机器学习在冲突预测方面的局限*

——基于对暴力预警系统的再检验与讨论

刘辰辉　唐世平

【内容提要】 随着大数据时代的来临,各领域的研究越来越依托于大数据,机器学习算法已成为社会科学研究的热门工具之一。在和平与冲突研究领域,作为预测冲突的重要手段,机器学习算法认为通过大量数据训练得到的模型能够准确预测国家的冲突行为和冲突事件。一些项目并没有事前公开其数据、算法和代码,因此无法对其预测能力进行评估。另一些项目和研究虽然公布了它们的数据和算法等,但在关键指标上表现欠佳。通过检验暴力预警系统,发现该系统基于机器学习的相关分析对非洲地区冲突的预测效果并不理想。事实上,无论是基于统计学建立的因果关系模型还是机器学习挖掘出的关联性规则,都无法精准预测未来发生的冲突,其原因在于国家的内外部环境会随着时间推移而发生变化。新冲突的发生可能受全新因素的影响,也可能受因素之间相互作用的影响,还可能是由不同机制驱动所致,但这些因素都无法被机器学习与算法所捕捉,这导致相关模型和算法缺乏外部效度。通过机器学习算法的方式预测冲突这一涌现性结果的前景并不乐观,研究者在对冲突进行预测时还需要融合多种方法。

【关键词】 机器学习;冲突预测;社会结果;非洲地区冲突;行动

【作者简介】 刘辰辉,复旦大学复杂决策分析中心高级数据分析师;唐世平,复旦大学国际关系与公共事务学院教授(上海　邮编:200433)。

【中图分类号】 D815 【文献标识码】 A 【文章编号】 1006-9550(2023)12-0114-30

* 本文系2019年上海市哲学社会科学规划青年课题“非洲国家族群冲突风险预测”(项目批准号:2019EGJ002)的阶段性成果。感谢《世界经济与政治》匿名审稿专家提出的宝贵意见与建议,文中疏漏由笔者负责。

一 引言

暴力冲突尤其是大规模暴力冲突对任何社会都会带来灾难性后果,除阻碍经济增长外还会导致大量人员伤亡并引发难民潮。研究者通过模拟五种“共享社会经济路径(SSP)”,揭示了暴力冲突对人均国内生产总值(GDP)的影响,认为不同场景下的冲突会使全球经济的增长率下降 20%—30%。① 根据乌普萨拉冲突数据项目(UCDP)的编码,暴力冲突分为国家冲突(SB)、国内非政府力量冲突(NS)和针对平民的暴力活动(OS)。冷战结束后,世界范围内的各类冲突数量呈上升趋势。提前预测可能爆发的暴力冲突对预防、缓解和化解冲突都具有重要意义。在和平研究领域,有学者认为“预测始终是和平研究的首要任务”。②

哈弗德·赫格(Havard Hegre)等认为,随着大数据时代的到来,和平领域的研究改变了过去数据收集不全面、不成熟、无法对武装冲突进行准确预测和预警的状况,机器学习一度成为预测暴力冲突的有效方案。③ 虽然通过大量数据训练得到的相关模型可以一定程度上预测国家的冲突行为和冲突事件,但冲突的爆发是一个非常复杂的过程,相关预测往往无法应用于实际政策。有学者指出,冲突过程中的各种因素和参与者的互动非常复杂,因此冲突可能无法被预测。相较于结构化的决策过程(如投票)而言,冲突预测要求有更精确的数据作为支撑和对介入各方互动模式的全面了解,目前的研究只能通过不断增加与冲突概率相关的结构特征进行风险评估。④ 换言之,冲突预测不仅需要综合各类数据并做出风险评估,还需要考虑冲突各方互动的复杂过程。

在冲突预测领域,乌普萨拉冲突数据项目开发了一款基于机器学习的冲突预测系统即暴力预警系统(ViEWS)。暴力预警系统具有数据公开、算法代码公开和预测结

---

① Kristina Petrova, et al., “The ‘Conflict Trap’ Reduces Economic Growth in the Shared Socioeconomic Pathways,” *Environmental Research Letters*, 2023, DOI:10.1088/1748-9326/acb163.共享社会经济路径是为了研究和预测全球气候变化对社会经济发展的影响而设计的一组情景,将不同的社会经济发展趋势和气候政策组合在一起以描述未来可能出现的发展路径,包括可持续发展路径、中间路径、分化路径、不平衡增长路径和强化不平等路径等。

② David J. Singer, et al., “The Peace Researcher and Foreign Policy Prediction,” in Jody B. Lear, Diane Macaulay and Meredith Reid Sarkees, eds., *Advancing Peace Research*, London: Routledge Press, 2012.

③ Havard Hegre, et al., “Introduction: Forecasting in Peace Research,” *Journal of Peace Research*, Vol.54, No.2, 2017, pp.113-124.

④ Lars-Erik Cederman and Nils B. Weidmann, “Predicting Armed Conflict: Time to Adjust Our Expectations?” *Science*, Vol.355, No.6324, 2017, pp.474-476.

果公开的特点,会定期更新和发布数据和预测结果。然而,对该系统进行分析和检验后发现,其在预测非洲地区冲突方面的效果并不理想。事实上,无论是基于统计学模型建立的因果关系模型(冲突因子与冲突爆发),还是机器学习挖掘出的关联性规则,两者都不能对未来发生的冲突进行准确预测且实际预测的精准率远不如预期,主要原因在于相关研究在传统统计、机器学习和深度学习等算法中忽略了冲突各方互动的过程。本文研究采用的数据基于乌普萨拉冲突数据库,具体而言,在空间层面的数据分为国家级别数据(C)和来自奥斯陆和平研究所(PRIO)提供的地理信息系统(GIS)的网格点(PG)数据,时间层面的数据则是月度(M)级别数据。需要说明的是,由于测量的问题,时间层面的大部分数据实际上是从年度(Y)级别数据转化而来的。从暴力预警系统公布的结果来看,在国家—月度(CM)级别上的预测表现远好于空间网格—月度(PGM)级别上的表现,因此本文只采用 CM 级别数据的预测结果。由于最终目标是将预测结果应用于国家层面的政策制定并应对地区高强度的政治风险,因此本文更加关注国家冲突。

二　预测研究:解释还是预测

(一)冲突研究的发展

特雷西·范·霍尔特(Tracy Van Holt)等指出,20 世纪 60 年代的冲突研究主要关注的是战争和国家冲突,随后的研究将对象逐渐扩展至国内冲突和其他国内政治暴力等问题领域。随着冲突理论的发展,冲突研究的目标也从早期的解释性研究逐渐发展至预测性研究,并涉及国内、国际、民族、环境、天气和地理等主题的连贯性研究领域。① 数据更新和技术迭代对冲突领域的研究起着至关重要的作用。

数据的作用是让研究者通过因果推断方法来探究冲突爆发的机制。随着数据采集和编译技术的进步,研究者创建出越来越多可靠、符合政治学理论和更加精细的数据库。数据资源的发展既推动了对现有冲突理论的评估和验证,也进一步增强了学界对研究问题的认识。② 对数据的挖掘与分析涉及冲突事件各方的行为和策略等精细化的内容,该方法能将冲突爆发的过程分解到一个极小的颗粒度,这有助于研究者更加深层次地理解冲突并对相关理论进行有效评估。战争相关因素(COW)数据库由戴

① Tracy Van Holt, et al., "The Role of Datasets on Scientific Influence Within Conflict Research," *PLOS ONE*, 2016, DOI:10.1371/journal.pone.0154148.

② Kristian Skrede Gleditsch, et al., "Data and Progress in Peace and Conflict Research," *Journal of Peace Research*, Vol.51, No.2, 2014, pp.301-314.

维·辛格(David J. Singer)等于 1963 年建立。该项目开启了对战争触发性因素的定量研究,同时也催生了对战争的预测研究。① 学界随后不断对数据的颗粒度进行深化和细化。例如,在 20 世纪 80 年代初期,学界构建了冲突与和平数据集(COPDAB)以及军事化的国家间冲突(MID)数据库。② 1993 年,彼得·瓦伦斯滕(Peter Wallensteen)等建立了乌普萨拉冲突数据项目,记录了冷战末期爆发的约 90 起武装冲突的数据。③ 1995 年,基思·贾格尔斯(Keith Jaggers)等发布并记录了 161 个国家的制度指标数据——政体 3(Polity Ⅲ)数据。④ 这些研究数据和相关研究工作为后来的冲突分析与预测研究奠定了基础。⑤ 自 2012 年以来,奥斯陆和平研究所构建了以地理信息系统网格点为观察对象的数据库,许多基于空间网格点级别数据的研究随之将冲突预测精度提升到国内局部冲突上。⑥ 总体来说,随着时代进步和数据提取技术的发展,可用的战争数据变得更加细致。相关数据既包含了参与冲突的个人、组织和国家,还包含了冲突各方的动机和能力等方面的参数,这些数据为研究者理解冲突爆发提供了数据基础。

以数据为基础,冲突领域的相关研究出现了爆发性增长,同时也推动了和平研究的发展。菲利普·施罗特(Philip A. Schrodt)利用神经网络(ANN)展开冲突预测研究,他认为随着模型参数和学习算法的不断优化,网络模型在冲突预测问题上的表现

---

① David J. Singer and Melvin Small, "The Wages of War, 1816-1965: A Statistical Handbook," *Journal of Peace Research*, Vol.11, No.1, 1974, pp.76-78.

② Edward E. Azar, "The Conflict and Peace Data Bank (COPDAB) Project," *The Journal of Conflict Resolution*, Vol.24, No.1, 1980, pp.143-152; Charles S. Gochman and Zeev Maoz, "Militarized Interstate Disputes, 1816-1976: Procedures, Patterns and Insights," *Journal of Conflict Resolution*, Vol.28, No.4, 1984, pp.585-615; Daniel M. Jones, Stuart A. Bremer and J. David Singer, "Militarized Interstate Disputes, 1816-1992: Rationale, Coding Rules and Empirical Patterns," *Conflict Management and Peace Science*, Vol.15, 1996, pp.163-213.

③ Peter Wallensteen and Karin Axell, "Conflict Resolution and the End of the Cold War, 1989-93," *Journal of Peace Research*, Vol.31, No.3, 1994, pp.333-349.

④ Keith Jaggers and Ted Robert Gurr, "Tracking Democracy's Third Wave with the Polity Ⅲ Data," *Journal of Peace Research*, Vol.32, No.4, 1995, pp.469-482.

⑤ 20 世纪 80 年代,学界对定量数据集的研究取得了较大发展,博弈论也被应用到冲突预测中。博弈论方法主要通过预测行为体可能的冲突行为预测冲突。本文聚焦基于统计分析和机器学习的冲突预测,不对博弈论模型进行讨论。参见 Bruce Bueno de Mesquita, "An Expected Utility Theory of International Conflict," *American Political Science Review*, Vol.74, No.4, 1980, pp.917-931; Bruce Bueno de Mesquita, "Forecasting Policy Decisions: An Expected Utility Approach to Post-Khomeini Iran," *Political Science & Politics*, Vol.17, No.2, 1984, pp.226-236。

⑥ Andreas Forø Tollefsen, et al., "PRIO-Grid: A Unified Spatial Data Structure," *Journal of Peace Research*, Vol.49, No.2, 2012, pp.363-374.

将优于线性模型的预测。[①] 深度学习模型能够展现传统回归模型无法捕捉的复杂机制,而网络模型能够证明变量间存在的相互作用机制。此外,还有研究分析了冲突各方互动的过程,如巴里·西尔弗曼(Barry G. Silverman)等利用代理人模型(ABM)并结合行为博弈论方法研究了冲突各方领导人选择不同政策导致的结果。[②] 还有学者将研究目标直接放在预测上,如帕特里克·布兰特(Patrick T. Brandt)等利用马尔科夫转换模型(Markov switching model)与贝叶斯向量自回归模型(Bayesian vector autoregression model)来预测国内和国际冲突。[③]

(二)解释与预测:从研究到应用

长久以来,冲突研究聚焦于解释冲突爆发的原因和过程,如斯科特·贝内特(D. Scott Bennett)等把军费开支、贸易和联盟等变量与冲突理论联系起来,通过回归分析模式验证它们之间存在的因果机制。[④] 但贝内特的研究在预测方面的表现欠佳,其解释模型仅通过检验结果参数 P 值而得出,因此预测性能较差。[⑤] 该问题的出现主要有两点原因:一是这类研究挖掘出的因子其实是"边缘变量",虽然这些变量在研究中存在较强的因果机制,但它们在实际的社会与政治过程中影响较小。二是由于社会发展、国际关系演变和技术迭代等原因,基于旧数据所衍生的理论在新环境下并不适用,而新的预测工具和方法可以用来探究暴力冲突发生的因果机制并建构新的冲突理论。有学者认为,预测有助于发现与既有理论不符的新事实,可以通过对数据的进一步挖掘,发展并演化出新的理论。[⑥] 新的冲突理论除了能够解释冲突爆发的原因外,还应具有预测的能力。

还有一些学者认为,大多数机器学习算法之所以出现解释性不足的问题,是因为它们缺乏透明度,宛如一个无法理解的"黑箱"。马亚·克里希南(Maya Krishnan)等

---

① Philip A. Schrodt, "Artificial Intelligence and the Study of International Politics," *American Sociologist*, Vol.19, No.1, 1988, pp.71-85; Philip A. Schrodt, "Prediction of Interstate Conflict Outcomes Using a Neural Network," *Social Science Computer Review*, Vol.9, No.3, 1991, pp.359-380.

② Barry G. Silverman, et al., "Modeling Factions for 'Effects Based Operations': Part Ⅱ-Behavioral Game Theory," *Computational & Mathematical Organization Theory*, Vol.14, No.2, 2007, pp.120-155.

③ Patrick T. Brandt, et al., "Real Time, Time Series Forecasting of Inter-and Intra-State Political Conflict," *Conflict Management and Peace Science*, Vol.28, No.1, 2011, pp.41-64.

④ D. Scott Bennett and Allan C. Stam, *The Behavioral Origins of War*, Ann Arbor: University of Michigan Press, 2004.

⑤ Michael D. Ward, et al., "The Perils of Policy by P-Value: Predicting Civil Conflicts," *Journal of Peace Research*, Vol.47, No.4, 2010, pp.363-375.

⑥ Ad Feelders, "Data Mining in Economic Science," https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=fd08484d50f4bfafd8078bf5fdee57e3e948f471,访问时间:2023 年 10 月 13 日。

认为,“可解释性”有时并不是必需的,而透明度高的算法容易被控制、监测和纠正。[①] 然而,可解释性的机器学习算法在高风险决策上是十分必要的。可解释性的内涵有两点:一是有助于理解决策过程及原因,[②]二是运用预测模型进行多次检验时其结果能表现出一致性。[③] 冲突预测不仅被用于挖掘冲突爆发的因果关系,而且能为决策者制定政策时提供支撑。决策者如果想要了解在不同资源配置条件下各种冲突危机的后续演化情况,就需要构建一个充分且透明的预测模型。同时,该模型还应保持一定的稳定性,能充分反映出各种机制在冲突爆发前的运作情况。总之,如果要将预测结果应用在决策中,该冲突预测模型涉及的内容需要同时具备可解释性与预测性。[④]

在应用层面,只有具备可解释性与预测性的冲突研究才具备辅助决策者制定政策的价值。对于决策者来说,一国可以调配的资源总是有限的,因此他们既要考虑政策的短期影响,又要考虑政策产生的长期影响。短期影响事关冲突是否会爆发甚至加剧等问题,而长期影响事关政策的结果和后续走向。然而,既有的冲突预测研究普遍存在能力极限的问题,即随着预测模型变量的不断增加,其对准确性的边际贡献会逐渐减小并趋近于零。[⑤] 突破这一极限问题需要考虑多种因素,如模型、数据、假设和不确定性等。既有研究往往通过采用高频率数据或新的预测模型等方式解决多因素问题,但改进效果十分有限。

之所以将冲突预测与政策制定密切相连,主要是因为随着基于新闻报道的分类算法日益完善,与战争相关的数据在空间—时间上的分类变得越来越精细化。实证分析和预测研究逐渐从年度转向时间颗粒度更细的研究中,这也使得决策者开始意识到实现早期预警的可能性和建立冲突预测系统的必要性。在 20 世纪 80 年代,美国中央情报局(CIA)建立了期望效用模型波利康(Policon)系统并以此预测外国领导人的决策。[⑥] 进

---

① Maya Krishnan, “Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning,” *Philosophy and Technology*, Vol.33, 2020, pp.487-502; Brent Daniel Mittelstadt, et al., “The Ethics of Algorithms: Mapping the Debate,” *Big Data and Society*, 2016, DOI: 10.1177/20539517166796.

② Tim Miller, “Explanation in Artificial Intelligence: Insights from the Social Sciences,” *Artificial Intelligence*, Vol.267, 2019, pp.1-38.

③ Been Kim, et al., “Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability,” https://dl.acm.org/doi/10.5555/3157096.3157352,访问时间:2023 年 10 月 13 日。

④ Galit Shmueli, “To Explain or to Predict?” *Statistical Science*, Vol.25, No.3, 2010, pp.289-310.

⑤ Thomas Chadefaux, “Conflict Forecasting and Its Limits,” *Data Science*, Vol.1, No.1-2, 2017, pp.7-17.

⑥ Sean P. O'Brien, “Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research,” *International Studies Review*, Vol.12, No.1, 2010, pp.87-104.综合早期危机预测系统目前为美国军火巨头洛克马丁公司所拥有。关于波利康系统的介绍,参见 Stanley Feder, “Faction and Policon: New Ways to Analyze Politics,” in H. Bradford Westerfied, ed., *Inside CIA's Private World: Declassified Articles from the Agency's Internal Journal*, New Haven: Yale University Press, 1995, pp.1955-1992。

入21世纪,美国中央情报局又资助建立了基于数据和计算的高烈度政治动荡(PITF)大型预测项目,旨在提前两年预测全球范围内的政变、革命和武装冲突等政治不稳定事件。① 此外,美国国防高级研究计划局(DARPA)还资助建立了综合早期危机预测系统(ICEWS),旨在实现自动监测、评估和预测所有可能影响美国国家安全的危机事件并为决策者提供建议。虽然综合早期危机预测系统具备丰富的测量维度和先进的算法,但该系统的相关研究结果、数据和算法并未全部公开。② 因此,学界很难对这类项目进行跟踪评估或研究改进。只有预测系统可以为决策者提供可靠而精准的预测并辅助他们提前进行有效干预,该系统才称得上一个合格的冲突预测系统。笔者认为,合格的冲突预测系统需要满足四点要求:一是具有准确度与精准性,即预测结果要在准确度与精准性上达到一定水平;二是具有科学性,即相关模型能系统地捕捉冲突爆发前的环境变量和事件,识别其中的因果关系;三是具有稳健性,即相关预测结果不能相差太大,预测系统应当是一个可控模型,能够及时对冲突事件做出反应或实现动态调整;四是具有拓展性,由于冲突预测直接服务于决策,预测结果既要包括冲突发生的多种可能性,还应涵盖出现的多种情景假设,这样才能为决策者提前干预冲突提供参考和帮助。

三 暴力预警系统的架构、数据、算法和评估体系

暴力预警系统是目前唯一实现了数据、算法、代码和预测结果公开且保持不断更新与优化的研究系统,也是一个透明度较高且受学界长期关注的冲突预测项目。该系统于2019年推出,经过不断更新和优化已发展成一个比较完整和成熟的冲突预测项目。基于暴力预警系统在开源和私有软件项目的托管平台(GitHub)上公开的项目代码,暴力预警系统已经有ViEWS 1.0、ViEWS 2.0和ViEWS 3.0三个版本,但ViEWS 3.0并没有进行数据和技术的改动,只是增加了对冲突导致死亡人数的预测。考虑到预测冲突爆发这一议题更为关键,因此本文着重说明前两个版本(ViEWS 1.0和ViEWS 2.0)所采用的系统结构、数据、算法和评估方法。

① Jack A. Goldstone, et al., "A Global Model for Forecasting Political Instability," *American Journal of Political Science*, Vol.54, No.1, 2010, pp.190-208.

② 综合早期危机预测系统的数据会公开一部分,但不是实时公开。值得注意的是,综合早期危机预测系统并非单纯用某些指标计算冲突的风险值,而是更关注国家内部及外部(国家之间)相互作用产生的可能结果。

(一)系统结构

ViEWS 1.0 是暴力预警系统研究团队在 2018 年完成的系统,其中包含两类数据和三种预测方法。① 暴力预警系统的研究团队随后对整个预测框架进行了评估,也对原有主题模型进行了优化,将 ViEWS 1.0 进行较大幅度的修正与改良后于 2020 年上线了新的 ViEWS 2.0。②

虽然暴力预警系统从 ViEWS 1.0 发展到 ViEWS 2.0,但暴力预警系统的大体框架并没有发生太大的变化。该系统包含五个步骤:一是收集并清洗不同数据;二是将数据转化为空间网格—月度和国家—月度供后续研究的计算和预测使用,并创建相关数据集;三是根据场景模型通过领先一步预测(one-step-ahead forecast)或动态模拟的方法预测其结果;四是将各模型的预测结果进行校准,再经过聚合算法计算平均结果和预测爆发冲突的风险值;五是发布结果或根据评估结果进行后续改进。

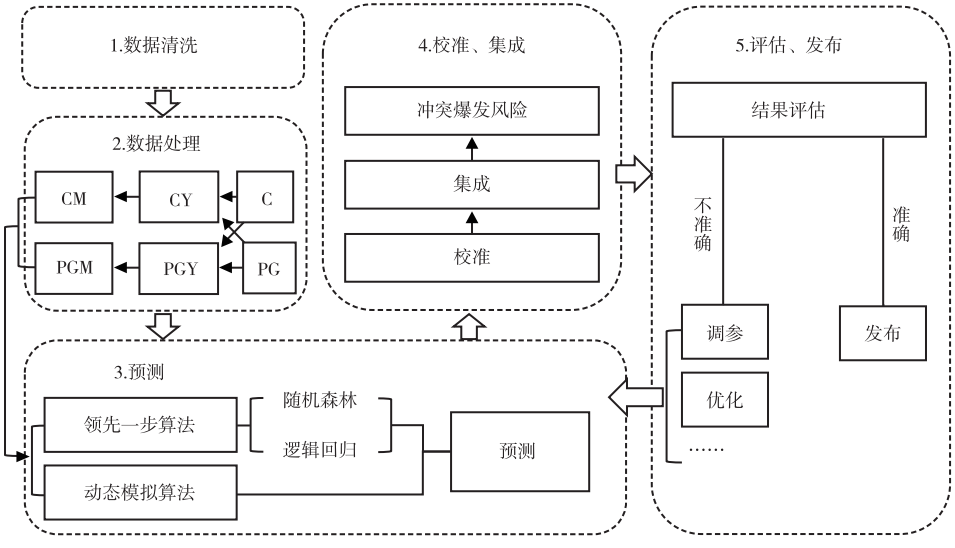

图 1 ViEWS 1.0 预测过程中的五个主要步骤

资料来源:笔者自制。

注:在步骤 2 中,研究团队使用国家与空间网格级别数据转化或合成国家—年度与空间网格—年度级别数据,再转化成国家—月度与空间网格—月度级别数据。

① Havard Hegre, et al., "ViEWS: A Political Violence Early Warning System," *Journal of Peace Research*, Vol.56, No.2, 2019, pp.155-174.

② Havard Hegre, et al., "ViEWS2020: Revising and Evaluating the ViEWS Political Violence Early-Warning System," *Journal of Peace Research*, Vol.58, No.3, 2021, pp.599-611.

(二)数据

暴力预警系统的主要输入数据来自各种和平研究数据库,涵盖了近 50 年来相关国家或次国家地理级别(空间网格点)的地区冲突数据以及具有相同颗粒度的政治、经济、人口、环境和矿产资源数据。暴力预警系统也整合了包括武装冲突地点和事件数据项目(ACLED)和社会冲突数据项目等暴力冲突数据库。此外,暴力预警系统还纳入了世界银行的全球发展指数(TIMG)和全球民主多样性数据集(V-Dem)等相关数据。暴力预警系统研究团队通过对来自不同数据库的数据进行变量组合后形成了多个预测的变量主题,如基准主题、历史冲突主题和经济主题等。暴力预警系统团队随后在空间网格—月度和国家—月度两个级别上构建了两类模型集并为预测非洲地区不同类型的冲突提供了验证与评估工具。

在 ViEWS 1.0 中,国家—月度级别预测包含 6 个专题和 2 个混合(包含空间网格—月度级别数据)数据集。而空间网格—月度级别预测包含 5 个专题、2 个混合数据集和 3 个国家—月度级别数据集。在 ViEWS 2.0 中,国家—月度级别数据包含 24 个模型,空间网格—月度级别数据包含 30 个模型。虽然暴力预警系统将预测的时间颗粒度提升到了月度级别,将空间颗粒度提升到空间网格级别,但由于数据受统计方法的限制,构建的一些分数据集出现了模型高估或低估的情况,这一问题在 ViEWS 1.0中表现得尤为明显,主要出现了两种情况:一是以年度(Y)为时间统计单位的部分数据在处理步骤中,其冲突风险被均匀分散到国内的每个网格单元上,导致部分地区的冲突风险被严重高估。以描述国内经济状况的人均 GDP 指标为例,如果以年为时间颗粒,部分国家—月度级别数据集和空间网格—月度级别数据集会将国家—年度(CY)的数据转化为国家—月度级别数据和空间网格—月度级别数据,即一年中每一个国家—月度级别数据都等于该年测量国家—年度值,空间网格—月度级别数据则是该国每个网格点在每个月的值都等于国家—年度值。这就导致在国家—月度级别数据层面,某年 12 个月的月度人均 GDP 指标为同一个值。而在空间网格—月度数据层面,某国境内所有网格点在 12 个月内的月度人均 GDP 指标也为同一个值。二是网格点(PG)级别数据层面的结构型数据(如“森林面积”“金矿储量”等)通常以年度为单位,部分数据可能在数年内都不会有太大的变化,虽然结构型变量对捕捉长期冲突爆发风险具有较高的可信度,但如果运用这类变量过多反而会影响预测结果的准确性。在 ViEWS 1.0 中被纳入的变量主题多由这类变量构成,进而导致对国家冲突爆发的预测值在一年甚至几年内的变化并不大。

为避免上述两种情况的发生,暴力预警系统的研究团队采取了四点修正措施:

其一,对原有的主题数据集进行删减并新增了几个新的数据主题。在 ViEWS 2.0 中,国家—月度级别数据只包含了 16 个主题(原本是 24 个),空间网格—月度级别数据包含了 12 个主题(原本为 30 个)。在国家—月度级别数据部分,暴力预警系统的研究团队扩大了一些随时间缓慢发生变化的结构性主题数据集,如民主多样性数据集提出的身体暴力指数(the physical violence index)。其二,研究团队纳入了一些可达到月度更新频率的新数据,如武装冲突地点、暴力冲突与游行数据以及国际危机组织(ICG)发布的危机观察报告等数据。其三,为了减少系统出现过分估计爆发风险的情况,暴力预警系统加入了冲突历史主题,从某国上次爆发冲突的时间和烈度等方面预测近期可能爆发的冲突情况。其四,增加了聚合算法,该算法本身能够实现对多个模型预测的结果进行融合,根据多方面因素预测未来冲突爆发的情况。

在空间网格—月度级别数据部分,研究团队调整了原有模型,在保留部分结构性数据的情况下修改或新增了一些主题。其中,部分更新主题被纳入月度级别统计的变量中,如将标准化降水蒸散指数(SPEI)纳入统计。此外,由于冲突爆发这一事件在统计学上出现的概率较低,如果将冲突是否爆发作为因变量,那么相关数据可能会出现严重不平衡。因此,暴力预警系统中也包括了数据处理的模块,如通过下采样(under-sampling)的方式来处理数据不平衡问题。在 ViEWS 2.0 中,该研究团队调整了冲突判定标准的阈值,增加了不同类型冲突爆发的观察数量,变相增加了模型对捕捉冲突爆发因子的敏感程度。总之,ViEWS 2.0 的预测能力较 ViEWS 1.0 有了大幅度的提升。

(三)算法

在具体的预测技术(算法)上,暴力预警系统包含 3 个模块的预测算法,分别是动态模拟(dynasim)模块算法、领先一步预测模块算法和聚合贝叶斯模型平均算法。

1.动态模拟模块算法

动态模拟模块算法是一种基于统计模型的模拟预测方法,其核心是利用冲突爆发的统计模型获得参数估计,然后通过模拟的方式预测未来某个月各国爆发不同类型冲突的概率。赫格等曾用该方法预测了 2010—2050 年的全球武装冲突,还预测了相关国家的国内冲突。① 动态模拟模块算法的优势在于:通过动态多项回归

① Havard Hegre, et al., "Predicting Armed Conflict 2010-2050," *International Studies Quarterly*, Vol.55, No.2, 2013, pp.250-270; Havard Hegre, et al., "Forecasting Civil Conflict Along the Shared Socioeconomic Pathways," *Environmental Research Letters*, 2016, DOI: 10.1088/1748-9326/11/5/054002.

模型可以捕捉冲突爆发前的风险走势,在面临突发状况时还可以通过外部系数人工干预模拟的预测结果。例如,2007 年 4 月,美国第二大次级房贷公司破产,在预测后续月份金融市场走向时可以在模拟结果外乘以一个小于 1 的萎缩系数以表明未来市场的下行趋势。同理可以预测新冠疫情暴发对非洲地区冲突产生的影响,该模拟算法可以在预测的结果上加一个-5%—5%的风险增加值(该值可以根据过去非洲地区传染病对本地区冲突的影响程度而定)。

动态模拟模块算法本质上是一个类机器学习的算法,动态模拟算法包含三个步骤。第一步是通过逻辑回归(logit model)建立自变量与因变量的参数估计,其公式为:

$$P(Y=1 \mid x)=\frac{1}{1+e^{-(W^{T}x+b)}} \qquad \text{式 1}$$

其中,$P(Y=1|x)$代表在给定特征 x 的情况下,模型预测 Y 为阳性(冲突会爆发)的概率 P;而$\frac{1}{1+e^{-(f(x))}}$是逻辑函数,其意义是将线性回归 $f(x)$ 的预测结果映射到 0—1 的概率空间,从而得到一个(冲突爆发)概率的预测结果;而 $W^{T}x+b$ 则是线性回归的表达式,W 是输入特征 x 的权重,b 则是偏置值。第二步是用 t 代表参数估计得到特征对应系数记为 W_t,对其进行拟合并抽样,用以构建下一期(用 $t+1$ 表示)直至未来第 s 期(用 $t+s$ 表示)特征对应系数 $W_{t+1}\cdots W_{t+s}$,再通过对应系数乘以本期特征 x_t 得到对应 $t+s$ 期的预测。第三步是重复第二步 n 次,最后通过平均法得到未来 s 期的预测 P_{t+s}。①

2.领先一步预测模块算法

领先一步预测模块基于经典机器学习算法而搭建。领先一步预测基于 t-s 时刻(这里指 t 时刻向前 s 步即 t-s 时刻)的数据来预测 t 时刻的不同类型冲突爆发概率,s 的范围是[1,36],即 t 时刻前 1—36 个月的数据。

假定预测目标(或称因变量)是 y_t,预测变量为 X_{t-s},领先一步预测时刻 t 的结果 y_t 将基于 $t-1$、$t-2\cdots t-36$ 时刻的冲突爆发情况,其公式为:

① 详细代码可从 ViEWS 团队公布 1.0 系统代码中“ds”文件夹中查询,代码链接为 https://github.com/UppsalaConflictDataProgram/OpenViEWS/tree/master/ds/mn.py,访问时间:2023 年 7 月 8 日。该算法并未在赫格的论文中进行详细介绍。在“ds”文件夹中还有一种模拟方法,与文中图 2 提到的模拟算法在第二步上有所不同,采用的是对残差项拟合抽样,参见 Havard Hegre, et al., “Predicting Armed Conflict 2010-2050,” pp.250-270。

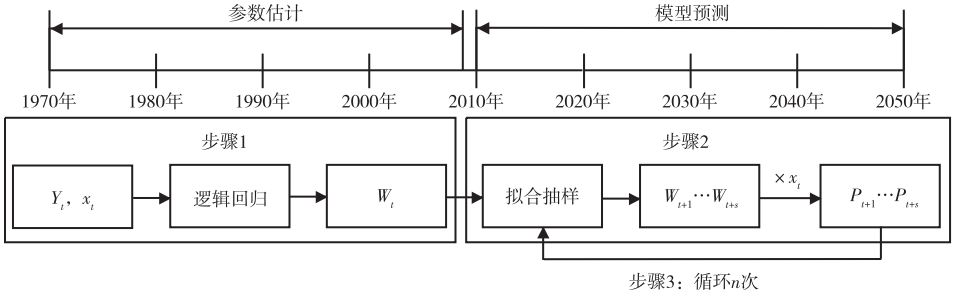

图 2　ViEWS 1.0 的动态模拟流程

资料来源：笔者自制。

$$y_t = f(X_{t-s}), s \in (1,36) \qquad \text{式 2}$$

因此，领先一步预测的数据集实际是将因变量进行 s 个单位的事件移动，造成 X_{t-s} 对应 Y_t 的数据集。对数据集进行训练后便得到能够预测未来 s 个月后的各国冲突爆发风险值。在这一预测模块中，该研究团队主要采用了两种算法：一是随机森林（random forest）算法，二是逻辑回归算法。

此外，在空间网格—月度级别数据部分，暴力预警系统研究团队还采用了极限梯度提升（XGBoost）算法对“全主题”数据集进行预测，并宣称获得了有价值的结果。但根据公布的评估结果，空间网格—月度级别数据在国家冲突层面的精准召回率曲线下面积（AUPR）低于 0.4，因此在现阶段无法应用于政策制定与实施层面。

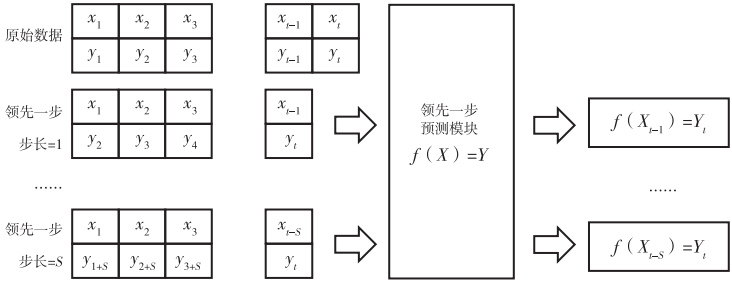

图 3　ViEWS 中的领先一步算法

资料来源：笔者根据 ViEWS 的公开发表、系统、代码自制。参见 Havard Hegre, et al., “ViEWS2020: Revising and Evaluating the ViEWS Political Violence Early-Warning System,” pp. 599-611。

3.聚合贝叶斯模型平均算法

许多基于机器学习的预测都是通过聚合不同预测模块(方法)结果而得出的,暴力预警系统的预测结果也是由多个主题数据集对应不同预测模块(方法)的结果聚合而成。在 ViEWS 1.0 中,实现聚合采用的是最简单的无权重平均(the unweighted averaging)方法。假设共有 K 个主题,$\hat{p}_i^k$ 代表不同主题的预测结果,i 是观测对象(某个国家),$\hat{p}_i^e$ 为最终的聚合结果,相关公式为:

$$\hat{p}_i^e = \frac{\sum_{k=1}^{K} \hat{p}_i^k}{K} \qquad \text{式 3}$$

平均算法在聚合方面并不是最优的选择。无论是传统统计的因果判断还是分析与预测,决定其研究结果准确度的上限是数据。不同的主题只能满足从多个角度挖掘冲突爆发的内因这一要求,但不同主题做出的预测上限是不同的,其模型表现也不一样,平均算法通常无法得到最优解,因此依旧会引发系统出现预测高估或低估的情况。在 ViEWS 2.0 中,针对 t 时刻在领先一步预测结果(以过去 1—36 个月的数据进行预测)实现的结果聚合与动态模拟进行的结果聚合均采用了聚合贝叶斯模型平均算法。聚合贝叶斯模型平均算法给予表现较好的主题分配了较高权重,而对表现较差的主题分配了较低权重,这为我们提供了一个在不同方面进行预测的方法,而不只是选择其中的某个最佳模型。①

聚合贝叶斯模型平均算法将从 K 个模型$\{M_1, M_2 \cdots M_K\}$中选择预测结果最优的模型。其中,每一个模型 M_k 均来自一个先验概率分布 $M_k - \pi(M_k)$,而对于已经爆发的冲突事件 $y^t(t \in T)$ 的似然函数可以表示为 $p(y^t \mid M_k)$,然后通过贝叶斯公式得到的似然函数为:

$$p(M_k \mid y^t) = \frac{p(y^t \mid M_k)\pi(M_k)}{\sum_{k=1}^{K} p(y^t \mid M_k)\pi(M_k)} \qquad \text{式 4}$$

随后,可以得到关于预测$y^{t^*}(t^* \in T)$的边际预测概率函数:

$$p(y^{t^*}) = \sum_{k=1}^{K} p(y^{t^*} \mid M_k) p(M_k \mid y^t) \qquad \text{式 5}$$

① Jacob M. Montgomery, et al., "Improving Predictions Using Ensemble Bayesian Model Averaging," *Political Analysis*, Vol.20, No.3, 2012, pp.271-291.

边际预测概率函数的公式可以理解为根据模型在观测时间段 T 内表现的好坏,即 $p(M_k|y^t)$ 决定该模型在未来预测中权重的分配,其中未来预测记为 $p(y^{t^*}|M_k)$。在 ViEWS 2.0 版,聚合贝叶斯模型平均算法出现在领先一步预测模块和国家—月度级别数据的模型聚合中。具体来说,ViEWS 1.0 的领先一步预测模块只是通过输入 t 时刻前 s 步的数据便能得到某个时间点 t 的冲突爆发风险值。而 ViEWS 2.0 则根据不同前 s 月份数据预测为每个模型 $Model_s$ 分配了一个权重,使每个模型都对最终结果产生了一定的影响。ViEWS 2.0 在后续不同主题数据集建模的预测结果融合中也采用了聚合贝叶斯模型平均算法。这显然比 ViEWS 1.0 无权重平均的做法要更合适。

(四)评估体系

大多数分类学习算法由于在不平衡数据的分类问题上进行预测时会出现较大误差,[①]因此该算法对整体数据的评估标准和准确度是远远不够的,必须加入精准性与召回率来评估少数样本的指标。暴力预警系统对预测结果的评估采用了 5 个指标来衡量,分别为接受者操作特征曲线下面积(AUROC)、准确召回率曲线下面积、布里尔分数(BS)、准确率(ACC)和分数(F1)。

具体来说,AUROC 对受试者工作特征曲线(ROC)覆盖面积进行计算而得,而 ROC 则对真正类率(TPR)与真负类率(TNR)进行绘制而得(TPR 为 y 轴,TNR 为 x 轴)。AUPR 根据精确召回(PR)曲线覆盖面积计算而来,而 PR 曲线通过对精准度(precision)与召回率(recall)进行绘制而得(precision 为 y 轴,recall 为 x 轴),其值越接近 1 说明预测效果越好。布里尔分数反映的是预测的冲突爆发风险值与实际爆发冲突的平均差异,值越小说明其预测越精确;准确率的值越接近 1,说明其预测的准确性越高,由于数据不平衡(冲突未爆发样本数远高于爆发样本数),可能会出现预测未爆发但实际爆发的情况。F1 反映了召回率和精准度两个指标的差异,F1 的值越接近 1 说明上述两指标就越接近,也表明该预测模型在理论上越理想。

需要注意的是,冲突预测往往存在样本不平衡的问题,即阳性样本(冲突爆发)的案例远远小于阴性样本(冲突未爆发)的案例。如果将阴性预测纳入考量,我们通常会得到一个很高的评分,但针对阳性样本的预测才是最重要的。也就是说,预测应该更加关注与阳性样本相关的精准度和召回率。从公式来看,AUROC

① Yanmin Sun, et al., "Classification of Imbalanced Data: A Review," *International Journal of Pattern Recognition*, Vol.23, No.4, 2009, pp.687-719.

和准确率都有将真阴率(TN)纳入公式中,因此这两个指标实际上不适合评估冲突爆发的准确率(见表1)。

表1　评价函数及相关公式

指标	公式	说明
AUROC	$TPR=\frac{TP}{TP+FN}$	在实际阳性的样本中预测阳性正确的比例
	$TNR=\frac{TN}{FP+TN}$	在实际阴性的样本中预测阴性正确的比例
AUPR	$Precision=\frac{TP}{TP+FP}$	在预测阳性的样本中预测正确的比例
	$Recall=\frac{TP}{TP+FN}$	在实际阳性的样本中预测阳性正确的比例
BS	$BS=\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i-A_i)^2$	实际(A_i)与预测($\hat{p}_i$)的平均差异
ACC	$ACC=\frac{TP+TN}{TP+FP+FN+TN}$	在预测阴性和阳性的样本中,预测正确的比例
F1	$F1=2\frac{Precision*Recall}{Precision+Recall}$	精准度与召回率的调和平均,判断两指标接近

资料来源:笔者自制。

注:真阳率(TP)表明实际是阳性样本被预测为阳性样本的样本数。假阴率(FN)表明实际是阳性样本但被预测为阴性样本的样本数。假阳率(FP)表明实际是阴性样本但被预测为阳性样本的样本数。真阴率表明实际是阴性样本被预测为阴性样本的样本数。

从各个指标的计算公式可以看出,暴力预警系统的研究团队已经考虑了冲突爆发与冲突未爆发两者间存在严重的数据不平衡问题,所以他们采用了ROC曲线,将其作为主要的检验预测正确与否的手段并更加关注阳性样本,以此来分析冲突爆发预测的准确性。总体来说,暴力预警系统的评估体系比较全面,可以成为此类预测研究的一项重要参考。

四　对暴力预警系统预测结果的重新评估

暴力预警系统至今已陆续公布了15次冲突预测,通过重新分解这些预测,可以综合评估该系统的预测能力和水平。

(一)暴力预警系统公布评估

暴力预警系统研究团队在 2019 年与 2020 年公布了 ViEWS 1.0 与 ViEWS 2.0,其系统在国家—月度与空间网格—月度这两类数据上的表现可见表 2 和表 3。

表 2 ViEWS 1.0 与 ViEWS 2.0 系统的区间划分

版本	期间	训练集	校准集	测试/预测集
ViEWS 1.0	评估	1990 年 1 月至 2011 年 12 月	2012 年 1 月至 2014 年 12 月	2015 年 1 月至 2017 年 12 月
	预测	1990 年 1 月至 2015 年 8 月	2015 年 9 月至 2018 年 8 月	2018 年 10 月至 2021 年 10 月
ViEWS 2.0	评估	1990 年 1 月至 2012 年 12 月	2013 年 1 月至 2015 年 12 月	2016 年 1 月至 2018 年 12 月
	预测	1990 年 1 月至 2015 年 12 月	2016 年 1 月至 2018 年 12 月	2020 年 1 月至 2022 年 12 月

资料来源:笔者根据 ViEWS 1.0 与 ViEWS 2.0 整理。参见 Havard Hegre, et al., "ViEWS: A Political Violence Early Warning System," pp.155-174; Havard Hegre, et al., "ViEWS 2020: Revising and Evaluating the ViEWS Political Violence Early-Warning System," pp.599-611。

注:"评估"期为样本内数据,主要用以模型评估和参数调整;预测期的预测集为样本外预测。

表 3 ViEWS 1.0 与 ViEWS 2.0 系统的最终聚合模型表现

系统	集成模型	AUROC	AUPR	BS	ACC	F1
ViEWS 1.0	CM	0.9555	0.869	0.0932	0.847	0.745
	PGM	0.9484	0.277	0.0062	0.991	0.289
ViEWS 2.0	CM	/	0.864	0.075	/	/
	PGM	/	0.289	0.045	/	/

资料来源:笔者根据 ViEWS 1.0 与 ViEWS 2.0 整理。参见 Havard Hegre, et al., "ViEWS: A Political Violence Early Warning System," pp.155-174; Havard Hegre, et al., "ViEWS 2020: Revising and Evaluating the ViEWS Political Violence Early-Warning System," pp.599-611。

注:"/"表示研究文献中并未提供相关指标。相关指标分别为:接受者操作特征曲线下面积(AUROC)、准确召回率曲线下面积(AUPR)、布里尔分数(BS)、准确率(ACC)和分数(F1)。

暴力预警系统研究团队声称,他们的预测是相对成功的,认为新的预测在不同数据级别和预测跨度等方面的表现与之前测试的结果一样准确。① 但在实际结果中,暴力预警系统进行的预测是否如该研究团队所说的那样准确还需要做进一步的评估。

① Havard Hegre, et al., "Can We Predict Armed Conflict? How the First 9 Years of Published Forecasts Stand Up to Reality," *International Studies Quarterly*, Vol.65, No.3, 2021, pp.660-668.

(二)评估

由于预测因变量(冲突是否爆发)存在严重的数据不平衡问题,因此本文只关注准确召回率曲线下面积与布里尔分数的指标。暴力预警系统在国家—月度级别数据上的表现相对良好,但无法有效提高预测的精准度。而其在空间网格—月度级别数据上的预测则是完全失败的(见表4)。下文针对该系统历次的预测进行了更为详细的复盘。从2018年6月起(除2019年4月至2021年1月外),该研究团队一直保持着对月度系统的更新与预测。训练集与校准集数据一般会在预测公布之前就发布其所有数据,而预测的区间被限定在自预测公布的上月至未来三年的时间内。例如,当我们评估2018年6月暴力预警系统公布的预测结果时,会将文件中已公布的2018年5月至2021年6月的预测结果与这段时间内非洲地区实际冲突的情况做对比。囿于篇幅,本文只罗列了国家—月度级别数据在综合各个模型和主题后的最终结果。在最终结果聚合模块中,ViEWS 1.0采用了平均计算法,ViEWS 2.0则采用了聚合贝叶斯模型平均算法。

表4　暴力预警系统历届预测结果评估

验证区间	AUROC	Precision	Recall	AUPR	BS	F1
2018年5月至2021年6月	0.9265	0.6207	0.9329	0.7314	0.1662	0.7455
2018年6月至2021年7月	0.9226	0.5929	0.9683	0.739	0.1823	0.7355
2018年7月至2021年9月	0.926	0.5789	0.9666	0.7409	0.1935	0.7241
2018年8月至2021年9月	0.9267	0.6452	0.9685	0.7333	0.1481	0.7745
2018年9月至2021年10月	0.9284	0.7323	0.8643	0.7421	0.1184	0.7928
2018年10月至2021年11月	0.9179	0.714	0.6486	0.6861	0.1594	0.6797
2018年11月至2021年12月	0.9178	0.7431	0.5504	0.6864	0.1672	0.6324
2018年12月至2022年1月	0.9223	0.75	0.4637	0.7053	0.1812	0.5731
2019年1月至2022年2月	0.934	0.7677	0.4667	0.7441	0.177	0.5805
2019年2月至2022年3月	0.9322	0.7651	0.4597	0.7065	0.1788	0.5743
2021年7月至2022年6月	0.9461	0.86	0.5513	0.8265	0.1296	0.6719
2021年8月至2022年6月	0.9421	0.8571	0.5538	0.8232	0.1296	0.6729
2021年9月至2022年6月	0.94	0.8857	0.5962	0.8134	0.1157	0.7126
2021年10月至2022年6月	0.9373	0.88	0.5641	0.8141	0.1235	0.6875
2021年11月至2022年6月	0.955	0.8889	0.6154	0.8383	0.1111	0.7273

资料来源:笔者自制。

注:由于最新的冲突数据只更新至2021年12月,所以自2018年12月以后的预测区间不足三年。相关指标分别为:接受者操作特征曲线下面积(AUROC)、精准度(Precision)、召回率(Recall)、准确召回率曲线下面积(AUPR)、布里尔分数(BS)和分数(F1)。

表 4 显示,暴力预警系统在新冠疫情期间的预测结果并不理想。自 2018 年 11 月开始,召回率直线下跌且保持在一个较低的水平并持续到 2022 年 6 月的最后一次预测。这几次预测区间正好涵盖了全球新冠疫情流行的时期。召回率体现的是在实际爆发冲突的案例中暴力预警系统对冲突预测的正确比例。在 2018 年 11 月之后的几段时间内,暴力预警系统的预测模型一直存在严重的误判情况。从精准率来看,虽然 ViEWS 2.0 相比 ViEWS 1.0 有显著提升,但也只是保持在 76%的水平。也就是说,暴力预警系统的预测仍存在一定的误判情况。上述初步评估表明,暴力预警系统的历次预测结果充其量只能算是合格,不能算是一个理想的预测系统。

下文利用时间窗口动态分析和评估了暴力预警系统历次预测的情况。由于暴力预警系统的预测时间为未来的三年期,本文采用了一个一年期的时间窗口对预测结果做了一个动态评估。如果以 2018 年 9 月至 2021 年 10 月为预测区间,第一轮主要对 2018 年 9 月至 2019 年 9 月的预测进行评估,第二轮则对 2018 年 10 月至 2019 年 10 月的预测进行评估,以此类推到 2020 年 10 月至 2021 年 10 月为止。由于对 2021 年的预测是从 8 月开始,评估区间不足 1 年,因此对其不纳入 2021 年的预测(如图 4、图 5 和图 6)。图 4、图 5 和图 6 的 y 轴对应的是各项指标,而 x 轴对应的是时间。

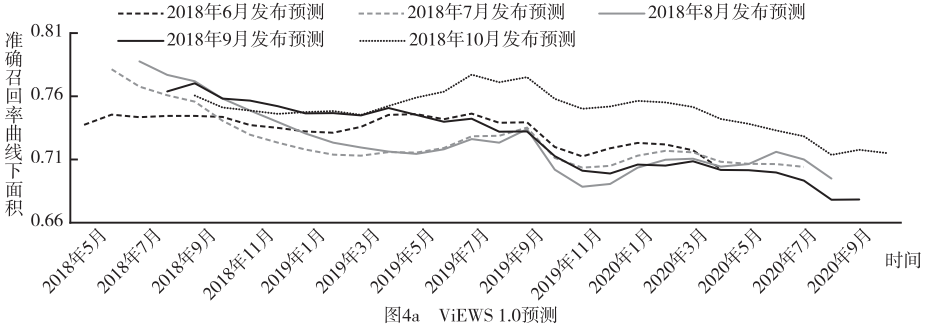

图4a ViEWS 1.0预测

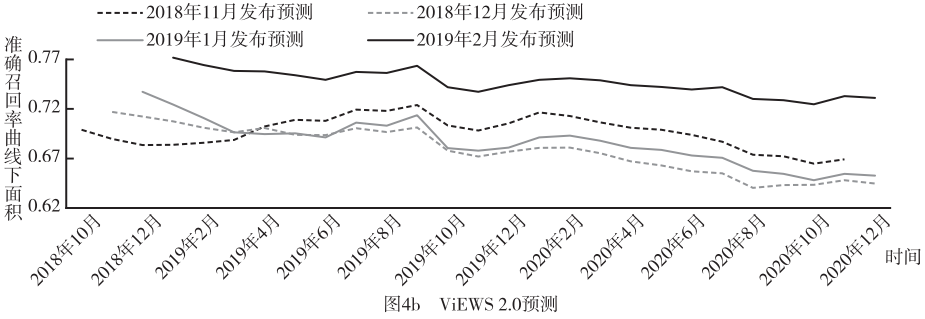

图4b ViEWS 2.0预测

图 4 暴力预警系统历届预测评估中准确召回率曲线下面积

资料来源:笔者自制。

图4显示,预测的时间跨度越大,其误判的概率就越大。暴力预警系统虽然可以实现对未来三年的预测,但从时间窗口的评估来看,其预测跨度越大,AUPR及相关指标的下降越明显。如果预测区间覆盖新冠疫情,召回率指标一般低于0.8(甚至低于0.7)。因此,暴力预警系统预测的可信度并不高,尤其是在面临像新冠疫情这种导致整个环境发生改变的突发事件时,其可信度会大打折扣。另外,对冲突爆发高危地区的高估以及对和平地区的低估问题在暴力预警系统2.0中也未能得到很好解决,这可以从图4、图5和图6的y轴区间值得到验证。其中,图4中从2019年9月开始,其预测的精准率存在明显的下降趋势,主要原因可能是新冠疫情的暴发影响了系统预测的稳定性,但暴力预警系统对此并没有进行说明。

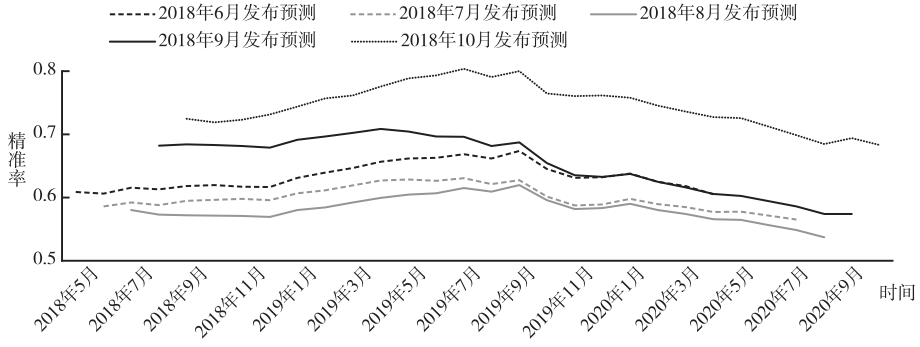

图5a　ViEWS 1.0的预测

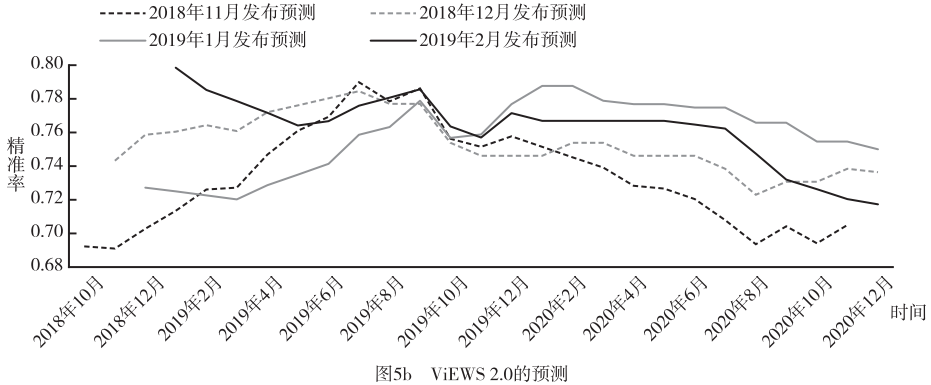

图5b　ViEWS 2.0的预测

图5　暴力预警系统历届预测评估中的精准率

资料来源:笔者自制。

目前,暴力预警系统对冲突的预测结果还不足以能为政策制定提供依据,主要原因有两点:一是缺乏精准度。虽然表4中最后5期预测的精准率均高于0.8,但实际预

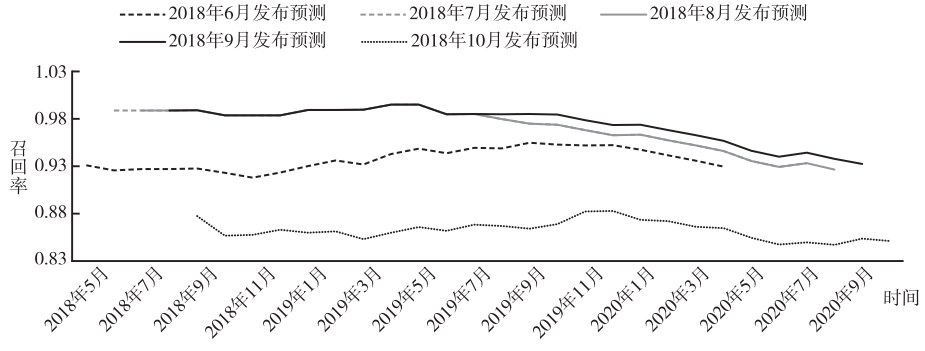

图6a ViEWS 1.0的预测

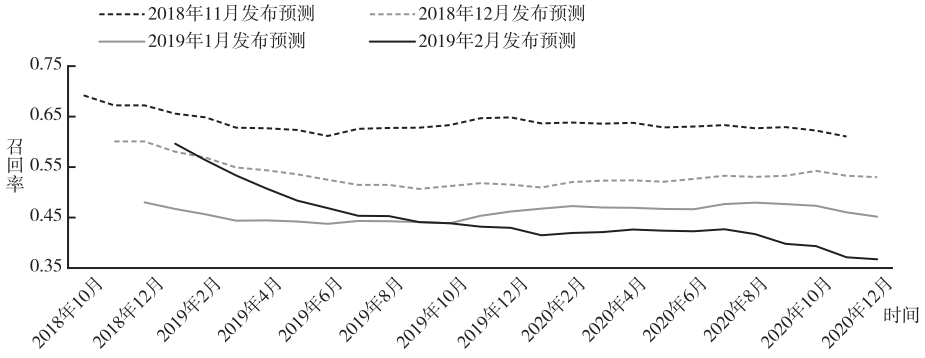

图6b ViEWS 2.0的预测

图6 暴力预警系统历届预测评估中的召回率

资料来源:笔者自制。

测区间不足三年,而之前的预测在召回率和准确召回率曲线下面积在指标的表现上均处于下降趋势且指标值低于0.8,因此相关数据指标不足以支撑决策者的政策制定。准确性过低的冲突预测可能会让决策者制定不符合实际的政策,进而浪费国内资源。二是缺乏稳健性。暴力预警系统缺乏对“黑天鹅”事件进行动态调整和预测的能力。虽然暴力预警系统模拟的是一个开放的过程,可以在运算过程中添加突发事件产生的影响,但暴力预警系统团队并未对新冠疫情在2020—2022年的大流行等“黑天鹅”事件进行动态调整并修正其预测结果。

五 更复杂的机器学习模型会更好吗

虽然ViEWS 1.0采用的预测模型是随机森林和逻辑回归,而ViEWS 2.0的空

间网格—月度级别使用的是极限梯度提升算法,但这些算法在机器学习领域都不是最先进的。由此,研究中存在的一个问题是:使用更加复杂的机器学习算法能否进一步提升暴力预警系统的预测能力?下文将使用更复杂的机器学习模型来检验该系统的预测结果。

(一)合成少数类过采样技术

暴力预警系统对训练样本不平衡的问题采用了下采样方法,本文则采用合成少数类过采样技术(SMOTE)的方法来改善样本不平衡的问题。传统的过采样(OS)方法只是单纯增加少数类样本个数,无法显著提升模型对该类别的识别度,反而会引发过拟合等问题。合成少数类过采样技术虽然也会增加少数类样本个数,但并不是单纯的复制,而是在一个由少数类样本决定的特征空间内进行采样,从而增加少数类样本的数量。① 由于抽样点都在两个临近样本点之间,可以保证样本特征的相似性,因而采样的相似程度在理论上高于传统的采样方法。

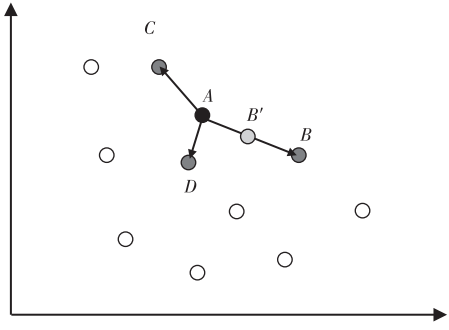

图 7　合成少数类过采样技术抽样法生成新抽样数据

资料来源:笔者自制。

在一个二维特征空间中存在两个不平衡的分类(圆形和菱形),其算法会先扫描每一个样本点,然后计算得到其 k 个最邻近(如图 7 中的 A 点与其最邻近 B、C、D 三点),然后连接样本点 A 与其最邻近点 B,在线段 AB 中生成一个随机点即抽样产生的新样本 B',在 AB 线段上生成一个离点 A 距离为 $\beta|\overrightarrow{BA}|$ 的点 B'(β 为 0 到 1 之间的随机数),其相关数学公式表示如下:

$$\overrightarrow{B'A} = \beta \overrightarrow{BA},\ \beta \sim U(0,1) \qquad \text{式 6}$$

① Nitesh V. Chawla, et al., "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, Vol.16, 2002, pp.321-357.

(二)其他机器学习算法及结果

预测研究还可以用更多的机器学习算法进行新的预测,如在模块中添加神经网络和极限梯度提升等对国家—月度级别的国家冲突进行预测,通过合成少数类过采样技术处理样本不平衡问题。

1.极限梯度提升算法在国家—月度级别数据的运用

尽管暴力预警系统研究团队在空间网格—月度级别数据的冲突预测上也使用了极限梯度提升算法,但他们并未公布国家—月度级别数据的预测结果。因此,本文将极限梯度提升算法运用在国家—月度级别的数据上进行了测试。极限梯度提升法是一种基于梯度提升决策树(GBDT)的算法。① 梯度提升决策树算法本身是以决策树(分类树或者回归树)为基础的算法,在训练过程中会先得到一个预测值,然后得到预测结果与实际结果的残差,再将残差作为下一次的训练目标。简单来说,梯度提升决策树算法的预测结果由训练好的 k 个基模型叠加计算得到,计算公式为:

$$F_m(x) = F_{m-1}(x) + f(x;a_m) \qquad \text{式 7}$$

其中,$F_{m-1}(x)$是前 $m-1$ 棵树的预测结果,$f(x;a_m)$是第 m 个基函数(也可以理解为第 m 棵树),a_m 是训练好的第 m 棵树的参数。极限梯度提升算法在 2016 年被提出,并在梯度提升决策树算法的基础上进行了很多优化,如损失函数上加入正则项、控制模型复杂度实现防止过拟合等。

2.神经网络

神经网络是一种模拟生物神经系统对真实世界做出交互反应的算法模型,②也是最基本的神经元模型,通过设置神经元接收带权重 ω_j 的数据 x_j 后与阈值 u 进行比较,再经过激活函数 θ 处理,最后得到输出 y。其数学公式是:$y=\theta(\sum_{j=1}^{n}\omega_j x_j-u)$。当众多神经元按照一定层次结构连接起来就形成了神经网络,每一层神经元的输出便成为下一层所连接神经元的输入,形成一种嵌套结构。同时,除输入层和输出层外,神经网络可以通过加入单隐藏层或多隐藏层解决非线性问题。

另外,神经网络是一种监督学习模型,因此我们可以通过训练集获得权重和阈值。神经网络训练的核心思想是不断通过正向传播和逆向传播中的梯度下降调整

① Tianqi Chen and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," The 22 nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, DOI:10.1145/2939672.2939785.

② Anil K. Jain, et al., "Artificial Neural Networks: A Tutorial," *Computer*, Vol.29, No.3, 1996, pp.31-44.

权重和阈值,直到代价函数的均方误差(MSE)达到收敛点或局部最小值,其数学公式如下:

$$MSE = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y} - y)^2 \qquad 式8$$

3.神经网络与遗传算法

在神经网络上的叠加遗传算法(GA)是一种流行的机器学习算法。[①] 遗传算法通过模仿自然界"适者生存"的法则,赋予神经网络在循环迭代中自我优化的能力。首先生成一个包含 n 个个体的种群,这里的种群指的是拥有不同参数的神经网络模型(包含神经网络层数、神经元个数、激励函数、优化器和学习率5个参数)。然后从种群中随机或按照固定顺序选取两个个体作为父代(用 s_i 与 s_j 表示)。之后便是交叉步骤,两个父代会生成一个子代,该子代会继承两个父代的特征,进而生成一个新的神经网络模型。该模型参数一部分来自父代 s_i,另一部分来自父代 s_j,正如图8所示的新个体 s_2'。在该步骤中可能会存在变异情况,产生的新个体 s_1' 在某一设定好的概率下发生变异,即某一个或多个参数并不继承于父代,这里 s_1' 的变异发生在神经元个数 d 上,没有继承来自父代的 d_i 与 d_j,而是生成了一个新的 d_m。遗传算法的选择、交叉和变异操作通常都是随机的。在每一代中,评估函数用来确定哪些个体更有可能会被选为下一代的父代。[②] 这就意味着表现好的个体更有可能被选为父代而产生下一代,表现差的个体则可能会被淘汰(如图8)。总而言之,种群演化会根据预测模型的精准度决定个体是否会被保留,而这将推动整个系统朝着提升精准度的方向演化并保留结果较佳的各项参数。

由于样本不平衡问题的存在,因此需要将精准度作为评价函数的指标,将以上的算法模块与不同抽样方法结合,利用暴力预警系统的数据集训练进行预测,得到表5的结果。表5显示,即便使用更加高级的机器学习算法,我们仍然无法得到较为理想的预测结果。尽管深度学习加入遗传算法优化后可以得到100%的召回率,但在整体上仍会出现严重的误判情况,尤其是在准确率问题上出现误判。因此,神经网络与遗传的算法总体上仍然无法显著提升暴力预警系统的预测能力,主要原因有三点。

① Annu Lambora, et al., "Genetic Algorithm-A Literature Review," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, DOI: 10.1109/COMITCon.2019.8862255.

② 对于评估函数,本文采用综合精准率和精准召回率曲线下面积作为指标。

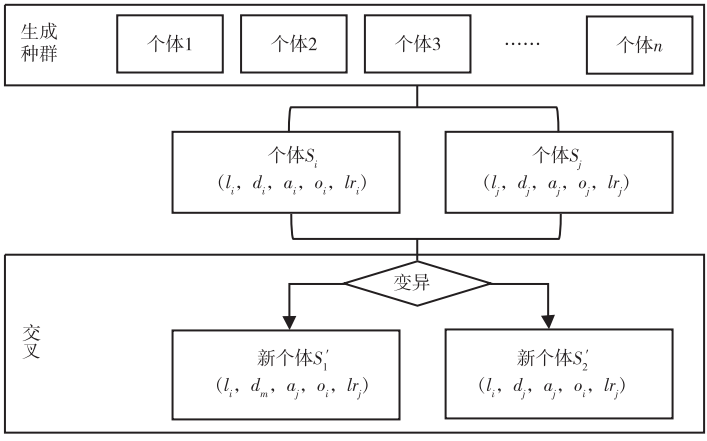

图 8　遗传演化神经网络

资料来源:笔者自制。

注:l 为神经网络层数,d 为某层神经元个数,a 为激励函数,o 为优化器,lr 为学习率。

表 5　加入新的机器学习模块后的预测结果

算法	Precision	Recall	ACC	F1
XGBoost+SMOTE	0.55	0.67	0.86	0.61
ANN+SMOTE	0.19	0.8	0.42	0.31
ANN+OS	0.18	0.89	0.38	0.3
ANN+US	0.19	0.9	0.35	0.32
ANN+GA+SMOTE	0.31	1.0	0.78	0.20
ANN+GA+US	0.16	1.0	0.16	0.27

资料来源:笔者自制。

注:通过机器学习模型与抽样算法,并结合外加遗传算法,共计可以尝试 12 种组合。根据召回率指标从大到小排列(由于无法得到精准率与召回率齐高的组合,因此只采用召回率作为筛选标准),排名第 6 的组合是 XGBoost+SMOTE,其召回率低于 0.8 合格线,因此只列举了前 6 个模型预测的评估结果。

第一,部分算法(如神经网络的结构和激励函数等)得到的结果虽然可以通过调整参数的方式提升预测的准确度,但通过调整参数后得到的模型将无法解决两方面的问题:一是模型的结构与各项参数缺乏可解释性,二是存在过拟合现象。以遗传算法为例,虽然可以扩大种群数量和迭代次数,经过几周训练后可获得预测能力较佳的模型,但这些经过迭代所得到的参数并没有任何规律。尽管测试集表现良好,但在预测时的表现甚至不如以前而出现过拟合的状况。就系统的预测表现来看,暴力预警系统

的研究团队所做的预测可能已经接近机器学习的极限。虽然研究者还可以继续尝试进行更多的建模和参数调节,但完全基于机器学习的预测受诸多技术瓶颈的制约,其算法与预测并未达到预期的结果。①

第二,机器学习的方法比较单一。这意味着在预测过程中需要增加新的方法。有学者利用地区(省份)级别的跨空间暴力冲突数据并结合时变因子对冲突预测取得了良好的效果。② 跨空间变量其实可以套用复杂网络的概念将每个观测区域视为节点,将边视为一种关系交互。对于地区冲突危机的演化也可采用类似模拟的算法,通过模仿历史冲突演化的方式实现预测的目标。

第三,预测窗口的选择不是很合理。预测窗口越小其预测的准确性越高,预测窗口越大其预测的准确性就越低。③ 目前暴力预警系统采用月度作为时间单位,但预测窗口是三年(36 个月),这容易导致系统预测失败,暴力预警系统合理的预测区间可缩至两年(24 个月)甚至更短时间。

六 机器学习的局限性与新的研究方向

(一)机器学习的局限性

随着大数据时代的到来,机器学习逐渐成为冲突预测领域的主要研究工具。暴力预警系统的核心主要是机器学习算法,但该算法不足以支撑后续对冲突预测的研究。系统预测往往会出现失败的情况,其原因主要源于深层次的本体论和认识论问题。

在社会科学研究中,诸多学者对观念、行动和结果进行了研究。有学者认为,三者间存在互动关系,行为体的观念(部分)驱动了行动,而行动在某种社会情境与其他行为体的行动产生互动后就会导致社会结果。④ 厘清它们的定义及相互关系有助于我

① 参见 Samuel Bazzi, et al., "The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia," *The Review of Economics and Statistics*, Vol.104, No.4, 2022, pp.764-779。有趣的是,他们的研究也表明,神经网络算法的效果并不理想。

② Samuel Bazzi, et al., "The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia," pp.764-779.

③ 庞珣:《定量预测的风险来源与处理方法——以"高烈度政治动荡"预测研究项目的再分析为例》,载《国际政治科学》,2017 年第 2 期,第 32 页。

④ 唐世平:《观念、行动和结果:社会科学的客体和任务》,载《世界经济与政治》,2018 年第 5 期,第 33—59 页。关于如何理解系统效应,参见罗伯特·杰维斯著,李少军译:《系统效应:政治与社会生活中的复杂性》,上海人民出版社 2008 年版;唐世平、王凯、杨姗:《理解国际安全战略中的"系统效应"——以中苏同盟破裂的多重影响为例》,载《世界经济与政治》,2013 年第 8 期,第 4—20 页。对系统效应预测能力的探讨,参见庞珣、刘子夜:《基于海量事件数据的中美关系分析——对等反应、政策惯性及第三方因素》,载《世界经济与政治》,2019 年第 5 期,第 53—79 页。

们理解为何以机器学习工具预测冲突会面临诸多困难。冲突的发生是一种社会互动的结果,是行为体的战略行为在一定环境下相互作用而出现的涌现性结果。行为体的观念是行为体实施行为的驱动力,行为则是行为体在某一时刻的自主行为。如果能够比较准确预测各行为体的行为,是否就意味着我们可以准确预测冲突?[①] 遗憾的是,尽管通过预测各方行为体的行为来预测具体的冲突可以取得成功,但该思路要发展成为一个相对自动化的系统却非常困难。机器学习难以准确预测冲突的原因主要有四点。

第一,冲突中的各行为体会考虑不同行为体之间的行为互动以及它们所处的环境,这就需要针对每一个可能的事件进行博弈建模,但构建一个基于机器学习且相对自动化的系统会非常困难。虽然冲突的确是行为体相互作用的结果,但是基于机器学习的大部分冲突预测研究由于其模型架构难以加入对冲突各方内部互动进行分析的环节,在最为常见的面板型数据中无法实现对行为体的互动分析。具体来说,在冲突预测上,机器学习是让机器挖掘各个国家冲突爆发的共有模式,再根据该模式判断未来是否会爆发冲突,这其实是一种“模仿”技术。这种模仿行为仅限于从输入到输出的过程,缺乏对行为体的互动分析。例如,在输入过程中,变量包括了预防冲突的相关数据(如宏观经济、人口和地区资源等),在输出时则机械地提取冲突爆发的风险值。这类技术不能很好地捕捉各行为体之间的互动,也无法反映危机时期不同行为体之间的复杂战略博弈情况。很显然,模仿技术并不关心行为体的运行机制。尽管机器学习是冲突预测的核心技术之一,且预测能力通常要比传统的统计预测方法表现得更好,但机器学习试图以一己之力预测全球冲突事件,其前景并不明朗。

第二,尽管统计分析和机器学习可以将冲突的起因归咎于饥荒和贫穷等因素,但国家的外交策略、国内对各类群体和反政府武装的处理方法以及领导人当时的心态也是导致冲突爆发和升级的重要因素。例如,菲利普·勒斯勒尔(Philip Roessler)对非洲政治进行研究后认为,中央政府与少数族群分享权力是维持社会和平与防止内战的必要条件,但也增强了对手在政变中夺取政权的能力。[②] 这一过程可以视为中央政府与少数民族进行博弈的过程:统治者牺牲多少权力可以换取和平又不使少数民族太过强大而威胁到中央统治;少数民族如何在不产生威胁的前提下争取最大权力。因此,无论是预测国内冲突还是预测国际冲突,各相关团体的相互作用都是不可忽略的重要

① Bruce Bueno de Mesquita, *The War Trap*, New Haven: Yale University Press, 1981, p.223.

② Philip Roessler, *Ethnic Politics and State Power in Africa: The Logic of the Coup-civil War Trap*, Cambridge: Cambridge University Press, 2016.

因素。

第三,冲突爆发并不是单一行动(或一系列行动)的直接产物,而是在某一社会环境下各行为体互动导致的社会涌现性结果。如果仅从行动的角度来解释或者预测冲突,其效用是非常有限的。以古巴导弹危机为例,通过统计分析或机器学习,我们也许能够确定危机一定会发生,但是很难预测危机是否会引发战争。苏联与美国在危机中必定会进行一系列的博弈,其间可能还会爆发某些突发性事件。例如,1962 年 10 月 27 日,一架美国空军 U-2 高空侦察机在古巴城市贝内斯上空被击落;1962 年 10 月 27 日晚上,美国海军对一艘苏联 B-59 潜艇进行深水炸弹攻击,而此时正值古巴导弹危机期间。这些突发事件都有可能影响美苏两国领导人的决策。可以说,战争爆发与否取决于美苏两国领导人是否下决心越过临界点。① 但是,决定战争是否爆发无法基于历史数据来进行推断,因而也无法通过机器学习进行模仿。

第四,即便机器学习确实预测了某次冲突,但很难对解决冲突问题提供有效方案。其原因在于,机器学习给出的预测往往缺乏解释,我们只是知道某个模型的预测似乎更好,但不知道为什么该模型预测得好,而且不知道基于这样的模型在未来的预测能力就一定更强。也就是说,机器学习给出的预测是实用的,而不是科学的或可解释性的。② 这是很多研究者没有意识到的问题,也是机器学习经常被人诟病的一个缺陷。这样的结果意味着机器学习给出的预测并不能告诉我们哪些因素和机制最为关键(类似于必要或充分条件)。因此,如果我们想要预防某次冲突,就必须可以干预这些因素和机制。尽管学界已经意识到这个问题的重要性并呼吁预测要将可解释性和预测能力结合起来,但截至目前这方面的进展非常有限。③

综上,我们无法确切知道要怎样才能实现对重要风险和事件进行精准科学的预测,因此得出的核心结论也是负面的,即机器学习不大可能以一己之力较好地预测重大的政治风险和事件。本文认为,从事这类预测的研究者应秉持一个开放的心态,对重要风险和事件进行预测时要融合多种方法,须结合博弈论、定性比较分析和社会网络分析等方法,基于更加细节的数据将领导人的博弈以及冲突各方交互的环节纳入预

① Chong Chen, et al., "Tipping Points: Challenges in Analyzing International Crisis Escalation," *International Studies Review*, 2022, DOI: 10.1093/isr/viac024.

② Keith Dowding and Charles Miller, "On Prediction Political Sciences," *European Journal of Political Research*, Vol.58, No.3, 2019, pp.1001-1018.

③ Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, Vol.1, No.5, 2019, pp.206-215; Jake M. Hofman, et al., "Integrating Explanation and Prediction in Computational Social Sciences," *Nature*, Vol.595, No.7866, 2021, pp.181-188.

测框架,唯有如此才能得到相对科学的可解释性预测,为预防可能发生的重大风险提供有价值的解决方案。

(二)新的方向:与模拟算法的结合

机器学习如何突破预测的难题?本文认为,机器学习作为一个模仿技术,如果能够与作为模拟技术的行为体建模技术(ABM)结合起来,可能会实现非常有趣的突破,尽管这两种技术的结合也面临着许多挑战。事实上,由于对复杂系统的了解和掌握缺乏足够的细节,人们希望机器学习能够帮助我们做出符合实际的预测。随着理论的进一步拓展以及相关数据的持续完善,从紧张局势发展到冲突爆发这一过程也逐渐变得更加透明,相关机制也更加明确。因此,模拟的方法在今后可能会成为更加有效的预测手段。

行为体建模技术模拟的优势在于其能够演算不同干预手段的效果并评估不同干预手段的作用,为决策者进行人为干预提供便利。此外,基于机器学习建模的预测往往与社会科学理论和既有实证结论并不存在明确的因果关系,因此机器学习的结果不容易被解释。这样一来,即便知道预测结果,我们也无法进行合理干预。如果我们基于预测结果来设计干预方案,就需要其预测具有很强的可解释性。① 由于行为体建模技术在建模过程中会高度依赖社会科学理论和既有实证结果,因此其结果在可解释性和可回溯性方面都比机器学习的表现更好。如果行为体建模技术能够大致预测某些危机或者危机发生的时间,那么我们就可以依据它模拟甄别的因素和机制来设计和实施可能的干预措施。如果行为体要进行建模技术模拟,这就要求我们对系统的理论和实证要有更多的了解,包括必须掌握足够多的细节并准备好消耗大量的算力。相较而言,机器学习在算力消耗及细节上并不如行为体建模技术模拟的要求那样高,如果将机器学习和行为体建模技术结合起来,两者可以实现取长补短的效果。

(三)数据颗粒度的挑战与前景

本文主要讨论了预测某个国家是否会爆发冲突的问题,但这些问题也可能同

① 相关争论参见 Michael D. Ward, "Do We Have Too Much Theory in International Relations, Or Do We Need Less? Waltz Was Wrong, Tetlock Was Right," *Oxford Research Encyclopedia of Politics*, 2017, DOI: 10.1093/acrefore/9780190228637.013.301; Robert A. Blair and Nicholas Sambanis, "Forecasting Civil Wars: Theory and Structure in an Age of Big Data and Machine Learning," *Journal of Conflict Resolution*, Vol.64, No.10, 2020, pp.1885-1915; Robert A. Blair and Nicholas Sambanis, "Is Theory Useful for Conflict Prediction? A Response to Beger, Morgan and Ward," *Journal of Conflict Resolution*, Vol.65, No.7-8, 2021, pp.1427-1453; Andreas Beger, et al., "Reassessing the Role of Theory and Machine Learning in Forecasting Civil Conflict," *Journal of Conflict Resolution*, Vol.65, No.7-8, 2021, pp.1405-1426。

样适用于试图用机器学习来预测某个国家是否会发生冲突的情况。有学者指出，由于预测冲突相当困难，研究者可能需要将预测范围缩小至特定地区。[①] 这种思路在近年来得到了越来越多研究者的认可。例如，塞缪尔·巴齐（Samuel Bazzi）等基于哥伦比亚和印度尼西亚的详尽数据，尝试预测这两个国家的国内冲突。[②] 他们的数据来源于各自国家的本地媒体报道，这可能比全球事件、语言和语调数据库（GDELT）的数据来源更精细可靠。陈冲和胡竞天则采用了不同的方法，基于缅甸的数据关注了冲突的空间依赖性并以此预测冲突的爆发。[③] 这些研究表明，预测特定地区冲突的出现是未来研究的一个方向，但这样的研究也同样面临着许多挑战。

在冲突预测领域，新研究几乎都伴随着新挑战和新发现。巴齐等试验了四种不同的机器学习方法后，得出了颇为有趣的结论。他们发现，在一年的预测窗口内，预测冲突（包括再次发生的冲突）的爆发地点相对容易。然而，精确预测冲突的爆发时间却极具挑战性。值得一提的是，他们的研究还发现，那些基于静态变量模型的预测能力居然比基于动态变量模型的预测能力更强。

近年来，学者们开始探索运用地理网格点数据进行冲突预测，以期实现更精确的冲突定位。然而，根据暴力预警系统团队在空间网格—月度级别数据上预测的实践结果显示，这种颗粒度的预测在准确度上表现较差，其实际应用价值存疑。正如前文所述，空间网格—月度级别数据因地理单元划分和测量问题，难以捕捉其异质性，也在准确度上面临挑战。比如，尽量用夜间光照占比度量网格点的经济数据是一种常见方法，但其存在明显的问题，如无法区分城市地区（包括近郊）的灯光差异，亮度变化也不够清晰。[④] 此外，某些结构型网格点数据难以被准确测量。例如，虽然森林面积和水源占比等可以通过卫星遥感技术获取，但对矿产资源和族群分布等进行精确定位相当困难。如果按照暴力预警系统的建模思路，将国家级别的结构型数据作为境内每个网格点的数值，可能会导致冲突爆发风险被均摊到所有网格点，从而引发严重的高估或低估问题。

对于大部分决策者来说，他们重点关注的通常是一个国家整体的情况，而不是冲

① Lars-Erik Cederman and Nils B. Weidmann, "Predicting Armed Conflict: Time to Adjust Our Expectations?" pp.474-476.

② Samuel Bazzi, et al., "The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia," pp.764-779.

③ 陈冲、胡竞天：《空间依赖与武装冲突预测》，载《国际政治科学》，2022 年第 2 期，第 86—123 页。

④ John Gibson, et al., "Which Night Lights Data Should We Use in Economics and Where?" *Journal of Development Economics*, 2021, DOI: 10.1016/j.jdeveco.2020.102602.

突会发生在这个国家的哪个特定地方。也就是说,预测的颗粒度越小,其政策意义也越小。因此,在冲突预测研究中,我们认为地区、省和城市已经是最小的颗粒度了,除非有更精细的数据,否则无须进一步细化。

七 结论

冲突预测对决策者有显而易见的价值。预测不仅具有实用价值,还是一种验证各种因素和机制的方法,因此有助于我们验证既有理论并为建立新理论奠定基础。① 冲突预测研究已成为学界的"显学"。② 在这一领域,暴力预警系统是其中一个较为透明的研究项目,该系统为其他研究者提供了从数据处理、可视化、建模到评估的一套完整研究体系,也为学界提供了一个公开且便于相互交流的平台。尽管暴力预警系统的预测模型是所有公开系统中表现较好的模型之一,但其总体预测能力并不理想。虽然机器学习是一项核心技术值得学界足够重视,但基于机器学习而对冲突进行预测的结果是不确定的,原因在于依靠机器学习无法充分捕捉引发冲突的行为(逻辑)、行为互动和不断变化的系统环境等动态信息。我们在保持开放心态的同时,不必迷信机器学习一定能够达成精准预测冲突或其他重大事件的目标,而是要运用多种方法将机器模拟技术与行为体建模技术结合起来,争取能在预测冲突或预测其他重大事件上实现突破。

(截稿:2023 年 7 月 责任编辑:赵远良)

① Michael D. Ward, "Can We Predict Politics? Toward What End?" *Journal of Global Security Studies*, Vol.1, No.1, 2016, pp.80-91; Michael D. Ward, et al., "Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction," *International Studies Review*, Vol.15, No.4, 2013, pp.473-490.

② 中国学者在这一领域也做出了诸多努力和探索,参见庞珣:《定量预测的风险来源与处理方法——以"高烈度政治动荡"预测研究项目的再分析为例》,载《国际政治科学》,2017 年第 2 期,第 1—32 页;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 100—117 页;陈冲、庞珣:《非洲恐怖袭击时空规律的大数据分析——基于 GIS 技术和分离总体持续期模型》,载《外交评论》,2020 年第 2 期,第 121—154 页;陈冲、胡竞天:《空间依赖与武装冲突预测》,载《国际政治科学》,2022 年第 2 期,第 86—123 页。