Personalized Treatment Selection using Causal Heterogeneity（基于因果异质性的个性化处理选择）

- ABSTRACT
  - 随机实验（也称为A/B测试或桶测试）在互联网行业中广泛用于测量在不同干预变量下对指标产生的影响。
  - A/B 测试找到表现出最佳性能的干预变量，然后使其成为整个群体的干预变量。
  - 然而，给定干预的效果可能因实验单元而异，处理选择的个性化方法可以大大改进通常的全局选择策略。
  - 在这项工作中，作者通过估计队列或成员级别的各种各样的干预变量所产生的效果来开发个性化框架，然后为通过（确定性或随机）约束优化获得的队列（或成员）选择最佳干预变量。
  - 作者对其提出的方法进行了双重评估。
    - 首先，进行模拟分析以研究在精心控制的设置下个性化干预的选择的效果。
    - 该模拟说明了所提出的方法之间的差异以及每种方法的适用性随着不确定性的增加。
    - 我们还通过一个与在 Linkedin 上提供通知相关的真实示例来证明该方法的有效性。
    - 该解决方案显著优于启发式解决方案和全局干预选择基准，从而在会员访问等顶级指标上取得了可观的胜利。
- INTRODUCTION
  - 在大型社交媒体平台中，每个成员都是几个随机实验的一部分，也称为 A/B 测试。
    - 他们的经验是由在每个实验中为他们选择的干预变量共同决定的。
    - 此类干预变量可能是不同的机器学习模型、复合推荐系统中的参数值选择和 UI 组件（例如，特定元素的字体大小、副本测试）。
    - 通常的做法是确定在整个人群中表现最佳的干预变量并将该变量推广到每个人。 我们将这种做法称为“全局分配”。
  - 全局分配可能是次优的。
    - 干预变量对个体成员（或成员群组）的影响可能非常不同。 例如营销电子邮件中的语气选择。
    - 平台的年轻用户（总体上）可能更喜欢非正式的语气，而年长的用户可能更喜欢更正式的语气。
    - 如果我们为不同的成员队列选择不同的复制变量，那么全局分配将使其中的一个组遭受更糟糕的体验。
    - 因此，个性化的干预选择可以实现更好的会员体验和更大的业务胜利。
    - 还有一个重要的副作用值得强调。
    - 当今社会我们正在努力打造更具包容性的体验。

    - 以个性化方式选择干预的能力对于改善代表性不足的群体的体验非常有帮助，尤其是当他们有不同的偏好时。
    - 由于平均效应主要取决于对大多数用户类别的影响，全局分配不仅会导致业务指标的丢失，还会（无意中）降低平台或产品的包容性。
- 让我们考虑一个通用的 A/B 测试，它设置了一个目标和一个护栏度量（均在全局用户级别），如图 1 所示。
    - Pareto边界是通过使用特定解决方案系列中的不同选择得出的。
    - 全局分配给出了较差的Pareto最优曲线，并且选择变量来分配给临时群组来达到更好的性能。
    - 差异将取决于群组的选择，因此明智地选择这些群组是我们的主要关注领域之一。
- 随机实验可用于通过估计变量对代理成员偏好（例如，总点击次数、总会话数等）的业务指标的因果影响来识别个人偏好。
    - 然而，传统的 A/B 测试只能为我们提供干预变量的平均效果。 获得变量的成员级因果效应极具挑战性，因为它从根本上是不可观察的。
    - 然而，最近的研究表明了我们如何利用随机实验数据得出描述因果效应异质性的成员队列。
    - 识别这种异质性很有用，但并不能达到改善用户体验和实现更好的业务目标的最终目标。
    - 我们建立在队列识别的先前工作的基础上，并稍微扩展以处理多种干预和多种指标。
    - 我们用优化公式补充该部分，使我们能够为我们的最终目标服务。
    - 此外，由于我们希望我们的最终解决方案能够很好地处理不确定性，因此方差感知队列识别方法非常适合我们的整体方法论。
- 在整篇论文中，我们假设我们有能力对具有不同干预变量的成员进行随机实验，并捕获处理对成员相关操作（即业务指标）的影响，例如点击、评论、 视图、滑动等。
    - 理想情况下，对于所有个性化推荐（例如，在用户的供稿中向用户展示哪个项目或广告），我们应该使用因果数据。
    - 然而，为此所需的数据量是不可行的，因此我们依赖于基于相关性的观察数据和预测模型。
    - 然而，对于干预变量的个性化，使用因果数据更可行，因为选择的数量要少得多。
    - 队列级别而不是成员级别的选择进一步促进了可行性。
- 某些应用程序可能需要全局干预分配。
    - 例如，当一个平台测试其移动应用程序的背景颜色时，可能会有强烈的偏好融合到单一的新颜色并围绕该颜色形成品牌标识。
    - 围绕稳定性和一致用户体验的相关问题也可能需要在所有用户之间增加一个单一的全局参数或处理。
    - 在此类应用程序中，虽然可能存在异类干预的空间，但可能需要基于每个成员的时间一致性（例如，每年仅更改给定成员的字体大小一次）。
    - 获得任何所需时间一致性的一种方法是调节异构干预分配模块的运行频率。

    - 最后，在来自随机实验的数据量很小的应用程序中，可能没有足够的数据来估计任何异质性，因此分配全局最佳变量是谨慎的。
    - 在这种情况下，我们的方法也收敛到这个结论。
- 我们进行模拟分析，以评估和比较不同场景（尤其是不同噪声级别）中提出的每种方法。
    - 我们讨论了如何为应用程序选择特定的方法，并在 LinkedIn 的通知上的实际应用程序中展示所选方法的好处（使用先前陈述的指南）。
    - 请注意，我们的整体框架是通用的，适用于许多性质相似的问题。
    - 离线模拟和在线 A/B 测试都表明，我们的解决方案的性能明显优于启发式解决方案和全局分配。
    - 虽然最近对因果异质性估计进行了深入研究，但我们在这项工作中的建议是第一个（据我们所知）提供有原则的端到端解决方案，用于识别和利用这种异质性来改善用户体验， 交付更多商业价值并打造更具包容性的产品。
- 我们论文的主要贡献如下。
    - 我们通过估计异质因果效应和解决优化问题，开发了一个为成员选择最佳干预变量的通用框架。
    - 我们讨论了确定应该为给定应用选择建议技术中的哪一种的方法。
    - 我们进行了广泛的模拟，以展示使用我们的框架与使用全局固定参数相比的优势，并突出显示一种技术优于另一种技术的情况。
    - 我们描述了为大型社交网络平台将此类系统投入生产所需的基础设施。
    - 我们展示了一个实际应用程序的结果，该应用程序在指标方面取得了重大胜利。

- 上面的都不是人话，下面是我自己的理解
- 异质性：Heterogeneity
    - 一个变量X对另一个变量Y的影响可能因个体而异。
    - 例：多上一年学让张三的收入增加了1000元，让李四的收入增加了1200元，那么教育年限对收入的影响就存在异质性；
- 因果异质性：Causal Heterogeneity
    - 因为变量X的改变导致变量Y受到影响，并且这种影响可能因个体而异。
- 问题背景
    - 现在在各大互联网平台上都存在着许多AB实验，服务每个用户所使用的机器学习模型、推荐系统参数、甚至UI界面等等，都由这些AB实验中配置的干预变量所决定的，通常我们会从众多干预中选择出表现最好的，进行全量，这种方式成为全局分配。
    - AB测试是为Web或App界面或流程制作两个（A/B）或多个（A/B/n）版本，在同一时间维度，分别让组成成分相同（相似）的访客群组（目标人群）随机的访问这些版本，收集各群组的用户体验数据和业务数据，最后分析、评估出最好版本，正式采用。
    - 全局分配这种方式往往会是次优的，因为不同的干预对于每个用户或者说每个用户群体的效用，可能是不同的，全量某种干预方式很可能会给某个用户群体带来较差的用户体验，个性化干预的选择，可以带来更好的用户体验以及更多的商业增长。

- 通过更个性化干预的选择，可以将Pareto前沿向前推。此处需要说明，并不是所有的场景都需要个性化干预选择，比如说需要考虑用户体验的稳定性和一致性，或者是用户对于产品的心智教育，这样的场景上更加适合全局分配。
- 本文中，作者假设可以进行用户粒度的随机试验，尝试多种不同的干预，并观测到用户级别的指标变化，如点击、评论、浏览等等。
- 本文的主要贡献有：
  - 提出了一个通过估计异质处理效应并求解最优化问题，来做用户粒度最优干预的选择的框架。
  - 讨论了如何在我们所提出的技术方案中，根据使用场景来进行选择。
  - 介绍了一种新的合并tree的算法，来解决多treatment和多metric的问题，并用一种随机近似的方式，考虑因果效应估计的方差的同时，求解多目标优化问题。
  - 做了仿真实验，证明不同场景下，我们所提出框架的有效性。
  - 介绍了在实际产品中，我们系统的整体架构。
  - 在真实应用中取得了显著收益。