计算扎根:定量研究的理论生产方法*

陈　茁　陈云松

提要:扎根理论的归纳逻辑和避免理论先入为主的原则与传统定量研究的演绎逻辑和假说检验大相径庭。在回顾传统定量研究理论生产局限的基础上,本文提出一种以定量方式直接助产理论的"计算扎根"方法:借助机器学习和归因算法,按照因果是可预测性的充分不必要条件之原理,根据对因变量的预测力筛选出以往研究未曾关注的自变量,以提出新的理论假说。本文对计算扎根的基本思路、逻辑前提、方法基础进行了系统阐述,并基于实际案例进行了演示。该方法弥补了定量研究理论生产的不足,在理论、学科、知识体系和社会治理等方面具有重要价值。

关键词:计算扎根　扎根理论　机器学习　归因算法　定量研究方法

一、导　言

依托于客观数据和模型的社会学定量研究对长期根植于逻辑思辨和历史情境的社会学传统研究方法形成了极为重要的补充,伴随着大型社会调查的开展和数据模型的普及,已成为社会学研究的重要范式。随着社会学学科共同体对定量研究科学性、规范性和因果解释力的追求,利用基于多来源数据的回归模型结果从样本中进行统计推断和因果识别,以此对提出的理论假说进行证伪检验,逐步成为定量社会学者普遍遵守的方法论范式。

定量研究原本只是对数字数据进行分析研究的统称(Scott & MarShall, 2009:538),但伴随着范式的固化,特别是定性和定量研究的二元分立,学术圈逐渐将量化研究自我限定在以演绎法为逻辑、以理论验证为目的、以统计推论为手段的单一面向。这种假设检验的方法无疑打破了无涉社会现象的纯粹哲学思

* 本文为国家社会科学基金重大项目"大数据驱动的网络社会心态发展规律与引导策略研究"(19ZDA149)阶段性成果。感谢匿名外审专家的重要修改建议。

辨,但它在和质性研究的长期角逐中,却似乎逐渐失去了理论生产和发展的主动权:质性研究者在不断地观察、发现和提炼新的概念与理论,并以此形成理论发展的"先导",而定量研究者则是对现有理论或基于文献和作者社会学想象力提出的假说进行"后置"的统计检验。定量研究者重视卡尔·波普对于科学的定义,也因此往往容易迷失在对证伪检验重要性的自我欣赏之中,不觉中忽视了数据和模型对于直接启发理论的价值、归纳逻辑对于定量研究的应用可能性。

有没有一种新的逻辑路径和模型,能让定量学者不仅能进行"后置"的科学检验,也能利用数据资料直接助产理论假说?事实上,使用量化资料直接助产理论假说的研究有非常悠久的历史,经典案例如涂尔干的自杀研究,统计上也有各种特征选择方法。但伴随着几十年量化方法的建制化,大家反倒忘了定量数据曾经也可以用来发展理论。究其缘由,对着变量列表进行随机的甚至遍历式 $N\times(N-1)$ 两两关联的头脑风暴,可能会产生大量了无实据甚至荒诞的假说;用传统的回归模型来进行各种 X-Y 组合的循环检验,无法解决控制变量的数量限制、选取组合和多重共线性等诸多问题。因此,借助数据和模型直接助产理论的研究路径在相当长时期内被定量学者所忽略。

但随着大规模社会调查数据的日益丰富和机器学习等方法在社会学中的应用,我们已经发现了这种可能。在本文里,我们将提出一种基于大量数据和机器学习模型的量化理论生产方法:对于既定的 Y 和大量的解释变量 X,通过监督学习方法对一系列 X 对于 Y 的预测能力进行量化分析。利用因果关系和可预测性之间的逻辑关联,我们可以对具有强大预测力的诸多 X 进行挖掘和筛选,从而直接助产理论假说,为 Y 寻找到潜在的具有理论价值的新 X,进而帮助社会学家生成、发展和修正理论。这一方法虽然是典型的计算社会科学方法,但其逻辑起点和扎根理论的核心原则有异曲同工之妙:打破理论的先入为主,在不做任何理论假说前提的条件下扎根于数据本身,从而打破"演绎-验证"的逻辑,打通经验研究到理论研究的生成路径。因此,我们将其命名为"计算扎根"(computing grounded theory)。

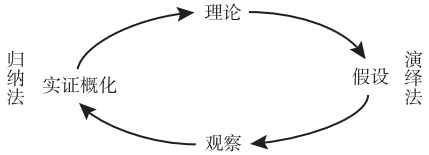
本文将首先对传统定量研究的假设检验路径进行简要剖析,然后详细介绍计算扎根方法的具体逻辑和思路。在此基础上,本文进一步从理论和方法层面分别论述计算扎根何以可能,并以"主观幸福感"为例进行案例演示,以检验计算扎根的信效度。最后,本文对计算扎根的方法意义和可能带来的潜在问题进行反思。

二、证伪的限度:传统定量研究的理论生产瓶颈

(一)假设检验与科学环

近四十年来,定量社会学研究的基本模式是基于调查问卷数据,采用多元模型回归的方法,对解释变量是否和被解释变量存在关联或因果进行统计推断。彭玉生曾形象地把定量研究比作"洋八股文"。他指出,国内外主流社会科学刊物,都采用了比较标准化的"模板式"格式,按照问题、文献、假设、测量、数据、方法、分析、结论八个部分各司其职,环环相扣(彭玉生,2010)。国外学者(Wells & Picou, 1981)通过对《美国社会学评论》的内容分析,也对假设检验模式作出类似的总结。有趣的是,定量研究的八股范式并非社会学独有,而是业已渗透进经济学、政治学、心理学等各个学科(林毅夫,1995)。尽管相关的环节可以合并或细化,但其基本思路就是对所提出的零假说进行证伪。

但检验理论并非科学研究的全部工作。华莱士在《社会学中的科学逻辑》中提出"科学环"概念,指出社会学研究是包括理论建构和理论检验的循环往复、螺旋上升、永无止境的过程(Wallace, 1971: 18)。很明显,假设检验的定量范式都集中在科学环的右半部分。从理论建构到理论检验本是一项科学研究的完整路径,但伴随着定量和定性的分野,理论建构似乎成了定性研究的专属使命,而定量研究日益将理论验证奉为圭臬。实际上,正如默顿所说,经验研究远远超出检验理论的被动功能,它不仅仅是证实或反驳假设,在塑造理论的发展上至少执行着四个功能:创立、修订、转变和澄清理论(默顿,2006:224)。



资料来源:Wallace, 1971: 18, Figure 1

图1 华莱士"科学环"的两种逻辑

(二)假设检验旨趣的历史渊源

假设检验范式起源于实证主义方法论传统,经过哥伦比亚学派对定量研究的规范建制化运动之后得到强化。拉扎斯菲尔德和斯托弗都主张用经验材料验

证理论的目的取向和科学化运动(Lazarsfeld et al.,1967;Stouffer, 1962)。斯托弗专门撰写了《检验思想的社会研究》一书,进一步使得用数据验证理论的方法在整个定量界得到充分普及。这一方法论传统针对传统理论话语包含大量形而上学的思辨和未经检验的论断等特点,将其视为不能提高有关社会事实的可靠判断的"空洞的陈述",也因此不属于精确的科学知识。在学术熏陶和方法训练之中,定量社会学研究者逐渐形成一种"实证型人格",他们要求自己不断地修正脑海中对于社会的构想,致力于提升社会科学对于实质性问题的回答效力(Pawson, 2000)。

假设检验的定量范式是社会学弥合理论与经验的鸿沟、确保结论科学性而形成的集体学科共识。但如果我们不加反思地将"理论先行 - 数据验证"的分析路径视为当然,那这种曾经作为知识解放力量的方法论就很容易转化为束缚,制约理论的生产创造力。事实上,使用量化资料探索理论并不始自今日,早在四十年前,一套从资料中自动筛选进行变量选择的分析的方法模式就已经具备。从统计学的角度看,有向前选择、向后选择、逐步回归等方法为模型挑选出最合适的变量,而后还出现了基于特征降维的偏最小平方、基于信息准则的 AIC 和 BIC 标准比较方法,基于正则化的岭回归、套索回归方法等。尽管这些方法或多或少遭受过批判,被指出变量筛选标准不科学、相关关系不等同于因果等问题(Rubin, 1974),但不得不承认,量化学者在实际的研究过程中都曾经得到过数据本身带来的启发,只不过很多研究者在从数据中得到新发现后并不会按照真实的研究过程来表述自己的研究,而是通过文献梳理的方式把自己的发现"装扮"成已有的理论假设,好像这些假设在分析数据之前就已经想出了,然后再按照假设检验的逻辑来证明它(Glaser,2008:15;吴肃然、李名荟,2020)。

(三)假设检验旨趣的后果

量化研究的"假设检验"会从两个方面对定量研究的知识生产形成束缚。

1. 导致定量探索性研究的缺位

长期以来学界逐渐形成一种实践者和旁观者的共同印象:经验研究是被用来校验理论的,理论则要通过研究者的奇思妙想来获得(Merton & Barber, 2011)。这尽管促成了不同研究范式的共同繁荣,但也导致了定量研究在科学发现之旅中的后置化甚至缺席:量化研究使得现有理论更为精致化了,但却很少产生新的理论建构(卡麦兹, 2009)。

2. 导致定量研究过度依赖常识而失去洞察力

定量研究所验证的假设基本来自现有理论的推导或社会学家的常识和灵

感。常识的矛盾之处在于它既能帮助我们理解世界,也会削弱我们的理解能力(Watts, 2011)。饶有趣味的是,一方面,社会学家需要带着与常识决裂的学科使命来怀疑并验证常识的科学性,但另一方面,在建立假设时又不得不在现有常识的窠臼中选取可能的解释变量,进而常常遭受"用复杂方法验证常识"的质疑(刘润泽、巩宜萱,2020)。

三、计算扎根:用机器学习助产理论

计算扎根的思路是打通从数据到理论的"逆向"路径,借助机器学习的预测能力和可解释的归因算法,基于因果是可预测性的充分不必要条件这一规律,实现用数据来直接生成关于既定因变量的机制理论。本章将分别对计算扎根的基本思路、逻辑前提和方法基础等进行详细讨论。

(一)计算扎根的基本思路

如图2所示,计算扎根的基本步骤可以由以下六个环节组成。

第一步,制定研究问题。根据社会调查问卷数据指标,结合研究兴趣和需要来确定研究对象 Y。理论上我们也可以不事先确定 Y,这样每个非先赋性的变量都可以成为我们预测的对象 Y,进而用遍历探索的方式来进行。

第二步,准备高维数据。社会调查数据往往是高维的,变量有上百个甚至更多。这些大量的指标,每一个都可能是潜在的 Y 的因,也即蕴含了扎根结果的可能性。不同层次的数据可以匹配起来,甚至可以纳入看不出任何与 Y 有关联的特征。

第三步,开展社会预测。基于高维数据,使用监督学习的方法训练 Y 的预测模型。算法可以是多样的,如支持向量机、随机森林、梯度提升树、神经网络,等等。只要能达到相对较好的预测效果,不必拘泥于算法是否复杂以及是否可解释。

第四步,比较预测能力。依赖机器学习模型的可解释性算法,对预测生成的黑盒模型进行归因分析,根据 X 对 Y 的预测力排序寻找可能的因。其基本思路是:打乱某些特征 X 是否影响模型预测的准确率,改变特征将如何影响预测结果。

第五步,寻找潜在理论。根据一组按照预测力排序的 X,寻找以往研究未曾涉及的社会关联。可以依据潜在关系模式将它们与既有研究比照,验证或澄清理论;亦可以对相似的解释项进行归类,抽象出概念或归纳理论命题。

第六步,补充交叉验证。验证计算扎根结果的稳健性和理论假说的适用

性。尝试使用不同数据、其他机器学习和归因算法对同一个因变量进行计算扎根,也可以对生成理论推导出的其他假说进行再检验,相互验证完成科学环闭环。

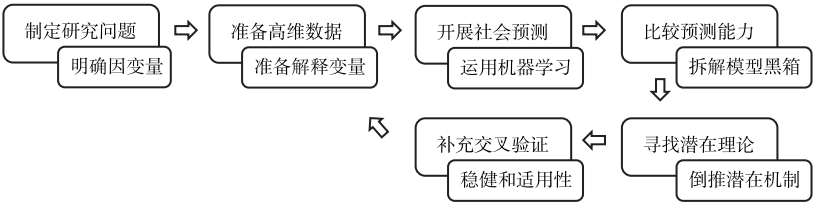


图 2 基于机器学习的计算扎根方法的总体路径

总体而言,该方法和假设检验式的定量研究大相径庭:它不预设理论假说,而是纯粹依赖算法和数据来训练预测模型,通过精确估算 X 对 Y 的预测力并比较排序,来为可能的因果关系提供一组备选的理论假说,保证了对数据的无差别开放态度和对实际问题的精确目标导向。根据霍夫曼等提出的社科研究方法四象限框架,计算扎根方法属于综合了预测和解释的综合建模研究(Hofman et al., 2021)。从周涛等人划分的五大社科方法类型来看(周涛等,2022),计算扎根的基本路径属于"基于(大)数据的探索性研究";如果在生成探索性理论假说后进一步使用其他数据验证,则又属于"先探索后验证的整合研究"。

事实上,机器学习辅助科学研究目前正在成为一股前沿的研究趋势,《自然》杂志曾以"AI-guided Intuition"为封面标题来预言人工智能将引导科学直觉,在数学(Davies et al., 2021)和管理学(Shrestha et al., 2021)[①]领域也都出现了使用类似的方法指导直觉和提出猜想的具体路径。同时,已有相关实证研究践行了机器学习和可解释性 AI 相结合的方法路径,如寻找导致焦虑的潜在因子(Nemesure et al., 2021)、挖掘提高乳腺癌生存率的潜在变量等(Moncada-Torres et al., 2021)。国内学者如罗家德等用相关方法对中国人脉圈层理论模型进行多轮修正和澄清(罗家德等,2021),周涛等也针对团队创新能力给出了类似的分析路径(周涛等,2022)。在这些相关实证探索的基础上,我们立足其运用于社会学研究的可能性,聚焦其通用的方法论价值和与传统研究模式的巨大差异,

① 该文章也提出了使用预测模型来建构理论的路径。不同的是,该文章主张使用不可解释的复杂算法和对模型复杂性加以约束的可解释性算法相互结合来平衡模型预测的准确性和可解释性。我们认为,预测和解释的平衡并不必然需要通过对算法复杂性的约束加以完成,可解释性 AI 算法本身已经可以对任何黑盒模型进行解释,且避免了任何不必要甚至可能具有误导性的变量筛选。

把算法模型的可解释性和理论生产在方法论层面进行了实质性关联,对这一思路进行系统化、标准化的提炼,正式提出完整的方法与实操路径。

(二)计算扎根的逻辑前提

作为理论生产方法的计算扎根理论,有着清晰的逻辑基础。对于计算扎根理论而言,其逻辑前提主要是两个方面。

一是扎根理论的归纳逻辑。扎根理论产生于实证研究对定量假设检验这一范式的反思。其创立者之一格拉泽指出,社会学过于强调理论验证,缩小理论与经验研究的努力主要集中在改进检验理论的方法,而试图从理论层面缩小这一差距的努力几乎没有什么成果;研究者应该从数据中发现理论,以打通经验研究到理论研究的生成路径(Glaser, 2008:7)。他主张以逐级归纳的方法从经验材料中直接创造出理论,再将其与现有理论和研究相比照。避免在研究开始前先入为主的观念或猜想是确保"扎根"有效的重要原则。

值得一提的是,创立者格拉泽强调扎根理论是一种普适的方法论,既适用于质性资料,也适用于定量数据;而定量数据和定性资料在目的与能力上没有根本冲突,对理论的产生和验证都是有帮助的(Glaser, 2008:17)。但随着扎根理论的实际发展,人们发现它似乎还是更适合做质性研究。扎根理论的另一提出者斯特劳斯甚至把它当作质性研究的"专属工具"(Strauss & Corbin, 1994)。其原因不难理解:质性资料的深度和可解读性往往更有利于运用社会学想象力直接提出理论假说,而定量数据作为一种数值指标具有高度简化的抽象特征,其内在的数理统计关联难以通过直观的方式加以发现。

实际上,为打破学界对扎根理论只适用于质性数据的刻板印象,格拉泽专门撰写了《做定量扎根理论》手册以详细阐述量化扎根的步骤。其基本思路是:在所有可能的两个变量间计算反映关系正负变化的粗指标;如果变量始终与一系列变量相关,将这些变量放在一起就会涌现理论的潜在模式;下一步是精化分析,即进行三个及以上的变量分析,通过发展类别的属性进一步饱和类别,实现更密集的理论(Glaser, 2008: 54)。但问题在于,大量变量难以通过人脑直接进行关联;使用统计方法时,对于哪些变量应纳入模型,实际上缺乏清晰的选取规则。特别是当自变量多到一定程度时,会出现自由度不够或共线性等诸多问题。总之,定量扎根理论逻辑可行,但当时尚无合适的方法来开展令人信服的应用。

二是因果关系的可预测逻辑。社会现象之间的可预测性和因果机制是两个不同但又高度关联的范畴。根据韦伯的定义,社会学是提供关于行为过程和结

果的因果性解释的科学(Weber, 1968: 4),可将社会学理论理解为指标之间的因果关系。按照这个逻辑进一步推演,社会学指标中的因变量对于自变量来说就一定具有可预测性。这是因为可预测性是因果关系成立的必要非充分条件,同时也是验证机制性原理的最基本手段(Watts, 2014)。

不过,由于受到数理统计工具的限制,社会学家们对于预测往往不太关心。在讨论到社会学中的因果、关联和预测等概念时,社会学家们多施以遁词:或强调预测不等于因果,但对因果必然可以预测的逻辑则束之高阁、不加利用;或者强调复杂的、纳入太多自变量的回归模型不够简约;或批评能进行数据预测的算法模型因其黑箱过程而无法解释,等等。针对这一类论点,邓肯·沃兹曾一一概括并加以严厉的批评(Watts, 2014)。

计算扎根方法的逻辑基础之一就是充分地运用预测和因果之间的重要关系,也即因果是预测的充分而非必要条件。这意味着,如果一个 X 可以很好地预测 Y,那么 X 的确可能构成 Y 的原因。虽然这一关系只是可能而非必然,但其构成因果的概率总比不具备预测性的关联要高得多。在社会学家逐渐把学科旨趣压缩到两两变量的分析而放弃社会预测的大背景下(Hofman et al., 2017),以机器学习的预测力来推动理论的生产对于定量研究具有重要意义。

(三)计算扎根的方法实现

计算扎根允许几十、上百甚至上千个变量的互动,其通过对算法模型的相关特征值的预测力进行比较,比通过个人思维灵感来发现理论的过程要稳定和可靠得多。其具体的实现过程包括社会预测和预测力比较两个方面。

1. 社会预测:运用监督学习的算法模型拟合

传统定量研究回归模型不擅长预测,而只擅长关联和因果推断。那么什么样的模型适用于复杂社会过程中的预测?著名统计学家布雷曼(Leo Breiman)曾将统计建模方法分为两种取向:一是数据模型,二是算法模型。数据模型事先假定数据服从某个函数分布 $f(x)$(如线性回归模型),然后对事先假定的 $f(x)$ 的参数进行拟合估计;而算法模型不假定数据的任何分布特征,旨在找到一个函数 $g(x)$,通过 $g(x)$ 可以对 y 进行预测(Breiman, 2001a)。实际上,这种分类恰恰切中了社会学的传统计量模型和机器学习之间的最本质差异。布雷曼进一步指出,当前社会和行为科学中广泛采用数据模型的思维方式,强调对模型参数的无偏估计而不是预测准确性。也就是说,社会科学中公认的实践模式,不是询问特定的数据和模型是否可以预测某些感兴趣的结果,而是询问理想化模型中的特

定系数是否具有统计显著性以及影响的方向。

但数据模型存在两个明显的问题:第一,为拟合特定参数模型,数据必须满足一定的假定。以线性回归为例,数据需要满足自变量和因变量关系是线性的、各自变量非多重共线性、残差服从正态分布、扰动项满足同方差、无自相关等多重假设。而现实社会复杂多样,要求数据满足严格假设未免过于苛刻,因此学术界采取了一种鸵鸟政策,逐渐将重要性转移至对显著性的强调,而对数据是否满足模型假定持开放或悬置态度(Freedman, 1991)。第二,结论是关于模型的机制而非关于事实的机制。将简单的参数模型强加于复杂系统生成的数据之上,会导致准确性和关键信息的损失。模型错误指定或研究者在数据分析中引入大量自由裁量权会导致潜在偏差(Simmons et al., 2011)。如果模型不能很好地模拟自然情况,则结论可能是错误的(Breiman, 2001a)。

以机器学习为代表的算法模型则为以上问题提供了一个非常好的替代方案。算法模型隐含的认识论假设是:事实数据的内在机制是未知和复杂的,关键是尽可能找到一个算法可以很好地通过 x 预测 y,即用算法拟合数据。算法模型往往采用非线性、非参数方法,通过一个或多个超参数来调整模型的复杂性。机器学习对数据复杂性的尊重使得被分析的数据可以服从任意分布,而不需要满足任何假设条件。我们认为,这种解放将至少从两个方面提高生产理论的能力。

第一,满足真实社会过程中的非线性数据关系。数据模型的线性假定往往难以符合社会真实情况。尽管模型的简约性可以作为理由,但简化只是手段而不是目的。大部分机器学习拟合过程不需要满足既有的函数设定,而是以追求预测准确性为最高宗旨(Breiman, 2001a)。

第二,满足真实社会过程中的高维复杂数据关系。传统计量模型只能纳入有限的解释变量。监督学习算法可以在单个学习模型中同时考虑数千个不同的因素和各种复杂的交互作用模式(Linthicum et al., 2019)。一个社会现象的影响因素纷繁复杂,纳入更多潜在的"因",发现新的解释维度的可能性也就更大。

2. 预测力比较:解决黑箱模型可解释性的归因算法

尽管机器学习打破了以往统计模型的种种预设限制,带来了数据生产力的解放,更好地模拟了事物的真实状态,但它最广为诟病的问题则在于其"黑箱过程"导致无法解释。不过,较新的机器学习文献中越来越多的证据表明,预测准确性和可解释性之间的矛盾并没有想象的那么严重。随着对复杂模型可解释性的迫切需求,越来越多"拆解黑箱"的方法得以发明且获得了成熟应用(Ribeiro et al., 2016)。哈佛大学教授的高被引论文指出,对机器学习的黑箱模型进行解释性分

析是一种基于数据驱动发现可解释因素的有效方法(Doshi-Velez & Kim, 2017)。

我们以沙普利值解释方法 SHAP(SHapley Additive exPlanations)为例详细介绍解释黑箱模型的具体路径。该方法根据联盟博弈理论来计算每个 X 的沙普利值,以此作为衡量其重要性的指标。考虑到不同参与者的数量和顺序都会影响最终的整体收益,该方法通过穷尽各种参与者的排列组合情况,对每种组合都计算包括该参与者和不包括该参与者的状态下整体收益的差值,记为该单个参与者的边际贡献;再对各种排列组合求该参与者边际贡献的均值,记为该参与者的沙普利值(Shapley, 1953)。所有参与者的沙普利值相加则为整体收益。

具体来说,每个参与者 i 的沙普利值的具体计算公式如下:

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}(v(S \cup \{i\}) - v(S))$$

其中,N 是所有参与者组成的集合,$|N|$ 是这个集合中包含参与者的数量;S 是一种参与者的组合方式,是 N 的一个子集;$S \subseteq N \setminus \{i\}$ 表示集合 N 删除特征 i 后的全部子集;$v(S \cup \{i\}) - v(S)$ 为包括参与者 i 的整体收益相对于不包括参与者 i 的平均边际贡献;该平均边际贡献在总的排列中出现 $|S|!(|N| - |S| - 1)!$ 次。为提高计算效率,数据科学家们开发了 SHAP 算法,通过条件期望函数来近似估计沙普利值,具体技术细节在此不再详述。沙普利值充分考虑了变量之间的交互作用,具有坚实的博弈理论基础,是同时满足有效性、对称性、虚拟性、可加性的公平分配方法(Lundberg & Lee, 2017)。

当然,除了沙普利值,对黑箱模型进行可解释性分析的方法还有很多,如置换特征重要性,即通过比较置换某列特征前后模型预测误差的变化来衡量该特征的重要程度(Breiman, 2001b);再如部分依赖图,即通过对一个特征反复修改,建构出本不存在的事实状态并再次预测,比较修改前和修改后因变量预测结果的差异(Zhao & Trevor, 2021);抑或使用可解释的代理模型来模拟原始的黑箱模型(Ribeiro et al., 2016)。这些方法的创立和成熟为重新平衡预测的准确性与可理解性提供了可能,为计算扎根奠定了坚实的方法基础。

四、计算扎根的实操和标准:理论生产示例

(一)研究问题与数据

我们以“主观幸福感”为例展示计算扎根如何助力于幸福感理论的启发和

澄清。本案例使用的数据为中国综合社会调查(CGSS)2017 年数据,共包含样本 12582 个、变量 783 个,为幸福感的计算扎根分析提供了相对全面可靠的数据依据。本研究的被预测变量为“总的来说,您觉得生活是否幸福——非常不幸福、比较不幸福、说不上幸福不幸福、比较幸福、非常幸福”,预测变量为问卷中除被预测变量以外的其他所有变量。①

(二)研究方法与步骤

第一步,数据预处理。首先,二分类的 Y 有助于提高算法预测的准确度,我们将“非常不幸福、比较不幸福、说不上幸福不幸福”记为 0,即非幸福样本;将“比较幸福和非常幸福”的填答者记为 1,即幸福样本。其次,我们将类别变量转化为虚拟变量。再次,我们删去缺失值大于 30% 的变量。最后,由于 1 类样本的明显数量高于 0 类样本的数量,数据不平衡可能带来算法偏差,我们采用 Bootstrap 采样对少数样本进行过采样补全,保证两类别的重新平衡。

第二步,模型训练。使用梯度提升算法 XGboost 训练预测模型,参数为 1000 个子决策树和其他默认参数。经 70% 的训练集迭代收敛后,余下 30% 的测试集显示,模型准确率为 0.92,召回率为 0.86,$F1$ 分数为 0.92,模型整体表现良好。

第三步,模型归因。主要采用沙普利值的 SHAP 模型全局可解释性方法进行可解释性分析,揭示影响预测的最重要因素和影响的方式。具体来说,针对每一个个案的每一个自变量 X,我们都为其计算了一个沙普利值(SHAP value)。该指标的意义是:对于该个案,加入这个 X 会比没有加入时对预测结果带来多少平均边际贡献。该值为正,则意味着该 X 的加入会导致幸福感的增加,该值为负,意味着该 X 的加入会导致幸福感的减少。

(三)研究主要发现

图 3a 展示了归因算法提取的对预测幸福贡献最大的前 20 个变量,衡量指标为所有个案在各个 X 上的沙普利值绝对值的平均值,也即变量 X 的平均边际贡献。图 3b 通过散点图的形式展示了不同预测变量影响的具体细节。图中的每一个点代表一个真实的样本;对于每一行来说,颜色代表该行变量特征值 X 的大小,X 越大,点的颜色越黑;②横轴为 SHAP 值大小;相同 SHAP 值的点越多,

① 幸福感评分和抑郁次数被剔除,它们几乎是因变量自评幸福等级的替代问题。

② 为便于读者阅读,变量特征值的大小均默认为从小到大排列,如公平感特征值为 1~5,1 代表非常不公平,5 代表非常公平。

“蜂窝”的截面积就越大,看起来就会越粗。总的来说,该图能反映出变量间作用的方式和大小,也能反映个案的分布状况。以公平感为例,散点图显示,公平感越高的个案(黑色点)往往集中在横轴的右侧,即 SHAP 值为正,幸福感增加;公平感较低的个案(灰色点)往往集中在横轴左侧,即 SHAP 值为负,幸福感减少。这意味着公平感对幸福的影响方式为典型的正向影响。

图 3 的分析结果显示,问卷的所有变量中,对幸福感预测最大的特征是公平感,SHAP 值约为 1.4;其次是健康程度。为便于归纳,我们按照变量含义的相似性将幸福感最主要的影响维度归纳为五个方面:(1)主观认知:公平感、信任感;(2)主客观地位:自我阶层定位、10 年后自我阶层预期、当地家庭阶层定位、自评社会经济地位、住房面积;(3)人口学和健康因素:健康程度、出生年、健康影响;(4)婚姻家庭:配偶同住、配偶工作小时、配偶年收入、夫妻应分担家务;(5)生活方式:休息放松、看电视、听音乐、每周工作时间。理论上,我们可以针对所有范畴进行层层归纳,抽象出更高层次的概念和关于幸福感的总体理论模型;也可以针对以往研究未曾关注的某一变量或某一具体维度作更深层次的挖掘和比较,探寻共同因素和共变规律,归纳出微观层面的理论假说。考虑到以上变量和维度涉及多学科领域,在以往的理论和实证研究中都或多或少的被关注讨论(刘军强等,2012;丘海雄、李敢,2012;Diener et al., 2018),本着预测力优先的原则,我们仅选择一个以往研究未曾关注过的,且预测力排在前十的变量“配偶每周工作小时”进行展示。

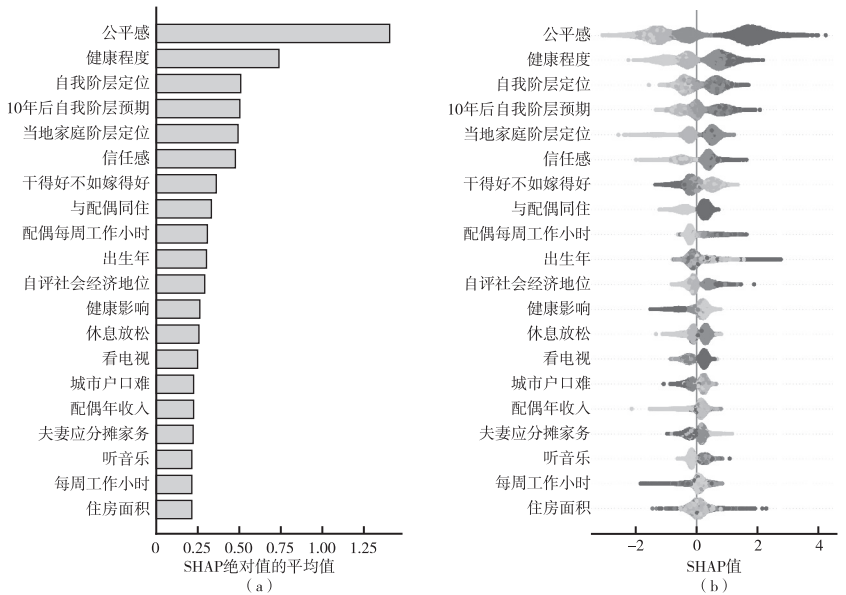


图 3　对幸福感预测平均边际贡献(SHAP 值)最大的前 20 个因素

1. 幸福感的新发现:寻找预测力强的新变量

从扎根结果生成理论假说的详细过程可以由以下几个步骤组成:(1)生成潜在假说的经验命题;(2)虚假相关的排除和因果关系的确立;(3)相关范畴的归纳与概念提炼;(4)与既有理论的对话和逻辑推导;(5)总结理论命题并使用其他数据方法进行再验证。具体来说,第一步是直接通过变量的预测力排序,发现关于变量间相关关系的事实命题。但命题还没有在现象和本质之间建立起一种基于因果的理性认识,我们可以进一步借助步骤 2 ~ 5 来相互补充,以填补命题到理论的鸿沟并增加理论的科学性。①

我们首先提出经验命题。根据图 3(a),“配偶每周工作小时”这一变量排在预测的第 9 位,但既往研究却没有充分关注。我们将两者的关系表述为一个新的经验命题:配偶工作时间会影响另一半的主观幸福感。

第二步,我们使用双向机器学习(Chernozhukov et al., 2018),尽可能排除可能存在的其他混淆变量,从而净化出两者的真实关系。以问卷中涉及的全部其他变量为混淆变量,Lasso,Random Forest,Decision Tree 和 XGboost 四种算法都显示两者之间存在显著的因果关系,限于篇幅,具体结果不再呈现。

第三步,为排除数据偶然性导致的预测力,应寻求和 X 内涵高度接近的其他变量,观察是否具有解释上的稳定性和逻辑性,进而创造出某个概念或某组概念来对数据间的关系模式归纳出一种统合理解。本例中,“配偶每周工作小时”指涉配偶在工作和家庭中的时间分配问题,我们筛选了 SHAP 值排在前列的,且都涉及夫妻双方在工作和家庭中的时间分配的其他类似变量:“每周工作小时(排名 19)”和“家人当面交流时间(排名 21)”。

我们通过进一步比较以上三个变量来启发理论直觉。图 4 通过全样本沙普利值的“宏观特征影响图”,展示了沙普利值在三个变量上的变化曲线。图中,灰色的点代表每一个样本,横轴代表这个样本的相关特征 X 的真实值,纵轴代表的这个样本对应的 X 的平均边际贡献也即沙普利值,黑色的线为该 X 在各个取值上的沙普利值均值的连线,连线的变化可以反映两变量间关系的变化。

可以看出,不管是配偶还是自身的每周工作时间,0 ~ 40 小时内的幸福感都随着工作时间的增加而增加。但超过 40 小时后,夫妻双方的工作小时形成了一种截然相反的张力:配偶更长的工作时间能明显提高幸福感,但自身工作时间的增多则明显降低幸福感。这意味着,配偶更多地承担社会角色并减少待在家中

① 这些步骤并不是必须的,而是为了相互补充以增加理论假说的科学性。

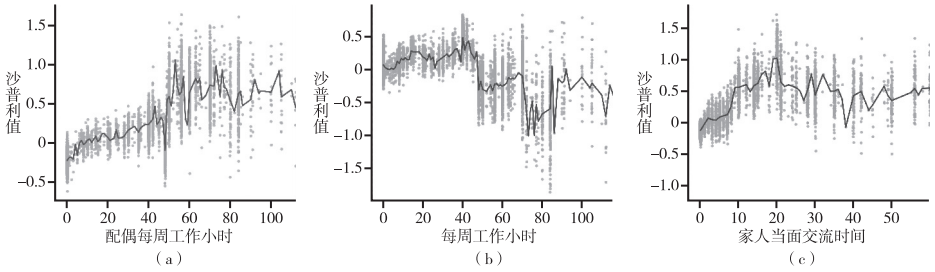


图 4 “家庭距离”新理论相关变量的沙普利值宏观特征影响图

的时间将有利于另一半幸福感的提高,我们将这种现象抽象化为“家庭距离”概念。同时,过远的家庭距离,也即配偶每周工作时间超过 60 小时将会导致幸福感的提升效应减弱。与之类似,“家人当面交流时间”也呈现峰状分布,即最好的家人当面交流时间在每周 20 小时左右,低于或高于 20 小时对幸福感的提升效果都相对减弱。通过以上变量的比较我们发现,提高幸福感需要配偶间保持一定的家庭距离,但这个距离不宜过低也不宜过高。这样,我们通过变量的比较和概念抽象构造出家庭距离影响幸福感的核心理论假说。

第四步,我们需要与既有理论展开对话,逻辑推导出变量间因果关系的详细机制,包括影响路径(中介效应)和不同群体影响的异质性(调节效应),形成更丰富的一系列逻辑递进的假说命题。[①]心理学的研究认为,时间的分配、个体的独立自主性和与他人的联系是影响个体幸福感的重要因素(Becker, 1965; Reis et al., 2000)。而家庭距离假说则主张夫妻双方在时间分配上保持独立和与他人联系的平衡状态。低家庭距离意味着家庭成员共处的时间增多,直接挤压另一半的独立自主性并增加夫妻双方产生矛盾的风险窗口。高家庭距离则导致与家人的疏离。同时,我们考虑还有其他压力伴随着家庭距离产生。

既有研究显示,出于对工作的重视和对个人责任的强调,社会上对无业者存在严重的污名化现象。失业的人常被视为懒惰、无用或不可靠的人(Brand, 2015)。工作除了赚取收入的明显后果外,还具有相当重要的“非金钱成本”,包括为一天提供时间结构、定义个人的地位和身份等(Jahoda, 1981)。这意味着工作时间较短的人或多或少也正在遭受家庭内部的污名化以及非金钱成本的损失。一方面,有着低工作时间的配偶可能会被定义为懒惰和失败,与其共同生活

① 我们还对男性和女性群体分别进行计算扎根,为丰富理论提供更多证据,限于篇幅,在文中不再具体展示。研究者可以根据性别、城乡、职业等不同群体做更精细化的分析,以进一步启发和丰富理论假说。

的另一半会产生强烈的被剥夺感;而工作时间较长的配偶则会被定义为勤劳、可靠或成功,与其共同生活的另一半进而会产生相对满足感。

但另一方面,过远的家庭距离也不利于提高幸福感。配偶过度工作则意味着与另一半联系的减少,而由于夫妻沟通是平衡冲突和婚姻满意度的重要中介变量(Carroll et al., 2013),家庭距离过远往往会导致情感疏离和矛盾的积累。同时,家庭距离过远会使得另一半不得不承担过多的家庭责任,导致家庭义务分配的转嫁与失衡(Bianchi et al., 2000)。

限于篇幅,本部分仅作为数据扎根步骤的案例演示之用,不再对产生的理论假说使用其他数据进行验证。[①]基于以上的计算扎根分析结果和推导步骤,归纳出"家庭距离理论",也即夫妻家庭距离过远或过近都不利于提高幸福感。我们进一步将其表达为逻辑递进的假说系列。

(1)配偶的家庭距离会影响个体的幸福感,但这种关系是非线性的。

(2)配偶过近的家庭距离会压缩自身自主时间,并增加共处矛盾;适当的家庭距离会增加自身自主时间,并减少共处矛盾。但过远的家庭距离会减少家庭成员的交流机会,带来情感疏离和矛盾累积。

(3)配偶的家庭距离会通过影响夫妻认同而影响幸福感。配偶过近的家庭距离会导致对配偶认同感的下降,与其共同生活会产生被剥夺感;适当或稍远的家庭距离会导致对配偶认同感的上升,与其共同生活会产生相对满足感。

(4)家庭距离影响家庭权利义务的分配,过远的家庭距离会使另一半被迫承担更多的家庭责任,由于家庭义务分配失衡造成双方矛盾积累,进而降低幸福感。

2. 理论的精细化发展:探究复杂关系的多元模式

前文发现"配偶每周工作小时"的沙普利值曲线呈现非线性模式,那么,这种 X 和 Y 之间的复杂关系还有哪些常见模式?我们另外选取了一些变量并绘制沙普利值宏观特征图(图5)。不难发现,我们能够从中找到大量传统回归分析模型所无法或者无力发现的细节,而这些细节对于进一步拓展、补充、验证和澄清理论非常重要。围绕 SHAP 值随 X 取值的变化,我们可以获得如下五种复

① 我们主张研究者在获得理论启发后再使用多重方式对理论假说进行验证,多种算法模型的交叉验证、算法模型和数据模型的交叉验证、使用其他数据的交叉验证、基于逻辑推导的其他理论命题的交叉验证等都能作为验证理论的方法,研究者应根据实际需要选择合适的验证方法。同时,我们不主张算法模型的结果一定要跟数据模型的结果完全对应,正如前文对数据模型和算法模型比较时所论述的,由于它们在变量数量、参数设置和数据关系模式假定等方面存在本质区别,两种模型的关系应该是互补而非互相竞争的。

杂关系的基本模式。

第一,“梯”状分布。X 对 Y 的影响在某个转折点后迅速变化,之后趋于平缓,如上一个阶梯一般。典型变量如“自我阶层定位”“10 年后自我阶层预期”和“10 年前自我阶层定位”(图 5a – 5c)。其中,“自我阶层定位”的关键性转折点是 4(图 5a),也即如果自我定位在 4 以上,则其对幸福感的影响是正向的,且阶层间相差不大(SHAP 值在 0.6 ~ 0.8)。而一旦定位低于 4 后,则迅速变成负向影响(– 0.2 左右),更低的阶层间(1 ~ 3)影响变化也不大(保持在 – 0.5 左右)。更有趣的是,这个转折点和人们对未来预期的阶层转折点(图 5b)不同:后者的转折点为 5。这个微妙的差异意味着:人们在当下生活中,只要认为处在社会中层(=5),就会觉得还不错,但人们对未来给予了更高的期望,未来处于第 5 阶层对幸福的平均边际贡献只有 0。

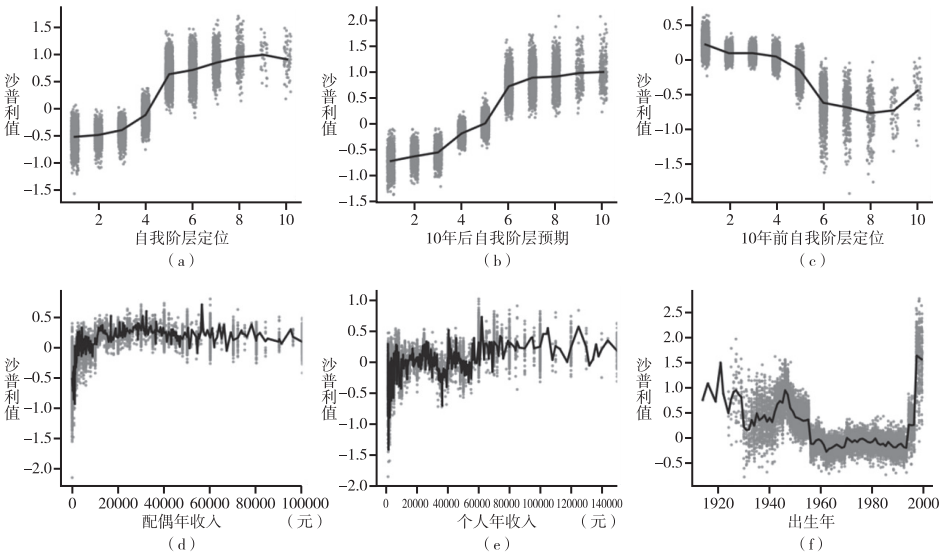


图 5　相关变量的沙普利值宏观特征影响

第二,“厂”字型分布。X 对 Y 的影响前期陡峭上升,后期趋于平缓,以“个人年收入”和“配偶年收入”为典型表现(图 5d – 5e)。这和幸福饱和理论所预期的一致:收入对幸福的正向影响服从平均边际贡献递减。这对社会治理政策具有重要的启发意义:扶贫应该把有限的资金投入到最为困难的群体中去。

第三,“谷”状分布。X 对 Y 的影响两端群体较高,而中间群体则比较低,形成谷状,以“出生年”为典型表现(图 5f)。出生在 1955 年前的老人和 1995 年后的年轻人的幸福感明显高于中间人群。另外,处于中间的人其年龄与幸福感没

有太大关系,沙普利值几乎呈水平分布。这些结果与近年来研究年龄与幸福感的突破性文献结果颇为符合(Blanchflower & Oswald, 2008)。

第四,“峰”状分布。X 对 Y 的影响中部群体较高,而两侧逐步降低形成峰状。如上文所述的“配偶每周工作小时”(图 4a)和“家人当面交流时间”(图 4c),这里不再赘述。

第五,“同质—异质”效应。同质效应表现为对同一类人群幸福感的影响一致,组内 SHAP 方差小;异质效应表现为对同一类人群幸福感的影响差异大,组内 SHAP 方差大。以“每周工作小时”为例(图 4b),工作时间为 0 ~ 40 小时的 SHAP 值均在 - 0.5 ~ 0.5,分布较均匀,工作时间对幸福感的影响较为同质;70 ~ 80小时的 SHAP 值则分布在 - 1.5 ~ 0.1,对幸福感的影响有较大的异质效应。这提示,工作时间较短的人基本上更幸福,但工作时间较长的人可能更幸福,也可能更不幸,可能存在其他重要交互变量一起影响幸福感。

3. 稳健性检验:罗生门效应的解决

数据和算法在计算扎根中被推到一个相当重要的位置。已有相当一部分学者注意到算法的罗生门效应,即因参数设定不同而带来的内部异质性和因算法不同而带来的外部异质性(Breiman, 2001a; 胡安宁等,2021)。计算扎根是否存在罗生门效应? 在多大程度上存在? 本部分从以下三个方面进行测试。

第一,数据的异质性。稳健的扎根结果不会随着数据量大小和构成变化而产生较大变化。我们通过 Bootstrap 自助抽样,分别在经过平衡后的原样本中随机抽取原始数据的 50%,60%,70%,80%,90%,100% 进行计算扎根。

第二,预测算法的异质性。稳健的扎根结果应该在不同预测算法下相似。我们比较了 XGboost、Catboost、LightGBM、Gradient Boosting 和 Random Forest 五种算法的计算扎根结果。

第三,算法参数的异质性。同一算法不同的内部参数也可能带来不同的分析结果。我们替换了 XGboost 算法的内部参数,包括最大树深度(max_depth)、正则化系数(alpha)、学习率(learning rate)、子样本比例(subsample),等等。

在每一种条件下,我们都得到了一个包括所有特征及其 SHAP 绝对值均值的表格。我们对不同条件模型计算的 SHAP 结果计算皮尔逊相关系数,具体结果如图 6 所示。总的来说,这些模型的训练结果高度相似,两两模型计算的相关系数基本都在 0.95 以上,相关系数的显著性都为 0.000。数据的异质性和算法参数的内部异质性基本不存在;预测算法存在一定程度的异质性,但最低也达到 0.88 以上。我们亦根据排序计算了斯皮尔曼等级相关系数,分析结果与皮尔逊

系数高度相似,故在此不再报告。综上所述,就幸福感这一案例来说,计算扎根方法具有相当大程度的稳健性。

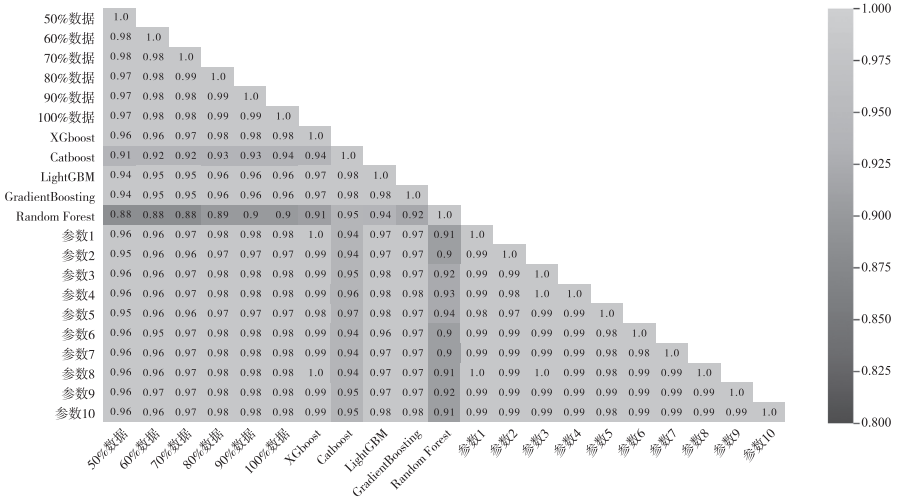


图 6　计算扎根方法的稳健性检验结果

(四)计算扎根的推荐技术标准

运用于社会科学领域的机器学习训练方法一直以来并没有较为统一的标准。为此,我们从 Web of Science 核心合集中,以“机器学习”为关键词筛选出 60 篇社科领域高被引论文,归纳出训练模型常用的变量数量、样本数量、模型选择、模型评估指标等信息,为算法模型训练给出经验参照标准。①

(1)样本数量。从文献统计结果看,样本数使用的中位数为 1888,2015 年之后的文献中位数为 11196。在保证样本可得性和代表性的基础上,我们建议探索性分析的样本数应大于 2000,探索加验证的分析样本应该更大。

(2)样本平衡。样本数应根据所要预测变量的类别数和难易程度调整。特别需要注意的是,罕见类别的样本数不能过少。对于非平衡样本,应通过重新采样的方法以平衡各类样本数量(Chen et al., 2022)。

(3)特征值数量。即用于训练的 X 变量的数量。从文献统计结果看,最多的使用了 1821 个变量,中位数为 22.5。更丰富的 X 变量会带来更好的训练结果,并更利于发现新的潜在理论;但也要考虑某些模型可能对数据噪声较为

① 限于篇幅,具体的筛选和统计结果不再详细展示,感兴趣的读者可联系作者获取。

敏感。

(4)算法模型选择。60 篇论文中,运用最多的算法是随机森林(29%),其次是支持向量机(26%)。神经网络类和梯度提升类算法约占 17% 和 15%。大部分论文都采用了不止一种算法并比较了模型的表现结果。因此我们建议,比较多种算法的预测性能和扎根结果,尽可能选择最优模型并进行稳健性检验。

(5)模型预测效果。统计的文献中模型准确度①的中位数为 0.79。计算扎根的有效性以模型预测的准确性为前提,考虑到既有论文的预测变量大多是二分变量,我们建议,二分变量的准确率应大于 0.8,连续变量准确率可以适当降低。

五、计算扎根方法的多重价值

计算扎根对传统定量研究的补充是多方面的。在数据层面,模型纳入的指标不再仅限于有限的几个变量,而是尽可能地纳入各类指标;在目标层面,不再强调模型系数的统计显著性,而是重新平衡社会预测的准确性和机制的可解释性;在观察视野层面,不再局限于回归系数大小和方向,而是细致挖掘变量间的非线性关系和群体异质效应。这些革新能够为计算扎根方法带来多方面的价值。

(一)理论创生价值:发现潜在模式

相比于传统数据模型,算法为导向的机器学习方法可以克服模型形式和变量选择的限制,并考虑变量间各种交互关系。对自变量数量瓶颈和关系限制的打破使得我们可以获得更完备的搜索、思考与检验解释变量的能力。这意味着只要数据本身足够丰富,“计算扎根”就能引导研究者通过发现新的解释变量启发新理论假说(陈云松等,2020)。通过一次“计算扎根”,我们就能够对整个调查数据的上百个指标进行筛选比较。

(二)理论发展价值:捕捉复杂关系

传统计量方法使用数据拟合模型,容易造成关键信息的丢失甚至错误(Varian,2014)。计算扎根方法通过超参数拟合数据,只要模型能尽可能地模拟

① 不同论文使用的模型评估指标有所不同,为保证一致性,这里采用模型准确率、精确率、召回率、AUC 和拟合优度中的最高值作为模型准确度的大致参照。

真实社会情境,就能充分捕捉变量之间的复杂关系,解放传统计量模型的线性枷锁,验证或者发展理论。前文的案例清晰地展示了其揭示和解释复杂关系的能力,更提醒我们,真实社会中数据的两两关系远没有我们预期得那么整齐划一:沙普利曲线几乎没有接近直线的分布。

(三)学科范式价值:第二种想象力

霍夫曼和邓肯沃兹等在2021年的《自然》杂志上发文呼吁在计算社会科学中整合解释与预测(Hofman et al., 2021)。他们指出,整合解释性和预测性思维的研究活动具有很大价值,但目前的研究屈指可数,该领域理应得到比迄今为止更多的关注。本方法正是整合社会科学解释性和预测性的全新尝试。对于定量研究范式而言,掌握计算机扎根方法不啻获得了米尔斯所提出的社会学想象力之外的补充。米尔斯的社会学想象力是基于个人体验的视角提升的思维(米尔斯,2017),而计算扎根则提供了一种基于数据的以算法模型来直接助产理论的思维能力。第二种社会学想象力蕴含了驱动理论新发现和放飞思维的磅礴力量。

(四)知识体系价值:自主知识生产

计算扎根方法天然具有一种更适合系统化知识生产的能力:有更多的新理论假说可以从数据中大量得到启发、更细微的机制和关系特征可以通过详实的预测力贡献分析被同时发现,以进行理论拓展和澄清。这对处于知识生产后发进程中的中国社会学来说尤其显得重要。要建立自主的知识体系,仅靠对具有先发优势的西方社会学的概念和理论进行异地验证是无法完成的。真正的自主知识体系需要一种足以对大规模社会、大时空跨度、高通量信息的中国数据进行复杂关系发现和理论提炼的工具。计算扎根无疑是这样的工具中最为重要的部件之一。

(五)社会治理价值:寻找干预因素

社会学是一门经世致用之学,社会公众和治理主体往往不会满足于概念提炼、过程解读和统计判断。这意味着定量社会学的学科使命不能仅局限于验证理论假说,还要掌握一种为社会现象找出关键干预因素的能力,才能真正为服务国之大者建言献策。计算扎根立足于社会预测的问题导向,通过不断模拟本身不存在的社会现象,对于为社会治理找出关键干预变量具有重要实用价值。

六、结 语

定量研究作为社会学领域的重要研究范式,深深根植于实证主义方法论传统,并形成了假设检验的单一路径依赖。我们强调,定量研究对于理论验证的过分强调很可能忽视了数据本身蕴含的巨大理论能量。基于此,本文提出了一种基于量化数据进行理论生产的方法:计算扎根。借助机器学习的预测能力和归因算法的可解释性,计算扎根恰恰能够在挖掘潜在关系模式、捕捉非线性关系等方面为定量研究的理论发展打开了一扇大门,打通从经验观察到理论生产的逆向路径。

回顾定量方法的发展历史,四十多年前,社会科学曾出现过基于数据资料和模型探索变量之间关系的学术风潮,但并没有形成成熟的研究范式。究其缘由,一是大量变量的纳入分析难以满足假设并导致多重共线性等问题;二是通过删除或添加单一指标筛选变量的方法仅仅是局部而非全局最优解,控制变量的变化会给结果带来较大扰动;三是预设的函数模式难以穷尽变量间复杂的关系作用模式。四十多年后,我们再一次呼吁定量研究补足其生产理论的缺角,吸取历史的教训,正视数据中蕴含的巨大理论能量。计算扎根方法的优势在于:第一,算法模型可打破模型预制的假设和关系模式,充分纳入大量变量并考虑变量间复杂的关系作用模式;第二,可解释性机器学习可凭借算法算力,在充分考虑变量各种排列组合的情况下得出全局最优解;第三,基于预测力的变量重要性排序比简单的变量相关性在分析逻辑上更能接近因果关系的范畴;第四,变量间各种非线性关系模式的挖掘和可视化呈现为引导直觉提供了更坚实细致的信息。

在为算法和数据可以直接助产理论而欢呼的同时,我们也提醒读者,本文并非否定传统的定量方法及其价值。任何一种方法都有前提、假设和局限,更有适用的特定的研究情境。它们都是定量社会学方法中的重要组成部分。我们强调,计算扎根不是对理论的拒斥,而是跳脱出已有理论和常识的限制,为提出新假说创造机会;计算扎根也并不排斥验证理论,而是同时强调将从数据中产生理论作为定量工作者检验理论之前的一个科学环节。

我们充分意识到,新的分析方法往往也会带来新的研究问题。计算扎根面临的挑战并不比它能带来的知识生产价值少。这些挑战包括:第一,数据维度的限制。就像遗漏变量永远不可能穷尽一样,尽管计算扎根尝试拓宽分析的数据维度,但这也无异于“戴着数据可得性的镣铐跳舞”。第二,社会预测的限制。

对社会复杂现象的可预测性一直有怀疑的声音(Taleb, 2010)。由于数据或模型的不足以及复杂社会系统固有的不可预测性,计算扎根方法并不适用于所有研究场景。第三,计算扎根的异质性。研究者知识生产的位置从研究的前端转移到后端,数据和模型被推到一个相当重要的位置,并可能导致潜在偏差。第四,相关性并非因果性。可预测并不等同于因果,对于因果关系和更深入的影响机制链条的挖掘仍需要进一步探索。

任何一种方法的成熟必然都要经历漫长的过程,要不断被实践和科学共同体所检验和修正。计算扎根方法未来需要探索和完善的有很多,如计算扎根方法的推荐标准和规范仍需进一步检验提升,计算扎根的适用场景和信效度尚需探索,计算扎根与统计推断和因果推断方法的对话有待推进……同时,本文提出的计算扎根方法主要基于结构化数据的分析。伴随着形式多样的大数据和人工智能的持续进步,计算扎根方法能否运用、如何运用于海量非结构化数据和更复杂的深度学习算法,也值得我们持续思考。作为混合了定性范式思维和逻辑的定量研究新范式,计算扎根需要学界更多的重视包容、推广实践与研究反思。我们呼吁在实证层面进行更多的检验和开拓,更为积极地把计算扎根这一方法在当前社会学研究中加以探索应用。只有当计算扎根方法能够实实在在地为当代社会学生成更多概念和理论,为中国社会学生成更多的自主知识,我们才会对计算扎根这一方法的力量和局限有更为深入的认识。

参考文献:

陈云松、吴晓刚、胡安宁、贺光烨、句国栋,2020,《社会预测:基于机器学习的研究新范式》,《社会学研究》第 3 期。

胡安宁、吴晓刚、陈云松,2021,《处理效应异质性分析——机器学习方法带来的机遇与挑战》,《社会学研究》第 1 期。

卡麦兹,凯西, 2009,《建构扎根理论——质性分析实践指南》,边国英译,重庆:重庆大学出版社。

林毅夫,1995,《本土化,规范化,国际化——庆祝〈经济研究〉创刊 40 周年》,《经济研究》第 10 期。

刘军强、熊谋林、苏阳,2012,《经济增长时期的国民幸福感——基于 CGSS 数据的追踪研究》,《中国社会科学》第 12 期。

刘润泽、巩宜萱,2020,《回顾与反思:定量研究在公共管理学科的滥用》,《公共管理学报》第 1 期。

罗家德、高馨、周涛等,2021,《大数据和结构化数据整合的方法论——以中国人脉圈研究为例》,《社会学研究》第 2 期。

米尔斯,赖特,2017,《社会学的想象力》,李康译,北京:北京师范大学出版社。

默顿,罗伯特·K. ,2006,《社会理论和社会结构》,唐少杰、齐心译,江苏:译林出版社。

彭玉生,2010,《"洋八股"与社会科学规范》,《社会学研究》第 2 期。

丘海雄、李敢,2012,《国外多元视野"幸福"观研析》,《社会学研究》第2期。
吴肃然、李名荟,2020,《扎根理论的历史与逻辑》,《社会学研究》第2期。
周涛、高馨、罗家德,2022,《社会计算驱动的社会科学研究方法》,《社会学研究》第5期。
Becker, G. S. 1965, "A Theory of the Allocation of Time." *The Economic Journal* 75(299).
Bianchi, S. M., M. A. Milkie, L. C. Sayer & J. P. Robinson 2000, "Is Anyone Doing the Housework? Trends in the Gender Division of Household Labor." *Social Forces* 79(1).
Blanchflower, D. G. & A. J. Oswald 2008, "Is Well-being U-shaped over the Life Cycle?" *Social Science & Medicine* 66(8).
Brand, J. E. 2015, "The Far-reaching Impact of Job Loss and Unemployment." *Annual Review of Sociology* 41(1).
Breiman, L. 2001a, "Statistical Modeling: The Two Cultures." *Statistical Science* 16(3).
—— 2001b, "Random Forests." *Machine Learning* 45(1).
Carroll, S. J., E. J. Hill, J. B. Yorgason, J. H. Larson & J. G. Sandberg 2013, "Couple Communication as a Mediator between Work-family Conflict and Marital Satisfaction." *Contemporary Family Therapy* 35(3).
Chen Y., G. He & G. Ju 2022, "The Hidden Sexual Minorities: Machine Learning Approaches to Estimate the Sexual Minority Orientation Among Beijing College Students." *Journal of Social Computing* 3(2).
Chernozhukov V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey & J. Robin 2018, "Double/debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21(1).
Davies, A., P. Veli?kovi?, L. Buesing, S. Blackwell, D. Zheng, N. Tomašev, R. Tanburn, P. Battaglia, C. Blundell, A. Juhász, M. Lackenby, G. Williamson, D. Hassabis & P. Kohli 2021, "Advancing Mathematics by Guiding Human Intuition with AI." *Nature* 600.
Diener, E., S. Oishi & L. Tay 2018, "Advances in Subjective Well-being Research." *Nature Human Behaviour* 2(4).
Doshi-Velez, F. & B. Kim 2017, "Towards a Rigorous Science of Interpretable Machine Learning." (https://doi.org/10.48550/arXiv.1702.08608).
Freedman, D. 1991, "Statistical Models and Shoe Leather." *Sociological Methodology* 21.
Glaser, B. G. 2008, *Doing Quantitative Grounded Theory*. Mill Valley: Sociology Press.
Hofman, J. M., A. Sharma & D. J. Watts 2017, "Prediction and Explanation in Social Systems." *Science* 355.
Hofman, J. M., D. J. Watts, S. Athey, F. Garip & T. Yarkoni 2021, "Integrating Explanation and Prediction in Computational Social Science." *Nature* 595.
Jahoda, M. 1981, "Work, Employment, and Unemployment: Values, Theories, and Approaches in Social Research." *American Psychologist* 36(2).
Lazarsfeld, P. F., W. H. Sewell & H. L. Wilensky 1967, *The Uses of Sociology*. New York: Basic Books.
Linthicum, K. P., K. M. Schafer & J. D. Ribeiro 2019, "Machine Learning in Suicide Science: Applications and Ethics." *Behavioral Sciences & the Law* 37(3).
Lundberg, S. M. & S. I. Lee 2017, "A Unified Approach to Interpreting Model Predictions." (https://doi.org/10.48550/arXiv.1705.07874).
Merton, R. K. & E. Barber 2011, *The Travels and Adventures of Serendipity*. Princeton: Princeton University Press.

Moncada-Torres, A., M. C. Maaren, M. P. Hendriks, S. Siesling & G. Geleijnse 2021, "Explainable Machine Learning Can Outperform Cox Regression Predictions and Provide Insights in Breast Cancer Survival." *Scientific Reports* 11(1).

Nemesure, M. D., M. V. Heinz, R. Huang & N. C. Jacobson 2021, "Predictive Modeling of Depression and Anxiety Using Electronic Health Records and a Novel Machine Learning Approach with Artificial Intelligence." *Scientific Reports* 11(1).

Pawson, R. 2000, "Middle-range Realism." *European Journal of Sociology* 41(2).

Reis, H. T., K. M. Sheldon, S. L. Gable, J. Roscoe & R. M. Ryan 2000, "Daily Well-being: The Role of Autonomy, Competence, and Relatedness." *Personality and Social Psychology Bulletin* 26(4).

Ribeiro, M. T., S. Singh & C. Guestrin 2016, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *Proceedings of the at 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, August 13 – 17.

Rubin, D. B. 1974, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of educational Psychology* 66(5).

Scott, J. & G. Marshall 2009, *A Dictionary of Sociology*. Oxford: Oxford University Press.

Shapley, L. S. 1953, "A Value for N-person Games." *Contributions to the Theory of Games* 2(28).

Shrestha, Y. R., V. F. He, P. Puranam & G. V. Krogh 2021, "Algorithm Supported Induction for Building Theory: How Can We Use Prediction Models to Theorize?" *Organization Science* 32(3).

Simons, J., L. Nelson & U. Simonsohn 2011, "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22.

Stouffer, S. A. 1962, *Social Research to Test Ideas*, New York: The Free Press of Glencoe.

Strauss, A. & J. Corbin 1994, "Grounded Theory Methodology: An Overview." *Handbook of Qualitative Research*. Thousand Oaks: Sage.

Taleb, N. N. 2010, *The Black Swan: The Impact of the Highly Improbable*. New York: Random House.

Varian, H. R. 2014, "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28(2).

Wallace, W. L. 1971, *The Logic of Science in Sociology*. Chicago: Transaction Publishers.

Watts, D. J. 2011, *Everything Is Obvious: Once You Know the Answer*. Australia: Currency.

—— 2014, "Common Sense and Sociological Explanations." *American Journal of Sociology* 120(2).

Weber, M. 1968, *Economy and Society*. New York: Bedminster Press.

Wells, R. H. & J. S. Picou 1981, *American Sociology: Theoretical and Methodological Structure*. Washington, DC: University Press of America.

Zhao, Q. & T. Hastie 2021, "Causal Interpretations of Black-box Models." *Journal of Business & Economic Statistics* 39(1).

Zhao, Q. & Trevor Hastie 2021, "Causal Interpretations of Black-box Models." *Journal of Business & Economic Statistics* 39(1).

作者单位:南京大学社会学院

责任编辑:刘保中