\} \cup \{R_{d-1},\ Y_{d-1} \in M_{d-1},\ T_{d-1} \in M_{d-1}\})$
  end
end

IV. 实验

我们采用半合成基准数据集，即 News [16], [27] 和 BlogCatalog [8]，进行持续因果效应估计。这些数据不包含任何个人可识别信息或攻击性内容。具体而言，我们考虑了三种情景，以表示增量可用的观察性数据之间不同程度的领域偏移，包括显著偏移、中等偏移和无偏移。

此外，我们生成了一系列合成数据集，并开展消融研究，以展示模型在序列数据集上的有效性。同时，我们还评估了保留不同数量特征表示时的性能，以及超参数的稳健性。

A. 数据集描述

News。News数据集由从《纽约时报》语料库[1]中随机抽取的5000篇新闻文章组成，用于模拟媒体消费者对新闻条目的看法。样本是不同的新闻条目，以词频 $x_i \in \mathbb{N}^V$ 和结局 $y(x_i) \in \mathbb{R}$ 表示，其中结局是读者对新闻条目的意见。干预 $t \in \{0,1\}$ 表示观看设备，即台式电脑（$t=0$）或移动设备（$t=1$）。我们扩展了 [16], [27] 中的原始数据集规范，以便模拟逐步获取的、具有不同程度领域偏移的观察性数据。假设消费者更喜欢在特定观看设备上阅读某些媒体内容，我们在大量文档上训练主题模型，并将新闻条目 x的主题分布定义为 $z(x)$。我们将随机抽取的一个文档的主题分布定义为移动设备的质心 z_1^c，并将所有文档的平均主题表示定义为台式电脑的质心z_0^c。因此，读者在设备 t 上对新闻条目x 的意见取决于 $z(x)$与z_t^c之间的相似度，即 $y(x_i) = C(z(x)\ z_0^c + t_i \cdot z(x)\ z_1^c) + \epsilon$，其中

[1]https://archive.ics.uci.edu/ml/datasets/bag+of+words

表 I

两种序列数据和 M=500 上的性能。我们给出了来自两个数据集的测试集上 $\sqrt{\epsilon_{\text{PEHE}}}$ 和 ϵ_{ATE} 的均值。标准差很小。**越低越好**。**在无领域偏移的情形下**，三种策略和 CERL 的性能相近，因为先前数据和新数据来自相同的分布。**在显著偏移和中度偏移的情形下**，CFR-A 在先前数据上的表现良好，但在新数据集上显著下降；策略 CFR-B 出现了灾难性遗忘问题；CERL 的表现与策略 CFR-C 相近，但 CERL 无需访问先前数据。此外，领域偏移越大，CFR-A 和 CFR-B 的表现越差。CERL 在面对偏移时仍保持稳定。越高越差。"↑"表示该模型的性能指标相较于 CERL 出现了统计学上的显著下降。

		News					**BlogCatalog**					**性能摘要**		
		先前数据			新数据		先前数据			新数据		先前	New	记忆
	策略	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}		$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}		$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	data	data	
显著偏移	CFR-A	2.49	0.80		**3.62**	**1.18**	↑ 9.92	4.25		**13.65**	**6.21**	↑ ✓	×	✓
	CFR-B	**3.23**	**1.06**	↑	2.71	0.91	**14.21**	**6.98**	↑	9.77	4.11	×	✓	✓
	CFR-C	2.51	0.82		2.70	0.92	9.93	4.24		9.77	4.12	✓	✓	×
	CERL	2.55	0.84		2.71	0.91	9.96	4.25		9.78	4.12	✔	✔	✔
中等偏移	CFR-A	2.58	0.85		**3.06**	**1.02**	↑ 9.89	4.22		**11.26**	**5.03**	↑ ✓	×	✓
	CFR-B	**2.98**	**0.99**	↑	2.65	0.92	**12.35**	**5.67**	↑	9.83	4.18	×	✓	✓
	CFR-C	2.56	0.85		2.63	0.90	9.88	4.21		9.81	4.16	✓	✓	×
	CERL	2.59	0.86		2.66	0.92	9.90	4.24		9.82	4.17	✔	✔	✔
No 转变	CFR-A	2.58	0.87		2.62	0.88	9.86	4.20		9.85	4.19	✓	✓	✓
	CFR-B	2.60	0.88		2.60	0.87	9.85	4.18		9.83	4.18	✓	✓	✓
	CFR-C	2.58	0.87		2.59	0.87	9.84	4.18		9.83	4.18	✓	✓	×
	CERL	2.59	0.87		2.60	0.87	9.85	4.19		9.83	4.18	✔	✔	✔

$C = 60$ 是缩放因子，且 $\epsilon \sim N(0, 1)$。此外，干预 t 定义为 $p(t=1|x) = \frac{e^{k \cdot z(x)^\intercal z_1^c}}{e^{k \cdot z(x)^\intercal z_0^c} + e^{k \cdot z(x)^\intercal z_1^c}}$，其中 $k = 10$ 表示预期的选择偏差。在实验中，我们从训练语料库中学习了50个LDA主题，数据集中包含3477个词袋特征。为生成具有不同领域偏移的两个序列数据集，我们将属于LDA主题1至25的新闻条目合并到第一个数据集中，并将属于LDA主题26至50的新闻条目合并到第二个数据集中。两个数据集的LDA主题没有重叠，因此被视为显著领域偏移。此外，我们分别使用属于LDA主题1至35的新闻条目和属于LDA主题16至50的新闻条目构建第一个和第二个数据集，这被视为中度领域偏移。最后，从50个LDA主题中随机抽取的条目组成第一个和第二个数据集，由此形成无领域偏移，因为它们来自同一分布。我们随机抽取60%的样本作为训练集，抽取20%作为验证集，其余样本作为测试集。

BlogCatalog。BlogCatalog [8]是一个管理博客作者及其博客的博客目录。每个个体都是一位博客作者，特征是从真实世界来源收集的博客作者描述中的关键词所构成的词袋表示。我们采用与 News 数据集相同的设置和假设来模拟处理选项和结果。从训练语料库中学习了 50 个 LDA 主题。该数据集包含 5196 个个体和 2160 个词袋特征。与生成存在领域偏移的 News 数据集的过程类似，我们为三种领域偏移场景中的每一种创建两个数据集。我们随机抽取 60% 和 20% 的个体作为训练集和验证集，并将其余个体作为测试集。

B. 结果与分析

评估指标。我们采用两种常用的评估指标。第一种是 ATE 估计误差，其定义为 $\epsilon_{\text{ATE}} = |\widehat{\text{ATE}} - \text{ATE}|$，其中 ATE 是真实值，而 $\widehat{\text{ATE}}$ 是估计得到的 ATE。第二种是异质性效应估计的期望精度误差（PEHE），[10], 其定义为 $\epsilon_{\text{PEHE}} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\text{ITE}}_i - \text{ITE}_i)^2$，其中 ITE_i 是个体 i 的真实个体处理效应，而 $\widehat{\text{ITE}}_i$ 是个体 i 的估计个体处理效应。

我们采用三种策略，将传统因果效应估计模型应用于逐步获取的观察性数据：(A) 将先前基于原始数据训练的模型直接应用于新获取的观察性数据；(B) 利用新获取的数据对先前学习的模型进行微调；(C) 保存所有先前数据，并与新数据结合，从头开始重新训练模型。在这三种策略中，预计 (C) 的表现最佳，并在 ATE 和 PEHE 方面达到理想效果，尽管它需要占用最多资源（先前数据集和新数据集中的全部数据）。我们基于反事实回归模型（CFR）[28],实现这三种策略；CFR 是一种具有代表性的因果效应估计方法。

如表 I 所示，在不存在领域偏移的情况下，三种策略与我们的模型在 News 和 BlogCatalog 数据集上的表现相近，因为先前数据和新数据来自相同的分布。CFR-A、CFR-B 和 CERL 所需资源少于 CFR-C。在显著偏移和中度偏移的情况下，我们发现 CFR-A 的表现

在先前数据上良好，但在新数据集上的表现显著下降；策略 CFR-B 则出现了灾难性遗忘问题，在先前数据集上的表现较差；策略 CFR-C 在先前数据和新数据上表现都很好，但它会使用先前数据和新数据重新训练整个模型。然而，如果存在内存约束或无法访问先前数据，便无法采用策略 CFR-C。我们的 CERL 与策略 CFR-C 的表现相近，同时 CERL 无需访问先前数据。此外，通过比较显著偏移和中度偏移情况下的表现，我们发现领域偏移越大，CFR-A 和 CFR-B 的表现越差。相比之下，无论领域偏移的程度如何，CERL 的表现都与理想策略 CFR-C 一致。

C. 进一步的模型评估

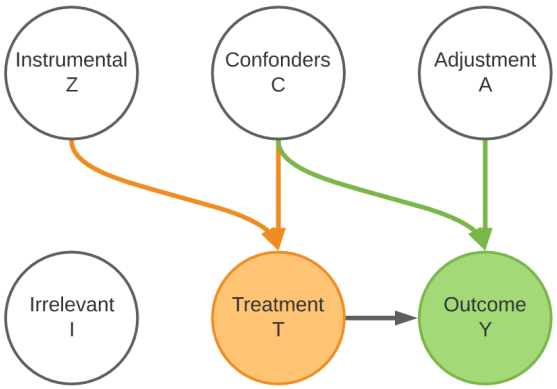

图 2. 变量、治疗与结局之间的相互关系。工具变量与处理分配相关，但除通过暴露外，与结局无关。调整变量可以预测结果，但与处理分配无关。混杂变量同时与治疗和结局相关。

合成数据集。 我们的合成数据包含混杂因素、工具变量、调整变量和无关变量。图 2 展示了这些变量、治疗与结果之间的相互关系。向量 $X = (C, Z, I, A)$ 中观测变量的数量设为 100，其中 C 中有 35 个混杂因素，A 中有 35 个调整变量，Z 中有 10 个工具变量，I 中有 20 个无关变量。本模拟中用于生成连续结局变量 Y 的模型为部分线性回归模型，该模型拓展了 [3], [15], [24] 中所述的思路：

$$Y = \tau((C^\intercal, A^\intercal)^\intercal)T + g((C^\intercal, A^\intercal)^\intercal) + \epsilon, \qquad (10)$$

其中 ϵ 是未观测协变量，服从标准正态分布 $N(0, 1)$，且 $E[\epsilon|C, A, T] = 0$，其中 $T \overset{ind.}{\sim}$ 伯努利分布$(e_0((C, Z)))$. $e_0((C, Z))$ 是倾向得分，表示基于其自身混杂因素 C 和工具变量 Z 的治疗选择偏倚。由于我们的目标是模拟多个数据源$\{\mathcal{D}_d;\ d = 1, \ldots\ldots, D\}$，因此所有观测协变量的向量$X = (C, Z, I, A)$ 从不同的多元正态分布中采样，其均值向量为 μ_C^d，μ_Z^d，μ_I^d，和 μ_A^d，协方差矩阵 Σ^d则随机生成且为正定矩阵。

对于每个数据源，除均值向量的大小和协方差矩阵的结构不同外，模拟过程均相同。令 D 为对角矩阵，其对角线元素为 Σ 对角线元素的平方根，即 $D = \sqrt{diag(\sigma)}$，则相关矩阵表示为：

$$R = D^{-1}\Sigma D^{-1}. \qquad (11)$$

我们使用 [9] 中的算法 3 来模拟由不同类型变量构成的正定相关矩阵。我们的相关矩阵基于枢纽相关结构，该结构中枢纽变量与其余每个变量之间的相关性已知 [17], [33]。每种类型中的各变量与枢纽变量的相关性从指定的最大相关性逐渐减弱至最小相关性；不同类型的变量则彼此独立生成，或在类型之间具有较弱的相关性。将第一个变量定义为枢纽变量，对于第i个变量（$i = 2, 3, \ldots\ldots, n$），同一类型中该变量与枢纽变量的相关性定义如下：

$$R_{i,1} = \rho_{\max} - \left(\frac{i-2}{d-2}\right)^{\gamma}(\rho_{\max} - \rho_{\min}), \qquad (12)$$

其中，$\rho_{\max}$ 和 $\rho_{\min}$ 分别是指定的最大相关性和最小相关性，而速率 γ 控制相关性的衰减速度。

指定枢纽变量与同类型其余变量之间的关系后，我们使用托普利兹结构补全枢纽相关矩阵的其余部分，并为其他类型的变量构造枢纽托普利兹相关矩阵 R_{type} 。这里，R 是一个$n \times n$矩阵，其对角线上包含块 R_Z，R_C，R_A，和 R_I，非对角线元素为零。由此得到的相关矩阵在同一类型内具有非零相关性，而不同类型之间的相关性为零。可加入正定相关矩阵 R的类型间相关性大小由其最小特征值决定。

函数 $\tau((C, A)^\intercal)$ 描述了真实处理效应如何随调整变量 A和混杂因素 C的取值而变化；即 $\tau((C, A))$ $=(\sin((C, A) \times b_\tau))^2$ 其中，b_τ 表示函数中每个协变量对应的权重，由 uniform(0，1) 生成。处理效应具有个体差异，意味着其强度因个体而异，因此取决于 C 和 A。无论处理分配如何，函数 $g((C, A))$ 都可能影响结局。该函数通过三角函数计算，使协变量呈非线性，其定义为 $g((C, A)) = (\cos((C, A) \times b_g))^2$。其中，$b_g$ 表示该函数中每个协变量对应的权重，由 uniform(0，1) 生成。偏差归因于未观测到的协变量，这些协变量服从随机正态分布$N(0, 1)$。处理分配 T服从伯努利分布

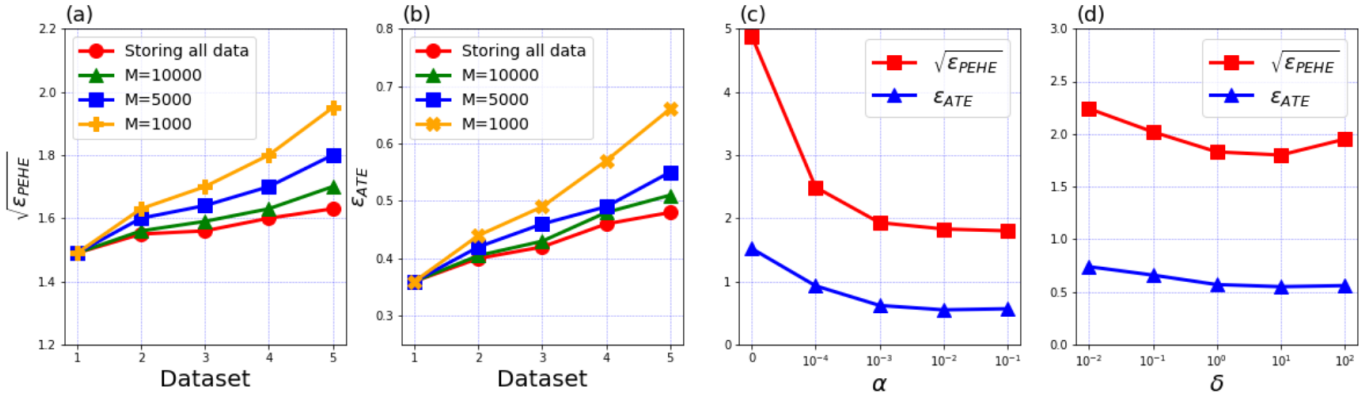

图 3. CERL 在不同设置下的性能。在 (a) 和 (b) 中，随着模型不断学习新数据集，每完成一个新数据集的训练，我们都会报告 $\sqrt{\epsilon_{\text{PEHE}}}$ 和 ϵ_{ATE} 在由先前数据和新数据构成的测试集上的结果。在 (c) 和 (d) 中，我们观察到，模型在很大的参数范围内性能稳定。

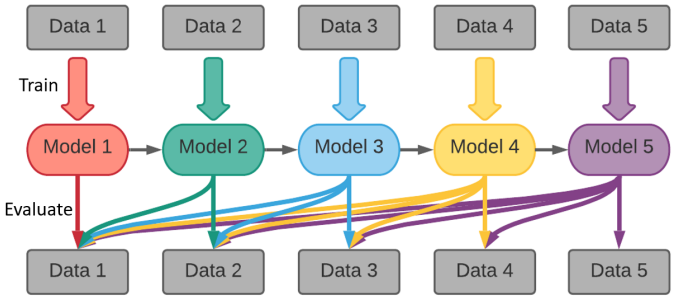

图 4. 五个观察性数据集按顺序逐步提供，模型将在无法访问先前数据的情况下持续估计因果效应。

分布，即 $T \overset{ind.}{\sim} \text{Bernoulli}(e_0((C, Z)))$，其概率为 $e_0((C, Z)) = \Phi(\frac{\mu(a)}{\sigma_a}())$，其中 $e_0((C, Z))$ 表示倾向得分，即基于混杂因素 C 和工具变量 Z 的标准正态随机变量的累积分布函数，也就是 $a = \sin((C, Z) \times b_a)$，其中 b_a 由 $\text{uniform}(0, 1)$ 生成。

我们通过五种不同的多元正态分布模拟五个不同的数据源，以表示逐步提供的观察性数据。在每个数据源中，我们随机抽取 10000 个样本，包括处理组样本和对照组样本。因此，五个数据集具有不同的选择偏差、协变量大小、变量的协方差矩阵，以及处理组和对照组样本数量。为确保对模型性能进行稳健估计，我们针对每个数据源重复模拟过程 10 次，得到 10 个合成数据集。

结果。 与 News 和 BlogCatalog 基准上的实验类似，我们仍使用两个序列数据集，在更复杂的合成数据上，通过三种策略将我们的模型与 CFR 进行比较。如表 II 所示，结果与 News 和 BlogCatalog 上的结论一致。我们的模型表现优于 CFR-A 和 CFR-B。CERL 的表现与 CFR-C 相当，而

表 II

在两个序列数据集和 $M = 10000$ 上的性能。我们给出来自两个数据集的测试集上 $\sqrt{\epsilon_{\text{PEHE}}}$ 和 ϵ_{ATE} 的均值。标准差很小。数值越低越好。结果与 News 和 BlogCatalog 上的结论一致。我们的模型表现优于 CFR-A 和CFR-B。CERL 的表现与 CFR-C 相当。此外，我们还进行了三项消融研究，以检验 CERL 中重要组件的有效性，即 CERL（不含 FRT）、CERL（不含聚类选择）和 CERL（不含余弦范数）。数值越高越差。“↑”表示与 CERL 相比，模型的性能在统计上显著下降。

	先前数据			新数据		
策略	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}		$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	
CFR-A	1.47	0.35		**2.51**	**0.73**	↑
CFR-B	**1.82**	**0.47**	↑	1.63	0.45	
CFR-C	1.49	0.36		1.62	0.44	
CERL	1.49	0.37		1.63	0.44	
不含 FRT	**1.71**	**0.43**	↑	1.63	0.44	
不含聚类选择	**1.57**	**0.40**	↑	1.63	0.44	
不含余弦	**1.51**	**0.38**	↑	1.65	0.44	

它不需要访问先前数据集中的原始数据。此外，我们还进行了三项消融研究，以检验 CERL 中重要组件的有效性，即 CERL（无 FRT）、CERL（无聚类选择）和 CERL（无余弦范数）。CERL（无 FRT）是去除特征表示变换后的简化版 CERL，其基于采用知识蒸馏的传统持续学习。由于先前的特征表示未被存储，也未被变换到新的特征空间中，因此我们仅利用新数据来平衡处理组和对照组之间的偏差。CERL（无聚类选择）采用随机子采样策略将样本选入记忆，而不是使用聚类算法。CERL（无余弦范数）移除了最后一个表示层中的余弦归一化。表 II 显示，与原始 CERL 相比，移除特征表示变换、聚类选择或余弦归一化模块中的任意一个后，性能都会变差。

更具体地说，移除特征表示变换后，$\sqrt{\epsilon_{\text{PEHE}}}$ 和 ϵ_{ATE} 大幅增加，这表明在持续学习任务中始终采用的知识蒸馏不足以实现持续因果效应估计。此外，使用聚类选择从处理组和对照组分布中选取代表性样本集合，对于特征表示变换至关重要。

CERL 性能评估。如图4所示，五批观察性数据按顺序逐步提供，模型在无法访问先前数据的情况下，仍将持续估计因果效应。我们进一步从三个方面评估 CERL 的性能，即内存约束的影响、余弦归一化的有效性，以及模型对超参数的鲁棒性。如图3 (a)和(b)所示，随着模型持续学习新数据集，每完成一个新数据集的训练，我们都会报告测试集上的 $\sqrt{\epsilon_{\text{PEHE}}}$ 和 ϵ_{ATE}；这些测试集由先前数据和新数据组成。带有内存约束的模型，其性能与理想情况相近；在理想情况下，所有数据均可用于从头训练模型。然而，我们的模型能够有效节省内存空间。例如，面对第五个数据集时，我们的模型仅存储1000、5000或10000个特征表示，而理想情况下则需要存储包含所有协变量的 $5 \times 10000 = 50000$ 条观测。对于余弦归一化，我们对 CERL （ M=5000，5 个数据集）进行了消融研究，在表示学习过程中移除了余弦归一化。我们发现，$\sqrt{\epsilon_{\text{PEHE}}}$ 从1.80上升到1.92，而 ϵ_{ATE} 则从0.55上升到0.61。接下来，我们探究模型对最重要参数 α 和 δ的敏感性，这两个参数控制表示平衡和表示变换。从图3 (c)和(d)可以看出，在较大的参数范围内，模型性能保持稳定。此外，特征表示蒸馏的参数 β 设为 1 [14], [23]。

V. 结论

尽管从学术角度来看，因果效应估计领域在克服相关挑战方面已经取得了显著进展，但基于观察性数据的工业应用始终更加复杂，也更具挑战性。不同于特定来源且平稳的观察性数据，大多数现实世界数据都是逐步获得的，并且来自非平稳数据分布。尤其是，我们还必须考虑现实中的可访问性问题。本文首次尝试研究终身因果效应估计问题，并提出了相应的评估标准。尽管大量实验表明了本文方法的优越性，但要围绕这一全新的问题构建全面的分析工具和相应的理论框架，仍需要付出不小的努力。

参考文献

[1] Zhixuan Chu、Jianmin Huang、Ruopeng Li、Wei Chu 和 Sheng Li。因果效应估计：最新进展、挑战与机遇。*arXiv* 预印本 *arXiv:2302.00848*，2023 年。

[2] Zhixuan Chu 和 Sheng Li。持续性处理效应估计：挑战与机遇。arXiv 预印本 arXiv:2301.01026，2023。[3] Zhixuan Chu、Stephen L Rathbun 和 Sheng Li。在选择性平衡表征空间中进行匹配以估计治疗效应。arXiv 预印本 arXiv:2009.06828，2020。[4] Zhixuan Chu、Stephen L Rathbun 和 Sheng Li。面向网络化观察数据处理效应估计的图信息最大化对抗学习。载于第27届ACM SIGKDD知识发现与数据挖掘会议论文集，第176–184页，2021。[5] Zhixuan Chu、Stephen L Rathbun 和 Sheng Li。利用真实世界数据学习信息最大化和领域无关表征以进行因果效应推断。载于2022年SIAM国际数据挖掘会议（SDM）论文集，第433–441页。SIAM，2022。[6] Prithviraj Dhar、Rajat Vikram Singh、Kuan-Chuan Peng、Ziyan Wu 和 Rama Chellappa。学习而不记忆。载于IEEE计算机视觉与模式识别会议论文集，第5138–5146页，2019。[7] Robert M French。联结主义网络中的灾难性遗忘。《认知科学趋势》，3(4):128–135，1999。[8] Ruocheng Guo、Jundong Li 和 Huan Liu。从网络化观察数据中学习个体因果效应。载于第13届国际网络搜索与数据挖掘会议论文集，第232–240页，2020。[9] Johanna Hardin、Stephan Ramon Garcia 和 David Golan。一种生成逼真相关矩阵的方法。《应用统计年鉴》，第1733–1762页，2013。[10] Jennifer L Hill。用于因果推断的贝叶斯非参数建模。《计算与图形统计学杂志》，20(1):217–240，2011。[11] Geoffrey Hinton、Oriol Vinyals 和 Jeff Dean。提炼神经网络中的知识。arXiv 预印本 arXiv:1503.02531，2015。[12] Saihui Hou、Xinyu Pan、Chen Change Loy、Zilei Wang 和 Dahua Lin。通过再平衡逐步学习统一分类器。载于IEEE计算机视觉与模式识别会议论文集，第831–839页，2019。[13] Guido W Imbens 和 Donald B Rubin。统计学、社会科学与生物医学科学中的因果推断。剑桥大学出版社，2015。[14] Ahmet Iscen、Jeffrey Zhang、Svetlana Lazebnik 和 Cordelia Schmid。通过特征适配实现内存高效的增量学习。arXiv 预印本 arXiv:2004.00713，2020。[15] Daniel Jacob、Wolfgang Karl H¨ardle 和 Stefan Lessmann。观察性研究中的群体平均处理效应。arXiv 预印本 arXiv:1911.02688，2019。[16] Fredrik Johansson、Uri Shalit 和 David Sontag。学习用于反事实推断的表征。载于国际机器学习会议论文集，第3020–3029页，2016。[17] Peter Langfelder、Bin Zhang 和 Steve Horvath。从层次聚类树定义簇：R语言的动态树切割软件包。《生物信息学》，24(5):719–720，2008。[18] Sheng Li 和 Yun Fu。在平衡非线性表征上进行匹配以估计治疗效应。《神经信息处理系统进展》，30，2017。[19] Zhizhong Li 和 Derek Hoiem。学习而不遗忘。《IEEE模式分析与机器智能汇刊》，40(12):2935–2947，2017。[20] Christos Louizos、Uri Shalit、Joris M Mooij、David Sontag、Richard Zemel 和 Max Welling。利用深度潜变量模型进行因果效应推断。《神经信息处理系统进展》，30，2017。[21] Chunjie Luo、Jianfeng Zhan、Xiaohe Xue、Lei Wang、Rui Ren 和 Qiang Yang。余弦归一化：在神经网络中使用余弦相似度而非点积。载于国际人工神经网络会议论文集，第382–391页。施普林格，2018。[22] Michael McCloskey 和 Neal J Cohen。联结主义网络中的灾难性干扰：序列学习问题。载于《学习与动机心理学》，第24卷，第109–165页。爱思唯尔，1989。[23] Sylvestre-Alvise Rebuffi、Alexander Kolesnikov、Georg Sperl 和 Christoph H Lampert。iCaRL：增量分类器与表征学习。载于IEEE计算机视觉与模式识别会议论文集，第2001–2010页，2017。[24] Peter M Robinson。根号n一致半参数回归。《计量经济学会期刊》，第931–954页，1988。

[25] Donald B Rubin。估计随机研究和非随机研究中的治疗因果效应。《教育心理学杂志》，66(5):688，1974。[26] Saeed Samet、Ali Miri 和 Eric Granger。隐私保护贝叶斯网络的增量学习。《应用软计算》，13(8):3657–3667，2013。[27] Patrick Schwab、Lorenz Linhardt 和 Walter Karlen。Perfect match：一种利用神经网络学习反事实推断表征的简单方法。*arXiv* 预印本 *arXiv:1810.00656*，2018。[28] Uri Shalit、Fredrik D Johansson 和 David Sontag。估计个体处理效应：泛化界与算法。载于国际机器学习会议，第3076–3085页。PMLR，2017。[29] Jerzy Splawa-Neyman、Dorota M Dabrowska 和 TP Speed。概率论在农业实验中的应用：原理论文，第9节。《统计科学》，第465–472页，1990。[30] Bharath K Sriperumbudur、Kenji Fukumizu、Arthur Gretton、Bernhard Sch¨olkopf、Gert RG Lanckriet 等。积分概率度量的经验估计。《电子统计学杂志》，6:1550–1599，2012。[31] Max Welling。通过聚类选择动态权重进行学习。载于第*26*届年度国际机器学习大会论文集，第1121–1128页，2009。[32] Liuyi Yao、Zhixuan Chu、Sheng Li、Yaliang Li、Jing Gao 和 Aidong Zhang。因果推断综述。《*ACM*数据知识发现汇刊》（*TKDD*），15(5):1–46，2021。[33] Bin Zhang 和 Steve Horvath。加权基因共表达网络分析的通用框架。《遗传学与分子生物学中的统计应用》，4(1)，2005。[34] Junting Zhang、Jie Zhang、Shalini Ghosh、Dawei Li、Serafettin Tasci、Larry Heck、Heming Zhang 和 C-C Jay Kuo。通过深度模型整合实现类别增量学习。载于*IEEE*冬季计算机视觉应用会议，第1131–1140页，2020。[35] Hui Zou 和 Trevor Hastie。通过弹性网络进行正则化与变量选择。《英国皇家统计学会杂志：*B*辑（统计方法）》，67(2):301–320，2005。