浅析国际政治大数据预测的限度

卢凌宇　张传基

内容提要：大数据的出现给国际政治预测带来了新的希望，但大数据并非无所不能。其成功预测的前提是事件具备起码的稳定性和连续性。作为一个过程，大数据预测大致包括数据准备、分析建模和模型应用与反馈三个阶段。数据准备主要由数据获取和数据预处理两个环节组成。对于数据获取而言，研究者既面临着出于保护个人隐私和国家安全的需要所施加的规范约束，又要努力克服数据资源的结构性缺陷所造成的现实约束。数据预处理则涉及通过数据挖掘技术从原始数据中提取特定数据的特征工程和旨在提高信噪比的数据降噪。在分析建模阶段，研究者设置的算法和模型会显著地影响到预测的效果。在模型应用和反馈阶段，研究者首先使用模型进行预测，然后根据预测结果来检验、评估和调试模型，其中事件背景条件和对象运行轨迹的变化是影响预测准确度的重要因素。从经验上看，上述诸条件满足得越多，预测准确率越高。本文按照大数据预测的工作流程，归纳并分析了国际政治事件预测实践中各个环节所面临的约束条件。文章创新之处在于较为深入地探讨了因果关系在大数据预测中的作用：它不仅是建模的基础，而且深刻地影响到预测的整个过程。

关键词：大数据　国际政治　预测　限度

一　导言

国际政治学是系统地研究外国政治和国际关系的社会科学，[①]包含描述、解释和预测（forecasting）三个功能。[②]国际政治学者对于预测在本学科的地位存在分歧，有的

① 若非特别说明，本文的“国际政治学”包括比较政治（外国政治）和国际关系两个政治学二级学科。

② “Forecasting”指对未来将要发生的事件或趋势的预告；“prediction”则是从理论演绎出的假设或含义，是对未知事件或趋势的预判，可能指向过去、未来或者现在。关于两者的区别，可参考 Keith Dowding, “Why Forecast? The Value of Forecasting to Political Science,” *PS: Political Science & Politics*, Vol.54, No.1, 2021, pp.104-106。

认为预测效果是国际政治学科学化水平的标志,有的则坚持解释才是学科最重要的功能。①尽管如此,预测对国际政治学者始终具有强大的吸引力。预测的准确度甚至可能变成了公众评价政治学科学价值的准绳。②不过,回顾国际政治的预测史,那是一连串失败的记录。而且越是重要和令人瞩目的事件,预测的成功率越低,③似乎预测失准才是常态。

"大数据"(big data)的出现为国际政治预测带来了新的希望。大数据简言之就是"数量特别庞大、来源特别广泛的数据"。④ 格雷(James Gray)认为,大数据的出现掀起了科学研究的第四个范式革命:"数据密集型科学发现"。传统的科学研究范式是实验归纳和模型推演,20 世纪中叶之后出现了仿真模拟(simulation),也就是"以数学方法、计算机技术、统计科学、信息科学和控制技术等为基础,运用计算机编程模拟的方式,在虚拟环境中模拟现实世界可能发生的现象、发展的状态,甚至是对未来变化趋势的预测"。⑤仿真模拟突破了社会科学研究对象无法实验或无法重复实验的限制,创造出了在现实世界中难以获得的操作和实施环境,是了解和掌握社会经济系统的结构和功能的一种有效的思考方法和实验工具。⑥大数据范式具有两个鲜明的特点:一是直接以真实世界为研究对象。模拟研究借助计算机编辑来模拟现实,但计算机模拟情境下的人类互动仍然是控制变量和给定规则下的互动,带有很强的"社会实验室"的性质,并非开放系统里的人类随机互动。而大数据则直接来自现实或取自现实,这是在模拟研究基础上实现了一个质的飞跃。⑦二是所谓的"数据驱动":只要有足够的数据,就可以用若干简单的模型取代一个复杂的模型,在程序上是先有数据,再建模型,然后用大量的简单模型去拟合数据。

大数据在社会科学研究中有着广泛的应用,主要分为三类:一是描述,最常见的是大数据的可视化展示,例如大城市交通控制中心实时路况监控;二是解释,比如通过收集人们的电子邮件数据及社交媒体文本数据研究人们的互动行为。大数据应用的核

① 毛莉:《理性看待国际关系预测》,载《中国社会科学报》,2017 年 9 月 22 日。

② 卢凌宇:《预测与国际关系科学》,载《欧洲研究》,2014 年第 3 期,第 143 页;Keith Dowding, "Why Forecast? The Value of Forecasting to Political Science," p.2。

③ 卢凌宇:《预测与国际关系科学》,第 143-145 页。

④ J. Craig Jenkins et al., "Political Behavior and Big Data," *International Journal of Sociology*, Vol.46, Issue 1, 2016, p.2. 关于大数据的定义,也可参考 Lilly Japec et al., "Big data in Survey Research: AAPOR Task Force Report," *Public Opinion Quarterly*, Vol.79, No.4, 2015, p.840。

⑤ 米加宁等:《第四研究范式:大数据驱动的社会科学研究转型》,载《学海》,2018 年第 2 期,第 14 页。

⑥ 董青岭:《机器学习与冲突预测——国际关系研究的一个跨学科视角》,载《世界经济与政治》,2017 年第 7 期,第 105 页。

⑦ 米加宁等:《第四研究范式:大数据驱动的社会科学研究转型》,第 11-15 页。

心是预测。①人类文明的进程其实可以简化为这样一个使用数据的标准流程:获取数据——分析数据——建立模型——预测未知。②预测未来或许是人类知识的终极目标。为此,我们既需要“望远镜”,又需要“显微镜”。“望远镜”让我们看得更远,发现新的“星系”;而“显微镜”则将更微小的世界展现在我们眼前。大数据就是数字时代的“望远镜”和“显微镜”,使我们看到并计量此前难于观察的现象。③

近年来,学者们开始广泛地使用大数据进行比较政治和国际关系预测,出现了一些颇为成功的案例。在联合国发起的“全球脉动”(Global Pulse)大数据计划中,有一个 2019 年的项目是预测索马里的难民和国内流离失所者(Internally Displaced People, IDP)。该项目汇总了有关内部和外部来源造成流离失所的潜在原因的数据,包括有关冲突事件和死亡人数、工资和商品价格、气候异常情况以及流离失所者流量的时间序列数据,最终集成为一个仪表板预警系统。该仪表板既可以显示历史数据,又可以预测索马里 18 个地区的 IDP 每月预计到达量,同时报告三个表现最佳的模型的预测结果。该系统下一步计划改善模型的预测性能,简化合并新数据和更新预测的过程,并在细粒度更高的子区域级别进行预测。④类似地,2012 年,斯维尔(Ned Silver)利用大数据成功地预测了美国全部 50+1 个州的选举结果,使常年做美国民调和选举预测的几个大公司震惊不已。斯维尔通过互联网尤其是各种社交网络尽可能地收集了所有和美国 2012 年大选有关的数据,其中包括各种新闻媒体上的数据、留言簿和地方新闻中的数据、脸书和推特上的相关发言和发言者社会关系的评论,以及候选人选举站的数据等等,然后依次对各州的情况进行挖掘分析,准确地预估了各州的选举结果。

尽管如此,大数据对国际政治的预测绝非无所不能。总体来看,大数据预测的效果即使不是远低于预期,也是差强人意。用诺斯科特(Robert Northcott)的话来说,预测的天堂依然遥不可及。⑤对于某些类型事件的预测而言,相比传统的小样本分析,大数据并没有体现出明显的优势。现今中国的人口普查基本是 10 年一次。在这十年间,为了及时掌握人口变化情况,还会进行几次传统的抽样调查,而各种社会大数据依然无法取代传统的小样本抽样。我们有理由追问:大数据预测不尽人意的绩效是偶然

① [英]维克托·迈尔-舍恩伯格、[英]肯尼思·库克耶:《大数据时代:生活、工作与思维的大变革》,盛杨燕、周涛译,浙江人民出版社 2013 年版,第 16 页。

② 吴军:《智能时代:大数据与智能革命重新定义未来》,中信出版社 2016 年版,第 12 页。

③ [美]史蒂夫·洛尔:《大数据主义》,胡小锐、朱胜超译,中信出版社 2015 年版,第 8 页。

④ Global Pulse, “Using Artificial Intelligence to Model Displacement in Somalia,” https://www.unglobalpulse.org/project/using-artificial-intelligence-to-model-displacement-in-somalia/.

⑤ Robert Northcott, “Big Data and Prediction: Four Case Studies,” *Studies in History and Philosophy of Science*, Vol.81, No.3, 2020, p.97.

的,还是系统性的?制约大数据预测效果的因素主要有哪些?换言之,大数据预测的限度是什么?本文旨在从国际政治学的角度,对上述问题做出尝试性回答。

二 政治学中的预测

国际政治是社会政治事件的一个子集。国际政治预测的现状在很大程度上反映了社会政治预测的总体水平。国际政治预测绩效差强人意是事实,有学者指出,这是由本学科研究对象的性质决定的。布莱斯(Mark Blyth)把政治事件按照其随机性从高到低分为三个类型:高斯分布事件、泊松分布事件和帕雷托-列维·曼德布洛特分布事件。其中高斯分布的特点是事件信息很容易获得,故而可预测性最强;泊松分布的事件有明显的随机性,可预测性次之,而帕雷托-列维·曼德布洛特分布的数据具有高度随机和易变的特点,几乎不可预测。由于政治事件绝大多数处于第二世界,呈泊松分布,其特点是观察不到事件的真正决定性变量,具有较高的不确定性,所以预测成功是概率性的。[①]布莱斯的研究是描述性的。他把三个世界的区别归结于数据获取的难易程度、数据分布、决定性因素的可观察性和现象的稳定性等因素,正是这些因素的组合决定了预测的效果。卢凌宇在借鉴布莱斯的类型学的基础上指出,国际政治可预测性依次取决于两个因素:一是事件的性质——事件属于"第一世界""第二世界"还是"第三世界";二是预测的方法,但他并未就特定预测方法的有效性高低展开论证,而是聚焦于论证绝大多数类型的国际政治事件是不可预测的。[②]

海因德曼(Rob Hyndman)和雅典娜梭普洛斯(George Athanasopoulos)的研究更进一步探讨了预测的约束条件。具体而言,无论事件性质如何,"可预测性"受到三个条件的约束:一是我们对产生结果的影响因素的理解程度;二是目前拥有多少数据;三是预测活动是否会改变预测对象的运行轨迹。王中原和唐世平在上述两位作者的基础上,增加了"预测手段和方法是否科学多元"作为第四个影响因素,[③]并借助这四个变量将政治事件分为"高度可预测""审慎可预测"和"高度不可预测"三个类型(见表1)。

---

① Mark Blyth, "Great Punctuations: Prediction, Randomness, and the Evolution of Comparative Political Science," *American Political Science Review*, Vol.100, No.4, 2006, pp.493-498.

② Ibid.

③ 王中原、唐世平:《政治科学预测方法研究——以选举预测为例》,载《政治学研究》,2020年第2期,第54页。

表 1 政治事件可预测的等级和限定条件

	理解程度 (变量关系和影响机制)	数据质量 (信噪比/丰富度/可及性)	预测是否改变 对象轨迹	预测手段和方法 是否科学、多元
高度可预测	充分	高	否	是
审慎可预测	部分	中	可控	发展中
高度不可预测	甚少	低	是	否

资料来源:王中原、唐世平:《政治科学预测方法研究——以选举预测为例》,第 54 页。

王中原和唐世平所列举的四种约束条件覆盖了影响预测效果的多数因素,包括数据质量、背景知识/理论、预测的反作用以及算法和模型设定。尽管如此,预测效果的影响因素并不是两位作者论证的重点,所以只用不到一个页面的篇幅匆匆带过。他们讨论的目的是为了提出政治事件的类型学,重点是分析选举这种“审慎可预测事件”的各类预测方法现状及前景。类似地,海因德曼和雅典娜梭普洛斯的论著是一部科学预测教科书,主体部分介绍主要的预测方法以及相关案例展示,预测的约束条件只是被当作理所当然的基本前提来对待,没有给予充分的讨论。

佩齐(Robert Pietsch)提出了实现数据成功预测的四个条件:一是了解所有在特定背景下影响预测对象的参数或变量;二是确保场景(context)稳定;三是确认参数(变量)处于稳定的因果关系中;四是案例足够多,覆盖现象所有可能的条件组合(configurations)。①他提出的第一、三个条件大致与王中原和唐世平的“理解程度”这一约束条件等同,第四个条件可视为与上述两位作者所强调的数据的“丰富性”大致相等。而佩齐所谓的“场景”可以被视为影响被预测事件的结构性因素,例如对于亚太地区的地缘政治而言,美国引发中美贸易冲突前后就呈现不同的场景。相对于王中原和唐世平,佩齐忽略了方法对预测效果的影响,但他提出的第二个条件——场景稳定性——则是王中原和唐世平未曾论及的。场景稳定意味着事件存在一定的连续性和稳定性。后文会论及,事件本身的不稳定会严重地影响大数据的预测效果。事件的连续性和稳定性决定了预测的上限。算法和模型的目的是尽可能地逼近这个上限。例如,以百分制为标准,如果实践中连续性和稳定性有 80 分,这就是我们预测的上限,算法和模型再怎么优秀,也只能使预测接近这个 80 分,而不可能超过它。缺了稳定这个前提,数据的数量及质量无论多高都难以实现准确预测。

① Wolfgang Pietsch, “Aspects of Theory—Ladenness in Data-Intensive Science,” *Philosophy of Science*, Vol.82, No.5, 2015, pp.910-911.

以上的文献回顾表明,预测的效果首先取决于事件的连续性和稳定性,其次才受制于数据和方法,其中数据涉及数量和质量,方法则主要覆盖算法和模型。但无论是数据的收集和识别,还是方法的选择,都取决于背景知识和理论。正是我们对事件的理解程度决定了哪些信息是预测需要的数据,以及需要使用何种预测方法。它们共同构成了预测的约束条件,决定了大数据预测的限度。

三 大数据预测:特点和基本前提

大数据有两项内涵:一是数据体量巨大。莱尼(Doug Laney)指出,大数据的特点是三"V":数量(volume)、速度(velocity)和类别(variety)。[①]这三个特点归根结底都是数量:速度快表明产生新数据的频率高,类别多的结果是在其他条件不变的前提下数据量会增加。二是有专门的数据分析方法。例如,构建回归分析模型时,为了求解模型参数,在小样本时代最常用的是最小二乘法(Ordinary Least Squares,OLS)。但在大数据分析中,OLS 会大大增加计算的复杂性,并且时间成本非常高,所以梯度下降法(Gradient Descent)才得到广泛的应用。[②]数据的体量就是统计学上的样本规模(N)或观察数。某些研究对象是有可能收集到所有样本的。例如,人脸识别系统基本能够收集到全体国民的人脸图像。所以,有一种观点认为,当样本规模(sample)等于研究对象(population, universe)时,描述就变成了预测,预测就一定会成功。

然而,情况远非如此乐观。调查数据增多确实改善了预测效果,但一来分析方法进步的作用不能被排除在外,二来对于某些类型的事件而言,大数据预测的效果并不显著优于传统的抽样预测。以选举问卷调查为例,理论上能够采集到选民的全样本,但一来数据采集的成本很高,二来即使忽略此成本,全样本也只是克服了由于样本过小所产生的抽样错误(sampling error),但它既不是导致预测失败唯一的原因,更不是最重要的原因。[③]根据统计学的"经验规则"(empirical rule):大约 68%的数据落在样本平均数的一个正负标准差以内;大约 95%的数据落在样本平均数的两个标准差以内;全部或者几乎所有的数据落在三个标准差以内。例如,澳大利亚有约 2200 万人

① Doug Laney, "3D Data Management: Controlling Data Volume, Velocity, and Variety," META Group Research Note, No.6, 2001, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

② 在梯度下降法中,梯度是一个向量,即为沿着变化率最大的那个方向(可形象地理解为下山最快的方向)行进,就需要选取学习速率,并通过多次迭代更新来求得最终结果。OLS 虽然不需要迭代,但其求解参数的公式中需要计算(X^TX)的逆矩阵,所以,当训练数据规模超过一定限度,求解过程的时间成本非常高。

③ Robert Northcott, "Big Data and Prediction: Four Case Studies," p.97, 99.

口。为了预测选举结果,绝大多数民意调查机构取样为 400 或 1000。原因在于,如果抽样人数为 400,那么大约 95%的数据落在正负 5%的置信区间;如果抽样人数为 1000,那么大约 95%的数据落在正负 3%的置信区间。①而且,从统计学上讲,要计算合适的样本量,仅有一个总样本数是不充分的,还需要至少一个指标来辅助。相比之下,样本平衡性是更重要的影响因素。比如,对于流感暴发而言,社交媒体使用者通常不具有代表性,在预测模型中过多地纳入此类数据会产生失衡的样本,导致出现偏差。②

类似地,即使政府掌握了全国居民衣食住行等各方面的全样本数据,也无法准确预测谁是全球主义者。这种深层次的居民属性,不是依赖描述就可以预测出来的,而是需要一定的分析方法才能挖掘出来。对数据中信息的挖掘常用的方法就是构建模型,而最复杂的模型本质上也是对现实的一种简化。大数据模型同样如此。任何模型都无法囊括现实世界的所有复杂因素,有些信息会不可避免地被遗漏。

布莱斯所谓无法预测的“第三世界”的特点是事件的发展缺乏起码的稳定性,呈现快速断裂和离散的状态。反过来说,无论是对自然界还是人类社会,大数据预测都基于一个默认的假定:过去、现在和未来具备连续性或规律性。唯有如此,我们才有可能通过归纳过去来预测未来。所以,如果国际政治事件满足不了这个前提,无论数据有多丰富,数据质量有多高,技术手段有多高级,预测结果都会低于预期。比如,生物世界中很多病毒的突变基本是不可预测的,而由此引起的流行病对人类社会威胁程度和影响程度也就相应地不可预测。

不过,问题的复杂性在于,事件的可预测性并不是一个要么是 0、要么是 1 的虚拟变量,而是一个从 0 到 1 的连续变量。换言之,可预测性或预测结果呈现为从 0 到 1 的连续的概率分布。很少有事件是绝对不可预测的。我们通常所谓的不可预测事件往往是指可预测性或概率很低的事件。地震就属于此类极难预测的事件。人类虽然开发出了地震预警系统,但该系统的功能并不是预测地震发生,而是预估地震发生后地震波到达目的地的时间,提前为人们预警。2010 年的北非阿拉伯国家的政治风波发生之后,学者们总结分析其产生的原因,对涉事各国国内政治冲突、国际经济的不景气以及西方势力的渗透和煽动等都给予了关注。然而,正是种种因素的叠加和混合致使如此大的政治剧变并没有被事先预测到。

而且,连续性本身也是发展和变化的,受到背景条件和外部冲击的影响。维克托

① Alan Agresti and Barbara Finlay, *Statistical Methods for the Social Sciences* (Third Edition), Prentice Hall, 1996, p.60.

② David Lazer et al., “The Parable of Google Flu: Traps in Big Data Analysis,” *Science*, Vol.343, pp.1203-1205.

(Jennifer Victor)指出,2020 年是多年来美国大选的一个转折点,选民政治态度的极化和新冠疫情以及特朗普这位大异常人的总统已经改变了美国的政治生态。①所以,非稳定性是成功预测所面临的根本挑战,也是大数据预测的真正瓶颈。不仅如此,在人类社会,连续性或规律不仅难以捕捉,而且会受到人的活动的干扰和影响,难以辨识。例如,权威机构公布对经济增长率和股市行情的预测往往会反作用于国民经济和股市,即出现"自我实现的预言"效果。无论预测是准确还是失准,我们都难以排除预测对预测对象的影响。

尽管如此,识别和利用事件的连续性或规律性是成功预测的前提。预测的整个过程以及所有要件,都与此项活动直接或间接相关。假定连续性或规律是客观存在的,它们的识别、挖掘和模拟就取决于数据和方法。如果基本符合预测的默认假设,同时我们对现实世界能够实现一定程度的量化,就有可能通过把这些量化的数据和相关专业的默会知识(tacit knowledge)整合起来,②建立模型,以便提高预测未来的准确度。

四 大数据预测的约束条件

预测是一个过程。如图 1 所示,预测始于数据收集,依次经历数据存储和管理、结合预测目标抽取数据、数据预处理——即将未加工的输入数据转换成适合分析的形式、机器学习和模型应用。这些环节构成了大数据技术的应用流程。③模型应用和数据收集之间的虚箭头表示模型应用在现实政治世界中自然会产生"新"数据,成为数据收集的一个来源。例如,对可能的冲突预测并采取一定措施后,现实世界的数据就会变化,成为新的数据。不同的预测技术路径肯定会在不同问题领域有不同的效果,差别可大可小。自然语言处理技术有一种分析方法叫作情感分析(Sentiment Analysis)。目前,它的实施主要借助两条技术路径:基于情感词典的无监督学习和基于分

① Jennifer Victor, "Let's Be Honest about Election Forecasting," *PS: Political Science & Politics*, Vol.54, No. 1, 2021, pp.1-3.

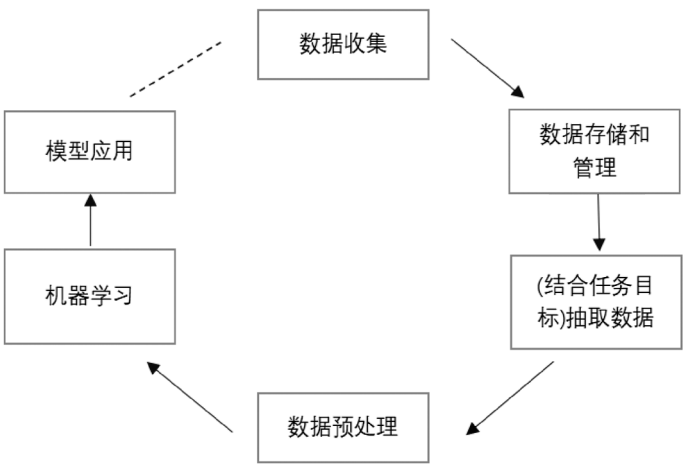
② "默会知识"又称"隐性知识",指我们知道但不能通过语言、文字、图表或符号明确表达的知识。波兰尼(Michael Polanyi)1958 年在《个人知识》中首先提出,参见 Michael Polanyi, *Personal Knowlege: Toward a Post-Critical Philosophy*, Routledge & Kegan Paul Ltd., 1962/1998, pp.80-85。

③ 实际上,不同的步骤/环节之间是有重合的。比如,大数据算法中包含着部分特征工程的工作。例如,神经网络算法在机器学习和认知科学领域,是一种模仿生物神经网络的结构和功能的数学模型或计算模型,用于对函数进行估计或近似估算。神经网络由大量的人工神经元联结进行计算。其中隐藏层的各神经元的一个主要工作就是在做特征提取,可能 A 神经元构建了变量的一个特征子空间(即一个高价值的特征),B 神经元构建了变量的另外一个特征子空间,C 神经元增加了变量的非线性等,但具体它们做了哪些变化及为什么这样做其实是不知道的,所以属于"黑箱"运算。

类算法的有监督学习。①在预测各国民众对经济全球化的态度或国内冲突的发生等问题上,都可以通过对网络文本数据的情感分析进行预测。采用不同的技术路径得出的预测结果肯定会有不同。情感词典方法处理法文文本数据的效果可能很好,但面对汉语文本数据,效果可能会大打折扣,因为汉语中一词多义的现象很常见。实际上,技术路径的选择不仅取决于问题的性质,还需要考虑很多非技术因素,比如操作的复杂性、时间成本和资金成本等。本文拟从大数据方法这一宏观的视角来做分析,所以假定在处理具体问题领域的对象时,我们已经选择了最优的技术路径。

依据各环节的主要功能,我们可以把上述六个环节总结为三个步骤:数据准备(数据收集到数据预处理这四个环节)、分析建模(机器学习)②、模型的应用和反馈——模型应用和数据收集。在大数据分析预测建模的实践流程中,每个环节都可能降低大数据的预测效果。接下来我们主要结合国际政治事件,对上述三个步骤中影响预测准确性的约束条件逐一进行分析。

图 1 大数据分析预测流程



注:表由作者自制。

① “情感词典”简言之就是把表达情感的单词汇总,并根据表达情感的程度进行赋值。基于情感词典指的是根据已构建的情感词典,对所分析的文本进行情感词提取,然后计算情感倾向,最终效果取决于情感词典的完善性。基于分类算法指的是通过对语料库的文本及对文本情感的标注,比如把“我们不支持特朗普总统”标注为0,即负面情感;反之为 1,是正面情感。然而,我们在标注后数据的基础上进行机器训练,得到分类模型,再进行预测,其中分类效果取决于训练文本的选择及情感标注的正确性。

② “机器学习”(machine learning)通俗地讲就是基于样本数据(称为“训练数据”)构建模型。关于机器学习在政治学中的应用,可参见 Justin Grimmer et al., “Machine Learning for Social Science: An Agnostic Approach,” *Annual Review of Political Science*, Vol.24, No.1, 2021, pp.395-419。

(一)数据准备

数据准备把原始数据加工为机器能够识别和运行的精练数据,严格来说包括数据收集、数据存储和管理、(结合任务目标)抽取数据以及数据预处理四个环节。鉴于中间两个环节偏向计算机技术处理,所以本文不做分析。我们将集中讨论数据获取和数据预处理对预测准确度的影响。

1.数据获取

数据获取的约束有两种:一种是规范约束;另一种是现实约束。所谓"规范约束",是指出于保护个人隐私以及国家安全的需要对数据获取施加的限制。数据获取的现实约束则是数据资源的结构性缺陷。

(1)规范约束

比较政治和国际关系预测经常会涉及与个人隐私尤其是国家安全相关的保密数据或信息。一方面,出于保护隐私或安全的需要,一些重要的政治活动都是秘密进行的,媒体并不知晓,或者即使了解也不被允许报道。利特鲁(Kalev Leetaru)创建的GDELT(Global Database of Events, Language, and Tone)数据库能够实时检测全球网络空间、门户网站、印刷媒体、电视广播、网络媒体、网络论坛中的新闻事件,对其进行分析提取,提炼出新闻事件相关的行为主体、人物、地点、组织和事件类型等关键信息。那些未被报道的政治活动无法被数据库收集到。早期有关朝鲜战争发生原因的研究有传统学派和修正学派之分。传统学派认为,苏联联合中国策划了这场战争,目的在于检验美国遏制国际共产主义运动扩张的决心。① 20 世纪 90 年代以来,随着苏联解体、大批当时保密的外交档案获得公布,研究者才发现,中国领导人只是原则上同意朝鲜采取军事手段。苏联和朝鲜的战争准备及其军事计划丝毫没向中方透露。②传统学派过高估计了中方的影响。由于档案保密的原因,研究者对当时朝鲜战争时期主要国家间关系产生了不少此类完全"错误"的认识,影响了后续的对相关国际政治事件的分析和预测。可见,不对称的重要信息会造成数据出现系统性的缺失,影响数据质量,降低预测的准确度。

另一方面,大数据对个人隐私和国家安全的威胁是现实存在的。棱镜计划(PRISM)是一项由美国国家安全局自 2007 年小布什执政时期起开始实施的对即时通信和既存网络资源进行绝密电子监听的计划。2013 年 6 月,美国前中情局雇员爱德

① 修正学派的观点是韩国侵略朝鲜或者诱使朝鲜入侵韩国。

② 沈志华:《冷战在亚洲:朝鲜战争与中国出兵朝鲜》,九州出版社 2013 年版,第 10-20 页。

华·斯诺登在英国《卫报》和美国《华盛顿邮报》曝光了该计划。计划中的监控信息包括电邮、即时消息、视频、照片、存储数据、语音聊天等十类信息,监听对象包括任何在美国以外地区使用或参与有关公司服务的客户,或是任何与国外人士通信的美国公民。美国政府通过监控获取了大量数据,然后对这些数据进行挖掘和分析,获得了有价值的情报信息。该计划严重威胁和侵犯了公民个人信息安全和被监控公民所在国家的国家安全,遭到广泛的国际谴责。

近年来不少学者提出“数据主权”的概念,认为在信息时代,大数据是实现国家主权的基础,“数据主权”将成为继边防、海防、空防之后另一个大国博弈的空间。①如果数据被广泛确认为国家主权,数据流动和跨界交流的难度会显著增加。实际上,大数据对国家安全的威胁经常被夸大,为数据获取增加了难度。2020 年 8 月 14 日,美国总统特朗普(Donald Trump)签署行政令要求字节跳动公司在 90 天之内出售或剥离该公司在美国的抖音(TikTok)业务,借口是抖音窃取美国用户的隐私,对美国构成“信息安全威胁”。字节跳动公司原计划结合美国用户的偏好和使用特点开发一些跨文化的互动短视频或音乐节目。但是,2020 年 3 月,特朗普政府宣称即将对抖音采取措施,字节跳动 4 月宣布计划将抖音剥离出母公司,成立独立运营的美国公司,基本断绝了数据跨界交流的可能性。

数据安全主要有三种类型:第一,数据不丢失、不损坏;第二,防止数据被盗;第三,数据的交流和使用安全。对于本文的分析相关性最强的是第三种安全。在大数据应用到国际政治分析中时,一个重要的障碍就是数据的跨界获取和交流以及数据使用的安全性。出于数据脱敏的需要,在使用有关个人信息进行分析时,有必要把身份证号这样敏感的个人隐私数据用无关的唯一标识符替代。为了达到数据加密的要求,要把国防机密发送给指定对象,就必须把文字加密成数字串。这两种操作是解决大数据应用所面临的数据安全问题的可行路径。②尽管如此,上述两种方法都存在难以克服的缺陷:由于不同应用场景的要求各不相同,对于数据脱敏到什么程度才算安全并不存在一个通行标准。数据加密后,除非知道解密的方式,否则我们难以确认数据是否涉及国家安全信息或违法犯罪信息。2020 年 10 月,麻省理工学院的信息与决策系统实验室(Laboratory for Information and Decision Systems)提出了一种新的解决方案,叫“合

① 蔡翠红:《国际关系中的大数据变革及其挑战》,载《世界经济与政治》,2014 年第 5 期,第 131 页。

② 董青岭:《反思国际关系研究中的大数据应用》,载《探索与争鸣》,2016 年第 7 期,第 93 页。“数据脱敏”指对某些敏感信息通过脱敏规则进行数据的变形,实现敏感隐私数据的可靠保护;数据加密是利用密码技术对信息进行变换,实现信息隐蔽,保护信息的安全。

成数据”(synthetic data)。合成数据与“低卡苏打水”模式相似。通俗地讲,低卡苏打水往往拥有和常规苏打水相似的外观、味道和泡沫,但两者在本质上是不同的。①

(2)现实约束

相比传统数据,大数据有两个显著的特征:一是完备性(completeness);二是多维性(multi-dimensionality)。完备性又称全面性,指随着样本数据量的增加,样本的代表性更充分。如果样本囊括了所有研究对象,抽样错误就会消失。对于某些类型的事件而言,预测不准的一个重要原因就是数据严重不足。大数据投票预测失准就是一个典型案例。由于过去的选举样本数量太有限,不足以训练机器学习。而且由于不计名投票意味着个人的投票结果是未知的,所以也找不到其他可替代的数据。②

多维性则指数据的维度或指标要尽可能多。多维性不是多源性,后者指数据的不同来源。多源的信息能够更加准确地反映分析目标的状况,便于交互印证,提高预测准确度。数据的多维性则取决于对相关参数的知识。例如,在美国的政治宣传中,大数据促成了微目标(microtargeting)宣传的兴起。微目标对应的概念是“宏观目标”和“中观目标”。宏观目标以选民整体作为政治传播的对象;中观目标以不同类型的选民群体(如黑人、中产阶级等)作为政治传播的对象;而微目标解决了选民群体内部的异质性给传播带来的困难,它把政治传播的对象聚焦到个人,能做到“把合适的消息传递给合适的人”。③ 2012 年,共和党总统候选人罗姆尼(Mitt Romney)竞选团队的数据负责人朗德(Alexander Land)曾表示,微目标与传统的目标群体最大的不同,就是由对着一个群体大叫,转变为与一个个选民亲密交谈。④

2016 年“全球脉动”实验室实施了一个项目,目的是“用邮政数据建立国民福祉的代理指标(Proxy Indicator)⑤”。这项研究使用了万国邮政联盟在四年内(2010-2014 年)收集的来自 187 个国家的汇总电子邮政记录,以创建一个国际网络来展示世界各

① Laboratory for Information & Decision Systems, “The Real Promise of Synthetic Data,” October 16, 2020, Synthetic Data Vault, https://news.mit.edu/2020/real- promise-synthetic-data-1016. 该方案的情况如下:2016 年,信息和决策系统实验室开发了一种算法,用以准确捕捉真实数据集中不同字段之间的相关性 —— 比如患者的年龄、血压和心率,并由此创建了一个合成数据集,该合成数据集留存住了这些相关性,但没有保留任何其他可识别信息。当数据科学家被要求使用这些合成数据来解决问题时,结果显示,他们的解决方案在 70%的时间内与使用真实数据生成的解决方案一样有效。之后,该团队继续进行开发。2019 年,Lei Xu 在于温哥华举行的第 33 届 NerulIPS 会议上介绍了新算法 CTGAN(conditional tabular generative adversarial networks),建立和完善综合数据表。其研究显示,CTGAN 在 85% 的案例中表现优于经典的合成数据创建技术。

② Robert Northcott, “Big Data and Prediction: Four Case Studies,” *Studies in History and Philosophy of Science*, pp.97-98.

③ 刘亚琼:《大数据时代美国政治宣传的特点及其启示》,载《新闻知识》,2020 年第 12 期,第 16 页。

④ Rasmus Nielsen, *Ground Wars: Personalized Communication in Political Campaigns*, Princeton University Press, 2012, p.144.

⑤ “代理指标”又称“替代性指标”,指当对某些现象无法直接度量时,间接度量时使用的指标。

地的邮政流向。它可以为 14 个社会经济指标建立代理指标,以新的方式模拟人类发展指数(HDI)和国内生产总值(GDP)等常规指标。此外,邮政数据与其他全球网络——贸易、移民、国际航班、互联网协议(IP 地址)和数字通信等——的数据相结合,会产生新颖的多维连通性指标。①

间接数据是数据存在的另一种形式,它也会显著地影响到预测的效果。根据指标反映变量的效度/真实度(validity)的水平,②数据可以分为两类:直接数据和间接数据。直接数据是直接反映真实指标的数据,如 GDP 和人口数量等。反之,无法直接反映真实指标的数据是间接数据。通常认为,以下两种情况需要收集间接数据:一是无法量化的指标。比如美国对中国的态度是友好、中立,还是敌对。二是由于保密等因素导致无法无效收集但在预测中又不得不使用的数据。例如,我们试图通过推特(Twitter)分析美国各州人民对联邦政府新冠疫情防疫政策的态度。在这个案例中,“地址”是最关键的变量之一。由于推特保护用户隐私,下载的数据中地址这个变量全部为空,所以只能结合推特用户个人主页填写的地址并比对其主要社会关系的主页填写的地址来对“地址”这个变量进行插补。间接数据胜过零数据,代价是目标指标的信息会遭受一定程度的损失,数据质量也会相应打一些折扣。所以,通过对候选人的谷歌搜索这一间接数据来预测选举结果的效果就远低于预期。③从逻辑上讲,公众在谷歌搜索某位候选人信息的结果可能是支持、反对或者中立,而搜索行为本身无法区分这三种结果。

2.数据预处理

数据预处理的目的是将未加工的输入数据转换成适合分析的形式。数据预处理涉及的步骤包括整合来自多个数据源的数据、清洗数据以消除噪声和重复的观测值、选择与当前数据分析任务相关的记录和特征。其中最重要的两个环节是降低噪声和特征工程(feature engineering),而特征工程借助统计学和工程学等专门领域的知识进行数据挖掘技术,以便从原始数据中提取特征。④从功能上讲,特征工程类似于常规定

① UN Global Pulse,“Building Proxy Indicators of National Wellbeing with Postal Data”.

② 在科学研究中,“概念”是被研究的抽象观念或现象(如教育绩效),“变量”是这个概念的(某个/些)性质或特点(如在校成绩),“指标”则是测量或者量化变量的方式(如年度成绩报告),而把抽象概念转化为可测量的变量和指标的过程称为“操作”(operationalization)。“数据”(data)则是收集起来用于参考或分析的事实和数据。

③ Robert Northcott, “Big Data and Prediction: Four Case Studies,” pp.96-104.

④ Jason Brownlee, “Discover Feature Engineering, How to Engineer Features and How to Get Good at It? ”September 26, 2014, Machine Learning Mastery, https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/.

量分析的数据操纵(manipulation)。特征工程主要有三个实现方式:一是整合在算法中。人脸识别运用拥有三个隐藏层的卷积神经网络(Convolutional Neural Networks, CNN)结构的算法,其中第一隐藏层习得线段等特征,第二层学习人的五官轮廓等特征,第三层学习人脸轮廓等特征。二是大数据研究者根据个人的数据经验以及对所分析数据的背后业务理解进行特征提取。如果我们要预测一个国家发生国内冲突的概率,就需要在模型中纳入人口总量这个自变量。由于世界范围内的人口分布严重偏离正态分布,我们通常会取这个变量的自然对数,以平衡数据内偏差。三是大数据专家和具有特定专业背景的专家合作实现特征提取。数据科学团队可以构建具有强大功能的模型,但不可能自行解决高度专业化的问题,所以需要与特定领域的专家开展密切合作。例如,中国人对电子邮件的使用不多;信用卡在印度尼西亚使用率很低。①数据专家会注意到这些细微差别,但没有专业知识来解释产生差别的原因,就会导致在营销上制定低效率或错误的策略。

数据降噪是大数据分析在数据准备过程中时间成本最高的一个环节。传统数据的收集需要先设计试验,然后调查收集,最后生成数据集。显然,大数据在收集的速度、范围、成本上要明显优于传统数据的收集方法。但由于大数据的主要来源是互联网上各种数字、文本以及音频和视频,还有各类传感器产生的数据等,数据中不可避免地夹杂着噪声。数据的质量是由信(号)噪(音)比来测量的。信号是测量信息的物理量,包括电信号、光信号等。信号是信息的载体,信号中携带着信息。噪声表示信号失真的程度。数据中信噪比越高,数据的质量就越高。2008 年,谷歌公司开发了旨在预测流感爆发的谷歌流感预测系统(Google Flu Trends)。该系统前期的预测结果与美国疾控中心的真实数据高度吻合。但在 2013 年 2 月,有专家批评 GFT 预测的流感病例门诊数超过了美国疾控中心公布的实际数字的两倍。随后,GFT 开发团队做了不少针对性的调整,但预测误差仍然偏高。2015 年 8 月,谷歌宣布停止发布当前估算。谷歌流感预测系统后期预测失灵的一个重要原因就是从数据中过滤新闻和社交媒体的影响不成功,降低数据中的噪声效果不明显,数据信噪比不高。②

为了让数据达到能够满足后续分析的需要,我们有时不得不选择牺牲掉数据中一些有价值的信息以提高信噪比。为了预测国民幸福感,需要收集月收入这个指标。如

① Huda Ali, "Five Insights about Harnessing Data and AI from Leaders at the Frontier," McKinsey Global Institute, March 25, 2021, https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/five-insights-about-harnessing-data-and-ai-from-leaders-at-the-frontier.

② 谷歌大数据工程师借助 5000 万搜索关键词训练 GFT 模型,然而很多关键词只是看似与流感相关,但实际上并无关联,将噪声误判为信号。

果让被调查者独立填写,有人可能会不负责任地提供不实信息,有些人可能略保守或略乐观:一个人真实月收入可能是 5300 元,保守的人可能就填写 5000 元,乐观的人可能填写 5500 元甚至更多。这些情况都会显著地增加数据中的噪音。如果我们采用让用户选择收入区间的方法,就会有效地降低噪音,提高整体的预测水平,但代价是可能会牺牲了对部分用户更加精准的预测。

在有些情况下,由于噪音难以消除,很多本来可预测的问题变得难以预测,甚至无法预测。对全球经济走势的预测就是一个经典范例。由于各个国家或地区的历史、地理、制度等各不相同,各种企业有着不同的发展和需求,诸多因素混杂在一起并且频繁互动,即使可以收集到很多经济相关数据,但数据中噪音太大,致使数据价值密度太低,无法实现有效预测。假如我们要对 2021 年的经济走势进行预测,新冠(Covid-19)疫情制造的噪音就是巨大的干扰因素,包括疫情引起的封城锁国、疫情的政治化、疫情暴发初期欧洲一些国家的"群体免疫"实践、新冠病毒的变异及其造成的欧美很多国家疫情的二次爆发,等等。

广泛存在的数据"噪音"警惕我们远离"大数据自大症",不能盲目地认为大数据胜过传统的数据收集与分析法。情况并不总是这样。须知数据收集的核心目的是使用,充分挖掘隐藏于数据中的信息和价值。"在大多数情况下,当我们需要使用大数据时,问题是要在大数据中找到合适的数据"。①价值密度过低的大数据在效用上很可能不如一个经过精心设计的试验收集到的小样本。

(二)分析建模

大数据和传统数据分析在方法上差异很大。大数据分析极大地推进了图像处理技术、自然语言处理技术和语音识别等技术的发展。从数学结构的角度来看,大数据操作和分析可以做得很巧妙,但其本质仍然是曲线拟合,虽然它可能很复杂、很困难。②分析建模的两个环节——算法设计和模型构建——都是曲线拟合。变量增多只不过意味着要在更高的空间维度中实现曲线拟合。

在数学和计算机科学中,算法(algorithm)是明确的有序序列,以使计算机实现指令,解决某类问题或执行计算。算法对预测精度的影响是显著的。例如,莱泽(David

① Andrej Zwitter, "Big Data and International Relations," *Ethics & International Affairs*, Vol.29, No.4, 2015, pp.377-389.

② Kevin Hartnett, "To Build Truly Intelligent Machines, Teach Them Cause and Effect," *Quanta Magazine*, May 15, 2018, https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/.

Lazer)等指出,流感暴发预测失准的一个重要原因是谷歌的搜索算法本身变化很快。[①]谷歌的搜索算法要关照更大范围的用户群体和不同的场景,所以会不时地根据现实的状态去调整算法。也就是说,谷歌搜索算法不是为预测流感而服务的,流感预测只是利用谷歌搜索的结果的一个附属服务产品而已。

大数据在商业、保险、医保等行业应用中,时常被诟病大数据技术或算法是有偏见的。实际上,算法是无辜的。造成所谓"大数据偏见"的原因一般有三个:一是收集到的数据本身就包含和反映了现实世界中的显性尤其是隐性的偏见,再经由模型展现了出来。在这种情况下,底层数据而不是算法本身通常是偏见的主要来源。[②]美国预防犯罪系统中对黑人的歧视本质上是社会中黑人较低的社会经济地位的以及对黑人的隐性歧视在数理模型上的表征。二是模型的使用者或决策者为了利益最大化或其他目的,操纵了大数据。比如,在某些非洲国家,经济发展滞后可能是国内民族冲突、不恰当的产业政策和政治文化等多因素共同作用的结果,但西方的一些政客为了自己的利益强行把这些国家的经济困境归因于缺乏民主和政治自由,从而对他国内政进行野蛮干涉。三是所选择的评价算法和模型指标不恰当。我们在评价分类模型的效果时,经常使用到的指标有准确率(Accuracy)、召回率(Recall)和精确率(Precision)等。但在很多实际预测中仅有这些评价指标是不完整的。在西方国家的选举中,竞选人都注重针对"微目标"即个人的宣传和争取,但假如在 A 地区有 5 位候选人,都通过自己的预测模型对一位选民进行争取,那么该候选人在信息"轰炸"下,他的实际选择可能与预测结果很不一致。类似地,在 20 世纪 40-50 年代,调查问卷刚兴起时,人们对这种新方式比较热情,应答率很高,但随着社会上各种问卷飞速增长,人们的心态发生了变化,出现了很多低应答率的情况,然后调查公司针对低应答率开发了多种应对方法。

大数据中的模型通常是数学模型。数学模型是用字母、数字及其他数学符号建立起来的等式或不等式,描述客观事物特征的图表、图像和框图等,以及揭示客观事物内在联系的数学表达式。模型的进步和发展对于分析预测的影响也是很显著的。自然语言处理技术(Natural Language Processing)中的预训练模型的发展即是很好的例子。自然语言处理任务可以分为三个模块:数据处理、文本表征和特定任务模型,其中只有文本表征模块可以作为一个通用模块来使用。自然语言处理领域可以预训练一个通

① David Lazer et al., "The Parable of Google Flu: Traps in Big Data Analysis," pp.1203-1205.
② Huda Ali, "Five Insights about Harnessing Data and AI from Leaders at the Frontier".

用的文本表征模块,这样对于文本的迁移学习(Transfer Learning)帮助较大。[①]以 Word2Vec 为代表的词向量技术是自然语言处理领域最常用的文本表征方法,其本质是一种静态的预训练模型,即不同上下文汇总同一词语具有相同的词向量,因而无法解决常见的多义词问题,直到 ELMO(Embedding from Language Models)提出了一种上下文相关的文本表示方法,并在多个典型任务中表现出色。此后,GPT(Generative Pre-Training)和 BERT(Bidirectional Encoder Representations from Transformers)等预训练模型相继被提出,自然语言处理进入了动态预训练技术时代。[②]与 ELMo 相比,GPT 采用了更强大的"转换器"(transformer)作为特征抽取器,但只能根据上文来预测词义。相比之下,BERT 则采用"转换器"作为特征抽取器,同时结合上下文预测词义。正是随着预训练模型的不断进步,自然语言处理在机器翻译、机器阅读理解方面才实现了今天的飞跃。

模型的基本构件是变量,而变量选择有赖于经验和理论。佩齐指出,了解所有潜在的相关参数是成功预测的一个基本前提。如果预测未能全面成功,该条件就没有得到满足。[③]对于社会科学家而言,要在实验室以外做预测非常困难,因为精准的预测需要控制每个相关因素。而相关因素既取决于经验,又取决于理论。政治竞选日益频繁地使用复杂的大数据方法,比如微目标选民。这是一种非理论的大数据方法。使用这种方法的关键是识别结果变量和公认/推定的解释变量之间的联系,前者比如实际投票,后者比如消费模式、媒体偏好等。[④]在大数据兴起之前,结果变量与因变量关系的识别在相当程度上取决于此前学者们得到证实和接受的因果推论。

提高模型预测的手段主要有两个:一是技术手段,主要依靠多模型组合或新的优化算法。集成学习(ensemble learning)方法中的随机森林算法(random forest)通过原始数据构建多个决策树,然后由决策树的大多数结果来确定模型预测的分类结果。例如,在冲突预测中构建了 10 个决策树,其中 8 个预测冲突会爆发,2 个预测不会,那么随机森林最后的输出的结果是会爆发。[⑤]二是利用专业知识和理论改进现有的指标,通常能够迅速地提高模型预测的准确率。例如,在卡哥(Kaggle)举办的泰坦尼克生存挑战赛中,正确率最高的模型之一的建模诀窍就是根据当时的社会背景设计了"是妇女还是儿童"这

① "迁移学习"是指在一个数据集上训练基础模型的过程中,通过微调等方式使得模型可以在其他不同的数据集上处理不同任务。

② 李舟军等:《面向自然语言处理的预训练技术研究综述》,载《计算机科学》,2020 年第 3 期,第 162-173 页。

③ Wolfgang Pietsch, "Aspects of Theory-Ladenness in Data-Intensive Science," pp.910-911.

④ Robert Northcott, "Big Data and Prediction: Four Case Studies," pp.97-98.

⑤ 集成学习是一种将多种学习算法结合使用的方法。

个虚拟变量。①

(三)模型应用和反馈

模型应用就是运用模型进行预测,反馈则是根据预测结果来检验、评估和调试模型,以便维持和提高准确度。从中长期来看,预测水平取决于预测低于预期时反馈的效果。而且,随着时间的变化,模型的效果没有新数据的增量训练,也会逐渐降低。商业银行在使用开发的信用卡评分模型时,每隔一段时间,都会根据新收集的数据重新评估和调整模型,以保持预测评分的准确性和稳定性。金融行业的实时反欺诈系统一旦在某地发现了新的欺诈案例,立刻实时更新到模型中,防止在其他地方再发生此类欺诈。

在国际政治预测中,反馈之后的模型调试主要包括两种情况:第一,数据的背景条件发生了改变,所以我们需要通过获取新数据来更新模型以便保持预测能力。20 世纪 60 年代,麦克莱兰(Charles McClelland)开发了一种编码系统,用于追踪政治危机中冲突与合作事件发生的速度和规模,被称为 WEIS(World Event/Interaction Survey)。该方案主要的功能是记录和描述国与国之间的行为。冷战后,尤其进入互联网时代后,由于非国家行为体在世界政治中日益活跃,WEIS 这种传统的事件数据编码方案逐渐过时。② GDELT 被视为 WEIS 的升级版,采用的是冲突与调解事件观察(Conflict and Mediation Event Observation,CAMEO)这个新的编码框架,重点是追踪和评价在全世界高度活跃的跨国公司、联合国各种附属组织和非政府间国际组织等非国家行为体。

第二,预测改变了对象的行动轨迹。在某些类型的预测中,这种反作用是影响预测准确度的一个重要因素。预测如果改变了对象的运行轨迹,不仅需要概念上的界定和理论上的阐释,而且要通过分析模型应用后的反馈数据来总结构建出有区分度的指标,然后再将数据或指标纳入现有的模型,这样才有可能同时提高模型的预测力和解释力,发现更多的因果关系。

模型反馈未必会提高预测准确率。投票预测的一个根本困难是选民的政治偏好和实际投票并不总是吻合。选民的性格、社交环境、投票日期前是否发生重要干扰事件等诸多因素都会影响到选民是否参加投票以及投票结果。专家们对 2015 年英国大选的预测很不准确,原因是选民的自我评价和政治偏好关联太强,导致样本偏差,所以有学者建议在今后的预测中纳入特定人口群体的历史投票率这个变量。2017 年,专家们使用了新的方法,但预测效果仍然很不尽人意,后来的调查发现,如果使用 2015 年的预测方法,

① 卡哥是谷歌旗下的一个全球性的数据建模和数据分析竞赛平台,网址是 https://www.kaggle.com/。

② 游祎:《基于大数据的世界政治冲突模型构建研究》,博士学位论文,外交学院,2020 年,第 45、56 页。

效果会更好。①

五 理论、因果关系和大数据预测

华尔兹(Kenneth Waltz)对理论的经典定义是“对规律的解释性表述”,而规律(law)则是(两个)变量间“一种被反复发现的关系”。②该定义的严密之处在于把理论界定为对某类事件(“反复发现的关系”)而不是一个事件(“一个被发现的关系”)的推测。这里的“理论”等同于因果关系。因果关系包括三个构件:一是变量在统计上具有相关性;二是在时间上有先后次序,自变量在前、因变量在后;三是在逻辑上自变量的发生导致了因变量的发生,③其中前两者是形式要件,后者则是实质要件。

大数据确实有助于进行因果推理和理论建构,但前提是数据质量要高。④首先,高质量的数据使因果关系的描述更加清晰。例如,棕榈油这样的原料物在发展中国家的数百万个农场中生产,在进入某跨国公司的一家工厂之前,它先要经过数千家炼油厂和工厂。这是一条很难追踪的供应链。但是,“轨道洞见”(Orbital Insight)公司能够使用地理位置数据和卫星图像来跟踪物理供应链。它的追踪并不基于准确性较低的文本,而是基于卡车行驶地点和森林砍伐发生地点的实时快照。⑤这样的话,如果某批货物脱链,就可以很快溯源,找到原因。其次,高质量的数据有助于形成解释,促进因果关系的发现和研究。高质量数据能显著地降低噪音对因果关系的干扰。开普勒(Johannes Kepler)的一个重要贡献是发现了行星围绕太阳运转的轨道是椭圆的。他的成就受益于从他的导师第谷(Tycho Brahe)那里继承的大量、精确的观测数据。再次,高质量数据有助于从相关关系中启发因果理解,“尿不湿与啤酒”就是一个经典案例。⑥最后,预测是应用和验证因果关系的一条重要途径,高质量的预测模型是以高质量的数据为前提的。

评价大数据模型预测效果的手段是交叉验证,根据结果的准确度来判断模型的质

---

① Patrick Sturgis et al., “Report of the Inquiry into the 2015 British General Election Opinion Polls,” 2016, Market Research Society and British Polling Council, http://eprints.ncrm.ac.uk/3789/; Anthony Wells, “Why the Polls Were Wrong in 2017,” 2018, http://ukpollingreport.co.uk/blog/archives/10002.

② [美]罗伯特·基欧汉编:《新现实主义及其批判》,郭树勇译,北京大学出版社 2002 年版,第 24、29 页。

③ 谢宇:《回归分析》,社会科学文献出版社 2010 年版,第 162-169 页。

④ Robert Northcott, “Big Data and Prediction: Four Case Studies,” p.103.

⑤ Huda Ali, “Five Insights about Harnessing Data and AI from Leaders at the Frontier”.

⑥ 沃尔玛一家分店经理在进行销售分析时,发现一个现象:啤酒与尿不湿的销量在周末总会出现成比例增长。随后,他派人对这些顾客进行观察并发现如下特点:已婚男士为主;家中有不到 2 岁的小孩;喜欢看体育比赛;而体育比赛一般都集中在周末。

量,不用深入分析变量间的关系。①大数据预测通常采取算法或黑匣途径,理论阐释被排除在外。换言之,大数据预测能够辨别因果关系的两个形式要件,即变量的相关性以及时序性。当这两个条件得到确认,我们只要能够在逻辑上推导出自变量发生导致了因变量的发生,大致就可以确认因果关系了。由于第四范式的这个特点,不少学者认为,大数据时代的科学研究只需要利用大数据技术直接分析海量数据,发现相关关系,就能获得新知识,所以"预测优于解释或者因果理解"。②还有的学者宣称大数据预测宣告了理论的死亡。③

诚然,因果关系对于大数据预测并不总是必要的。因果关系通常被用作建模的基础。但对于天气预报这样的问题领域而言,与其使用常规的大数据方法,④不如通过机器学习来盲预测。所谓盲预测,可以理解为机器学习是通过"训练数据"来构建预测模型。常用的气象预报模型是各种理论整合的结果,而盲预测模型几乎不依赖理论基础:"既然从现有的天气模型进行因果推理的能力缺失,那么,黑匣路径的机会成本降低了。假如天气系统确实是混沌的,那么只有概率性预报是可能的。"⑤

大数据的优势是对资料进行相关分析。现有大数据研究和应用重点致力于判断行为,而不是寻找因果关系,大数据分析因此被简化为准确预测。贝叶斯网络之父珀尔(Judea Pearl)表示:"随着我尽可能深入地研究深度学习,我发现它们都被困在了相关性的层面。也就是曲线拟合。"⑥陈定定在论及预测效果评价时提出三条评价标准:预测内容的现实性和精确性以及被预测事件或状态的发生机制。⑦这后一个标准指的就是因果关系。实际上,对因果关系的关注不仅有利于发现新知识、启发新理论,也会增强模型的预测能力。

实际上,大数据预测的整个过程都离不开理论的指导。数据收集貌似与理论没有联系,实际上也深受理论的影响。首先,预测者最好能够掌握预测目标、专业知识和理论等

① 漆海霞:《大数据与国际关系研究创新》,载《中国社会科学》,2018 年第 6 期,第 170 页。

② Robert Northcott, "Big Data and Prediction: Four Case Studies," p.96.

③ Tony Hey et al., eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, 2009; Victor Mayer-Schoenberge and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Eamon Dolan, 2014.

④ 常规的大数据方法通过大量数据来近似求解方程参数,然后再根据方程进行预测,其中参数选择和求解显著依赖理论。

⑤ Benedikt Knüsel et al., "Applying Big Data beyond Small Problems in Climate Research," *Nature Climate Change*, Vol.9, No.3, 2019, pp.196-202.

⑥ Kevin Hartnett, "To Build Truly Intelligent Machines, Teach Them Cause and Effect."

⑦ 陈定定、刘丰:《认真对待预测:国际关系理论发展与预测》,载《世界经济与政治》,2012 年第 1 期,第 25 页。

背景知识。漆海霞在分析中国与大国关系的影响因素时,依据自己的专业知识确定了以下因素:国际格局、权力转移、核武器、经济相互依赖和意识形态,然后根据各因素的相关知识和理论,来确定指标,选择要收集的变量数据。①其次,更多的数据对提高预测水平通常会有帮助,但数据需要工具来搜集,选择什么工具、如何安置则是由预测模型决定的,而模型又是由外生理论决定的。② 2012 年,IBM 与嘉露酒庄合作,希望构建一个使葡萄"既增产又增质"的分析预测系统。这样的期待超过了大数据天气分析和预报的范畴,要求借助地面传感器完成更广范围的数据收集工作,还需要利用高清卫星图像遥感技术来分析葡萄园的红外波以及时了解和跟踪葡萄园的详细状况,最后再整合这些数据进行分析,并在此基础上对葡萄树实施个性化管理,针对具体情况确定灌溉与施肥量。③

类似地,算法选择虽然主要根据计算目的和数据特点,但如果有理论的支持,效率也会显著提高。④"适应性提升"(Adaptive Boosting, Adaboost)算法就是一个典型的理论推动算法实践的案例。1988 年,科恩斯(Michael Kearns)在研究时提出了一个有趣的问题:"弱学习(weak learning)是否等价于强可学习(strong learnable)"。如果这个问题的答案是肯定的,那只要找到比随机猜测略优的弱学习算法,⑤就可以将其提升为强学习算法,而不必直接去寻找通常很难获得的强学习算法。沙皮尔(Robert Schapire)通过构造性方法证明了一个概念是弱可学习(weakly learnable)的,当且仅当它是强可学习的,从而对上述问题做出了肯定的回答。弱可学习与强可学习等价这个定理奠定了"提升"(Boosting)算法的理论基础。弗洛因德(Yoav Freund)等在更深入的研究中发现在线分配问题与提升之间存在着很强的相似性,引入在线分配算法的设计思想将有助于设计出更实用的提升算法。接下来,他们把加权投票的相关研究成果与在线分配问题结合,并在提升问题框架下进行对应推广,推导出了著名的适应性提升算法。⑥

高质量的搜索结果是各大搜索网站吸引和保留互联网用户的重要因素。2005 年之前谷歌主要是通过因果关系来提高预测率。因果关系的发现毕竟比较困难,所以,2005 年后,谷歌利用与"用户点击数据"的高相关性迅速提高了预测准确率。这个事件并不意味着相关性就代替了因果性,反而凸显了因果性的重要性,因为相关性作用的发掘是以已经习得的因果关系为前提的。程序员如果发现某些搜索结果低于预期,就会首先分

① 漆海霞:《中国与大国关系影响因素探析——基于对 1960-2009 年数据的统计分析》,载《欧洲研究》,2012 年第 5 期,第 61-78 页。

② Robert Northcott, "Big Data and Prediction: Four Case Studies," pp.99-100.

③ [美]史蒂夫·洛尔:《大数据主义》,第 180-188 页。

④ Robert Northcott, "Big Data and Prediction: Four Case Studies," pp.99-100.

⑤ 弱学习算法,通俗地讲就是比随机预测效果略好但准确率却不太高的算法。

⑥ 曹莹等:《AdaBoost 算法研究进展与展望》,载《自动化学报》,2013 年第 6 期,第 745-758 页。

析原因,然后做出改进。随着数据量的积累,研究人员发现搜索质量与用户大量的“点击数据”——即用户输出关键字后在搜索结果中点击的网页记录——相关性很强,利用这个特性可以迅速提高搜索结果的质量。现在,各个搜索引擎都有一个度量用户点击数据与搜索结果相关性的模型,被称为“点击模型”。随着数据量的增加,点击模型对搜索结果排名的预测越来越准确,虽然它在各个搜索引擎的搜索算法中的权重大概都在 60%以上。①尽管如此,搜索算法中必要的因果关系或规律是发挥“点击模型” 作用的基础。正如珀尔(Judea Pearl)所强调的那样,如果我们要对干预措施的有效性进行分析,那么我们必须引进因果模型;仅有相关性是不够的,这是个数学事实,而不是建议。②

大数据预测作为一个过程,其实包括三个实质要件:数据、(预测)模型和研究者,其中数据和模型是显性的,而研究者是隐性的。数据是由研究者搜集和处理的,模型也是由研究者来设定和调试的。不同的研究者具有不同的观念、知识背景和理论偏好,所以,面对同样的原始数据资源和技术手段,不同的研究者可能会做出大相径庭的预测。换言之,如果预测的基本前提得到满足,数据、模型和研究者偏好三个因素共同决定了预测的效果。

大数据分析是一门综合技术,是跨学科的。不同专业学者的知识和技术对于提高预测效果起着重要的作用。③ 国际政治大数据预测要求研究者既具备比较政治、国际关系或外交学的专业背景,又有计算科学、统计学等数据科学的专业知识。由于大数据产生的多源性以及大数据特征的多维性,为了更好地理解和应用数据,我们还需要和数据相关专业的技术人员开展合作。2002 年,IBM 斥资 35 亿美元,购买了普华永道会计师事务所的咨询部门,将数以千计拥有特定行业专业知识的咨询师收归旗下。IBM 的技术人员和这些行业专家合作,进行数据分析,不仅大幅度提高了公司的营业收入,而且为公司后续的大数据转型奠定了良好的基础。他们的专业知识有助于我们更透彻地理解数据本身,他们的专业经验更是一种难以量化的“大数据”。类似地,上文中提到的 GDELT 是目前数据最全、最及时的开源事件数据库。它把事件类型分为 20 大类和 300 多小类。这些类别制定需要拥有专业知识和理论的专家。相应地,数据库的使用者如果对这些分类具备一定的专业知识,数据分析和使用的效率就会得到显著提高。

① 吴军:《智能时代:大数据与智能革命重新定义未来》,第 137 页。
② Kevin Hartnett, “To Build Truly Intelligent Machines, Teach Them Cause and Effect”.
③ Robert Northcott, “Big Data and Prediction: Four Case Studies,” pp.102-103.

六 结论:国际政治大数据预测的特点和限度

数据是 21 世纪的“石油”。各个行业都在尝试开发和利用这块资源。大数据在国际政治领域的应用尚处于起步阶段,新尝试和新努力自然会遇到各种困难和限制。本文初步梳理和探讨了影响大数据国际政治预测准确性的主要因素,包括数据、方法和理论三个维度。作为一篇归纳性的论文,其主要贡献在于按照预测的工作程序,分析了大数据预测在各个工作环节面临的约束条件。文章的基本结论是在同等条件下满足的条件越多,大数据预测的准确率就越高。

过去半个世纪以来,气候预报水平的巨大提升为我们的论点提供了一个生动的注脚。气候预报显著地受益于大数据。七天预测的准确率和三天预报相差无几。自从 20 世纪 60 年代发射第一颗气象卫星以来,数据的质量和数量急剧增长,而气候预测模型的核心始终是过去几百年来已知的流体力学/动力学差分方程。与此同时,一系列的特别(ad hoc)条件被不断地纳入方程,来兼容那些形形色色的显著影响因子——如山川、云以及空气运动和洋流的整合,并在不断的试错过程中实现更新。①从 90 年代晚期开始,科学家开始控制随机变量,通过多重模拟来产生概率预测。由于气候是一个混沌系统,多次反复后,偏差会相互取消,概率预测变成无偏预测。而且,急剧提升的计算能力让数据得到了充分的利用。②

大数据预测在人类很多学科领域都得到了应用。相比天气预报这样的领域,大数据比较政治和国际关系预测尚处于较低的发展阶段。大数据政治学与非政治学预测的本质区别在于数据特点的差异。首先,国际政治相关数据有着显著的政治风险。相关数据采取和使用大概率会涉及政府机要和国家安全等敏感信息,数据收集的门槛很高,国家间的数据交流困难重重。相比之下,气象数据的收集是全球共享的,获取成本很低。其次,国际关系事件实例相对少而且有不可重现和复制。要预测当代世界政治中的国际或地区冲突,我们能收集的同类实例数量很有限。1946-2003 年间,世界范围内年均导致至少 1000 人以上阵亡的国际战争仅有 38 次。③由于数据稀缺,我们训练出来的模型预测

① Thomas Jung et al.,“The ECMWF Model Climate: Recent Progress through Improved Physical Parametrizations,” *Quarterly Journal of the Royal Meteorological Society*, Vol.136, No.650, 2010, pp.1145-1160; Jean Bechtold et al., “Progress in Predicting Tropical Systems: The Role of Convection,”Report No.686, Affiliation: ECMWF, November 5, 2012, pp.1-61.

② Robert Northcott, “Big Data and Prediction: Four Case Studies,” p.99.

③ 战争次数根据“战争关联数据库”的“国际战争数据”计算得出。数据来源:Meredith Sarkees and Frank Wayman, *Resort to War: 1816-2007,* CQ Press, 2010,https://correlatesofwar.org/data-sets/COW-war。

效果会大打折扣;即使预测效果令人满意,模型也得不到即刻验证的机会,因为高烈度的战争可能十几年才发生一次。当此类事件再次发生时,构建模型的条件和数据可能已经大面积发生了变化。相反,气象数据所预测的晴雨风雪则是自然常态,机器学习可以一直持续下去。

值得一提的是,对于大数据预测能力而言,数据、方法和理论不仅难以截然分开,而且三者的影响是可以"叠加"的。董青岭把基于大数据的冲突预测形象地称为"大数据安全态势感知"(Situation Awareness Based on Big Data)。他以 2013-2017 年的新闻报道数据和社交网络数据为研究对象,试图通过预测英国的恐怖袭击来考查大数据安全态势感知的预测能力。该案例的设计思路是以民众与政府之间的双向互动为聚焦点,考察政府对民众的言语和行为以及民众对政府的言语和行为,以便捕捉冲突的信号。①董青岭基于社会冲突、国际冲突相关知识和理论素养,构建了冲突特征向量,由信号内容和信号频率两个指标构成,其中信号频率主要包含以下指标的涉英政府每日新闻报道数和社交发帖量:民众诉求、政府赞成、民众抗议、政府拒绝、民众威胁、政府强制、民众攻击和政府打击,②并在此基础上建立了基于前馈神经网络算法的预测模型。

如前所述,我们在讨论数据准备对预测能力的影响时,省略了数据存储这个环节。不仅如此,在数据获取和数据预处理之间,还存在数据合并这个步骤,也就是把来源不同的巨量数据依据某个(些)指标合并成一个综合数据。③例如,汤姆斯库-杜布劳(Irina Tomescu-Dubrow)和斯洛姆茨恩斯基试图建立一个纵贯过去半个世纪、覆盖 142 个国家和地区的政治行为、社会态度和人口特征数据库,其数据来源包括 22 个知名的国际问卷调查机会,涉及 230 万被调查者。如此规模的数据合并工程在技术上未必复杂,属于劳动密集型的统计和计算工作,但很容易出现难以觉察的偏差,影响到数据的质量。例如,各国人民对冲突认定的标准不一样,造成同样性质的事件,在不同国家的问卷回答不一致,合并这样的数据难以避免偏差,关键是如何让偏差最小化。

值得一提的是,并不是数据越多、计算能力越强、方法越先进、理论越丰富,预测效果就越好。近半个世纪以来的 GDP 预测就是如此令人沮丧:假定真实 GDP 增长率不变

① 董青岭:《大数据安全态势感知与冲突预测》,载《中国社会科学》,2018 年第 6 期,第 179 页。

② 同上文,第 180 页。

③ See Nicholas Retih et al., "Building Cross-National, Longitudinal Data Sets: Issues and Strategies for Implementation," *International Journal of Sociology*, Vol.46, No.1, 2016, pp.21-41; J. Craig Jenkins and Thomas Maher, "What Should We Do about Source Selection in Event Data? Challenges, Progress, and Possible Solutions," *International Journal of Sociology*, Vol.46, No.1, 2016, pp.42-57; Irina Tomescu-Dubrow and Kazimierz Slomczynski, "Harmonization of Cross-National Survey Projects on Political Behavior: Developing the Analytic Framework of Survey Data Recycling," *International Journal of Sociology*, Vol.46, No.1, 2016, pp.58-72.

化,那么 12 个月的预测和基线预测相差不大,18 个月的预测则低于基线。而且,GDP 增长的拐点预测几乎总是失败。① GDP 预测的困境主要由两个原因造成:一是 GDP 的决定因素非常复杂,现存的预测方法都捕捉不到所有的因果过程(generating process);二是已经捕捉到的因果过程很不稳定,所以即使最先进的机器学习技术也于事无补。②宏观经济可能不仅是开放系统,不断地受到非经济因素的影响,而且无论经济因素还是非经济因素,很多都是观察不到的。不仅如此,国民经济还是个反思性系统,预测会影响到结果,而且是一个混沌系统,对于初始条件非常敏感,充其量只能做概率性的预测。不仅如此,GPD 测量偏差也很大,而且季节性调整会影响数值的精确度。由于上述这些原因,不同的测量方法会产生大相径庭的结果。任何大数据方法都无法同时解决所有这些问题。而且,即使非稳定性在一定程度上得到克服,大数据对 GDP 的预测效果仍然很不乐观。③

除此以外,大数据预测并不适合某些特定类型的事件,较有代表性的由实验结果外推(extrapolation)的田野预测。以美国政府对广播频谱的拍卖为例,20 世纪 90 年代中期以来,联邦通讯委员会对广播频谱进行拍卖。拍卖总体是成功的,频谱资源得到了有效的分配。然而,拍卖的成功既无关竞标人的新数据,也无法应用已经产生的拍卖数据,并且不依赖对这些数据的分析。相反,拍卖数据是在实验过程中产生的,其中起作用的是实际的操作知识,比如是否要求预付款,单独拍卖还是一揽子拍卖等等。更重要的是,由于拍卖具有很强的场景和事件具体性特征,其他来源的数据对于建模和预测很难起到作用,机器学习或数据挖掘也无法得到有效运用。预测所需的数据只能通过反复的实验和试错来产生。至于具体需要哪些数据,则要根据具体的场景和背景理论来决定,它们也是决定预测效果的主要因素。④

(**作者简介**:卢凌宇,云南大学国际关系研究院非洲研究中心教授;张传基,云南大学历史与档案学院世界史专业硕士研究生。**责任编辑**:宋晓敏)

① Gregory Betz, *Prediction or Prophecy? The Boundaries of Economic Foreknowledge and Their Socio-Political Consequences*, Deutscher Universitätsverlag Verlag, 2006, pp.30-38; Prakash Loungani, "How Accurate Are Private Sector Forecasts? Cross-Country Evidence from Consensus Forecasts of Output Growth," *International Journal of Forecasting*, Vol. 17, No.3, 2001, pp.419-432.

② Gregory Betz, *Prediction or Prophecy? The Boundaries of Economic Foreknowledge and their Socio-Political Consequences*, pp.101-108.

③ Robert Northcott, "Big Data and Prediction: Four Case Studies," p.100.

④ Ibid., pp.100-101.