

阅读报告

一、 文献信息

题目： ML and precision public health: Saving mothers and babies from dying in rural India

作者： Kasey Morris 、 Vincent S. Huang 、 Mokshada Jain、 B.M. Ramesh 、 Hannah Kemp 、 James Blanchard 、 Shajy Isac 、 Bidyut Sarkar 、 Vikas Gothwal 、 Vasanthakumar Namasivayam 、 Sema Sgaier

二、 问题意义

1、 研究问题

本文研究的问题为如何利用机器学习去进一步提高全球公共卫生健康，利用机器学习来拯救印度偏远地区的孕妇和新生儿。

2、 研究背景

更高的全球健康目标和有限资源之间的矛盾让我们不得不采取更有针对性的措施。在印度，孕妇和新生儿的死亡率居高不下，印度政府采取了广泛的系统的干预计划，目的是增加相应医院设施，提高孕妇在医院分娩概率，这些计划最初很成功但是近年来效果停滞不前，大约 20%的孕妇依旧在家进行分娩，为了解决这一问题，必须采取一种更精确的干预，即在正确的时间正确的地点对正确的人进行干预，且由于数据收集和处理技术的进步，增强了我们得到更精确方案的能力，机器学习就为这种方法提供了一条有希望的途径。

三、 思路方法

首先使用预测模型来发现与医院分娩相关的因素，并利用 causal ML 来确定这些因素的相对重要性和潜在因果结构。然后使用监督式 ML 算法对母亲的类型和影响婴儿分娩地点以及原因进行细分。

1、 数据来源

数据收集自印度北方的 5968 名母亲，这些母亲最近生了孩子，数据来自 75 个地区和 600 个街区。该调查包括评估怀孕期间健康行为的项目（产前检查（ANC）检查和铁和叶酸（IFA）补充剂的使用）；来 ASHA 的访问频率；计划生育；对可获得的产妇保健服务和基础设施的意见；风险感知；财务；对政府财政激励措施的认识。所有分析的主要结局变量是母亲是在设施中（公共或私人）还是在家中分娩。使用回归模型对所有变量进行分析，发现有 16 个变量显着增加了医院分娩的可能性。这些变量用于 Casual ML 分析。

2、 Casual ML

使用贝叶斯网络进行因果机器学习，贝叶斯网络是概率图形模型，代表了一组变量下的条件依存关系。贝叶斯网络利用这些条件依赖性对因果关系进行建模。因子的潜在因果顺序通过结构输出（即，有向无环图[DAG]，它表示变量及其“边”或变量之间的有向路径）来识别。这表明哪些变量直接是相关结果的因果，哪些是通过上游途径的因果，哪些是因果链之外的。

使用 GNS Healthcare 的逆向工程和正向仿真（REFS TM）平台来生成因果贝叶斯网络。REFS 使用马尔可夫链蒙特卡罗（MCMC）方法来建立因

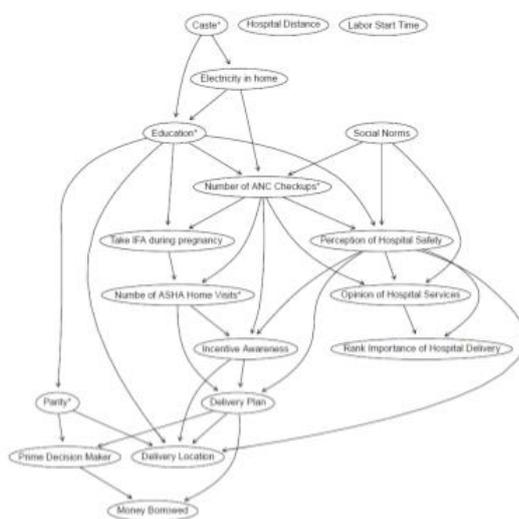
果模型的整体。

3、 监督学习进行细分

使用卡方自动交互检测分析 (CHAID) 决策树算法, 采用自上而下的修剪方法, 方法是依次修改停止条件 (每个节点的最小案例数, 最大树深度和 alpha 阈值) 以使其更加严格。10 倍交叉验证方法用于评估模型。总共检查了五棵树; 选择结构最简单, 预测和泛化误差最低的树作为最终模型。构造完最后一棵树之后, 将保存每个案例的细分标识符, 并根据设置的变量对细分进行配置, 以确定特征, 行为驱动因素和风险配置文件。

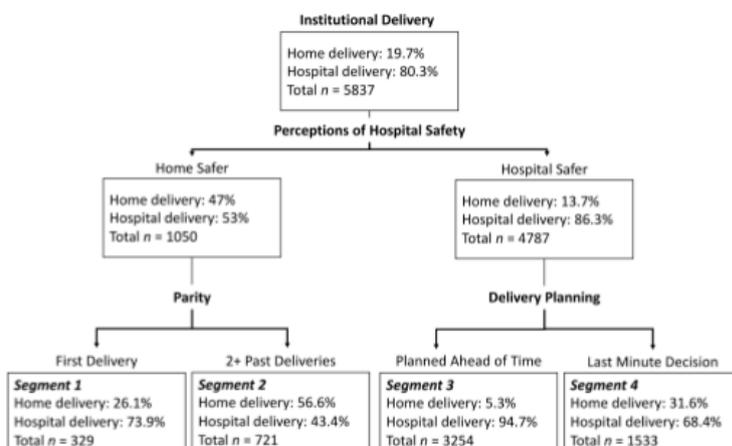
四、 实验结论

1、 贝叶斯网络得到的变量因果关系



ANC 检查的次数是一个重要的变量, 它对于网络中许多因果关系至关重要。此外, 对母亲的教育和将医院分娩作为社会规范似乎分别是重要的内在和外在原因, 这些因素调节了其他下游行为。尽管到最近医院的距离和分娩时间都与分娩地点有关, 但是它们不是因果关系, 同样家庭是否借钱以及对医院服务的一般性意见都不是分娩地点的因果关系。

2、 细分结果



细分方案基于三个因素将母亲类型分为四种。首先是对医院安全的看法，在相信医院更安全的母亲中，下一个分支根据分娩计划进行划分。在认为住房比医院安全的母亲中，下一个分支基于是否第一胎进行划分。

五、 启发思考

机器学习是当下最热门的研究方向之一，越来越多的领域开始应用机器学习去解决问题。此论文就是机器学习在社会服务上的一个应用，利用机器学习去解决实际社会问题。本论文中应用 PxPH 方法去解决印度偏远农村地区增加医院设施问题，从中可以得到两个有用经验。首先，这种方法需要收集数据并对数据进行优化，以涵盖各种驱动因素和行为障碍，机器学习方法通常遵循'garbage in garbage out'原则，如果没有优化的数据，洞察力和相关的干预建议可能会不足。其次，必须使用多种 ML 算法，此处使用的 ML 方法范围从广为人知和完善的算法（即监督学习）到新兴的方法（即因果 ML）。但是此方法有几个限制要注意。尽管我们可以根据自己的见识提供建议，但是如果没有评估指标，我们不能说这些建议是否会导致实际的有意义的变化。另外，从这些 ML 方法中得出的见解，根据定义，取决于馈入其中的数据。缺少完整或高质量的数据可能会导致不准确，不准确或有偏见的精确定位建议。为了减轻这些风险，需要研究人员，政策制定者和实施者之间认真合作。