AI 时代社会科学研究方法创新与模型“过度拟合”问题探索

冯帅帅　张佳星　罗教讲

摘要：大数据和机器学习的有效结合推动人工智能获得重大突破，同时也为社会科学开展量化研究方法创新带来新的发展契机。传统理论假设和统计知识驱动的量化研究对模型的过度拟合问题关注不够，导致研究结论的一般化能力受限，更使社会科学研究成果的社会预测功能为人所诟病。而基于交叉验证和正则化方法的机器学习建模方法可能有效解决过度拟合问题，为开展社会预测研究提供方法支撑。本文首先讨论了过度拟合问题的发生根源和内在机制，继而对模型过度拟合问题的机器学习纾解方法进行了介绍，最后分析了机器学习建模方法的不足和限制性因素。将新的机器学习方法运用于社会科学研究，这是机遇而非威胁，研究者需要保持客观态度，努力确保自己有能力根据实际需要将经典方法和新方法的组合应用于具体研究之中。

关键词：AI 机器学习建模 量化研究 过度拟合 预测研究

①作者简介：
冯帅帅，湖南师范大学公共管理学院讲师，博士；张佳星，南京大学社会学院博士研究生；罗教讲，武汉大学社会学院教授、博士生导师。
基金项目：
本文为国家哲学社会科学基金重大项目“大数据时代计算社会科学的产生、现状与发展趋势研究”（16ZDA086）阶段性成果。

Innovation of Social Science Research Methods and Exploration of Model Overfitting in AI Era

Shuaishuai FENG, Jiaxing ZHANG, Jiaojiang LUO

ABSTRACT

The effective combination of big data and machine learning techniques has promoted major breakthroughs in artificial intelligence, and brought new opportunities for the innovation of quantitative research methodology in social sciences. Quantitative research driven by traditional theoretical assumptions and statistical knowledge does not pay enough attention to model overfitting, which limits the generalization ability of research conclusions and ignores prediction in social science research. The machine learning models based on cross-validation and regularization methods may provide a way to address the problem of overfitting and offer methodological support for predictions in social research. This paper introduces the machine learning approaches to address the model overfitting with discussion on the reason and internal mechanism behind as well as its advantages and disadvantages. We argue that implementing machine learning techniques along with classic methods that fit research needs would offer an opportunity rather than a threat to social science researchers.

KEY WORDS

AI; Machine learning models; Quantitative study; Model overfitting; Prediction in social research

一、高科技发展推动社会科学方法创新

社会科学一直在进行研究方法的探索，可以说社会科学的发展与进步，也就是研究方法的创新与突破，具体表现为研究范式的转换，而社会科学研究新范式的出现与科学技术的进步密切相关。早在互联网时代到来之前，社会科学家已经发现了人类社会是一个复杂自适应系统（Complex Adaptive Systems，CAS），但限于当时的技术手段，人们只能把社会这个复杂系统人为地分割为政治、经济、文化等不同的部分，运用还原论的方法来进行研究。互联网、大数据和人工智能时代的到来，为社会科学家进行方法创新提供了难得的机会，计算社会科学（Computational Social Science，CSS）应运而生。随着计算机信息通讯技术的发展及互联网、移动互联网的普及，传统中被社会科学研究者视为“老大难”问题的数据困境已经得到相当程度的缓解，电子踪迹大数据、社交媒体大数据、互联网文本大数据、空间位置信息大数据和传统大规模调查数据等数据集为开展社会科学量化研究提供了坚实的数据基础(郝龙等，2017）。但是，数据不会自动呈现有用信息，需要研究者或源于理论驱动，或源于技术驱动，或源于数据驱动来开采数据的价值。

机器学习（Machine Learning，ML）技术是当前最前沿、最有效的数据价值挖掘工具，“它通过对数据建立抽象表示并基于表示进行建模，然后估计模型的参数，从而从数据中挖掘出对人类有价值的信息”（李德毅，2018）。发展到今天，基于计算机科学和统计学交叉融合的机器学习已经成为更新迭代速度最快的技术领域之一，并跃居当今世界人工智能和数据科学发展的核心位置。关于机器学习的内涵，Jordan 和

Mitchell 做了如下界定：从概念上讲，机器学习算法可以看作是在经验训练的指导下检索大量候选程序，以找到模型最佳拟合效果和优化性能的程序（Jordan et al., 2015）。就社会科学研究方法而言，该技术在建模、选元、分类、聚类等诸多领域拥有一定优势，引入机器学习技术是当前社会科学量化研究方法的重要创新。

有学者指出，大数据和机器学习算法技术将主要从模型建构和变量选择两方面实现对传统社会科学量化方法的优化和升级（黄欣卓，2019）。其一，模型建构方面。传统社会科学研究最常见的一般化流程是：在收集或使用任何数据之前，研究者必须有一个清晰的理论分析框架，然后从中推导出一组可证伪的命题 / 假设，继而通过统计建模证实或证伪命题并最后得出结论（King et al., 1995）。这种标准的演绎方法在评估和修正已有理论工具时特别有效。更重要的是，在传统社会科学研究中，能够直接服务于研究者开展量化分析的数据非常稀缺，因此追求解释能力高的“强模型”才能体现数据的价值。当然，这种建模过程对研究者理论素养的要求通常也比较高。然而，这种标准化的研究程序实际上具有许多潜在弊端，如研究者可能会错失提炼新概念、发展新理论和推导新命题的机会（Grimmer et al., 2021）。大数据时代来临之后，面对大体量、多元化、高维度和内容丰富的全新数据形式，传统“强模型”并不能实现最大化发掘数据深层价值的目标，更适合的做法是充分发挥新技术的算法和算力优势，在数据和理论双向驱动的基础上开发融合多条解释路径的集成模型。计算机科学家经常发现，结合多个不同模型和算法的集成模型，胜过任何单一模型的估计量（Montgomery et al., 2012）。其二，变量选择方面。既往小数据量化分析者主要依靠理论、文献、经验，

乃至灵感来进行变量选择，但随着数据量的增长，传统基于先验知识、专家经验和文献综述的人工选元方法将变得困难重重和效率低下（Wu et al., 2020）。而机器学习算法的奖惩函数，可以帮助我们突破原有选元方法，在海量数据、成千上万个变量中选择控制变量或工具变量，以帮助我们在实证研究中探索和建立更可靠的因果关系。特别是传统不被量化研究者所关注的文本、图片、视频、网络痕迹信息等大数据类型也可以借助机器学习技术进行降维转化，构造和生成具备大视野、大跨度、大历史特性的全新变量和测量指标，提升研究课题的科学性和说服力（陈云松等，2017）。

对于社会科学量化研究来说，除了模型建构和变量选择外，引入机器学习建模方法更突出的作用在于它可以有效应对传统基于普通最小二乘法（ordinary least square，OLS）统计建模所带来的模型过度拟合及社会预测困难问题。

二、量化模型过度拟合与预测难题

（一）从解释到预测：定量社会科学研究的象限划分

在主流社会科学和计算机科学领域存在两种截然不同的价值观念，社会科学家优先考虑因果解释，经常援引各式成熟理论作为因果机制尝试对个体层面和集体层面的人类行为作出令人满意的解释。而计算机科学家更关心开发准确的预测模型，无论它们是否符合因果机制，甚至是否可以解释（Hofman et al., 2021）。反过来，这种价值观差异导致社会科学家和计算机科学家在研究过程的侧重点上存在明显区分。社会科学

研究中的定量方法一般被用来识别变量间的因果关系或对具备理论价值的参数进行无偏估计，而计算机建模方法则以缩小模型泛化误差为终极目标。因此，社会科学家通常在样本数据中拟合他们的模型，因为他们关注的是解释而非预测，但计算机科学家则尤为重视模型在样本外数据的拟合和预测能力（Shmueli et al., 2010）。

从理想类型的角度出发，可以将社会科学量化研究建模的任务依据时空和干预两大维度划归到以下四个象限空间中，顺时针顺序依次为描述建模、解释建模、推断建模和预测建模。如表 1 内容所示。

表 1 定量社会科学研究建模功能象限划分

时空 / 干预维度	无干预意向	有干预意向
截面时空	象限 1：描述建模	象限 2：解释建模
未来 / 平行时空	象限 4：预测建模	象限 3：推断建模

描述建模（象限 1）主要侧重对截面数据样本统计特征（包括平均值、中位数、众数等集中趋势和标准差、方差、极差等离散趋势）的基本描述和总结，通常不涉及因果关系探讨。这类研究在政府部门统计报告和早期量化研究领域中出现较多。解释建模（象限 2）的目标是基于一定统计方法来识别和估计变量间的因果效应，为研究者的现象解释提供数据支撑。虽然解释建模的生成过程也通常源自固定时空数据，但对因果机制的探索为研究者提供了实施干预的可能，即假如发现了某输入变量对输出变量的显著影响，我们就可以通过调整自变量水平来使因变量提升或者降低到一定状态。随着各类高级计量方法和实验（及准实验）方法的引入，定量社会科学研究逐渐从解释建模过渡到更为高级的推断

建模（象限 3）形态，其基本问题可以概括为：如果我们干预了世界的某些特征，结果会有什么不同？推断建模源自平行时空理念的反事实分析思路，即通过可供观观测的特定时空数据，探讨输入变量的变化是否和如何因果性地改变输出结果。迄今为止主流的量化社会科学建模主要集中这三类研究中。近来在大数据和机器学习技术的推动下，传统为社会科学研究者所忽视的预测建模（象限 4）重又开始兴起（陈云松等，2020）。预测性建模指的是研究者建模的初衷主要在于预测结果是否发生以及在何时发生，而对变量间的因果关系不过分关注（Hofman et al., 2021）。预测建模的评估标准是，基于训练集数据建构的单一或集合模型在其他数据集（other sample）和未来时空（other time）数据集中的拟合效果。

直观上来看可能会认为推断建模和预测建模具有重合的地方，或者将推断问题视作预测问题的一个子类，但事实上两者有着本质的区分（Grimmer et al., 2021）：第一，根本目标不同。预测的目标是从不同可观察样本和特征变量中预测一个结果值的单位变化，但对于因果推断，我们感兴趣的是在世界的不同反事实状态下结果会作何不同。推断问题需要一个定义明确的干预变量(或行为)来解释世界在不同状态下的结果，而预测问题则不需要这种界定；第二，评估标准不同。对推断模型的效果评估非常困难，因为我们永远不可能实际观察到对计算因果效应至关重要的反事实数据。而在预测问题中，真相（预测成功或者失败）会在某个未来时刻或者某个其他数据集中被揭示；第三，解释变量的重要性不同。在预测问题中，解释变量仅仅是用来对还未观察到的结果进行预测的工具，不会刻意地人工区分所谓的核心变量和协助变量（次等变量），

只要对模型的预测能力有帮助即可纳入。但在推断建模中我们一般会有明确区分，使用核心变量作为干预手段，而协变量的使用只是为了帮助研究者获得对因果关系更加准确和无偏的估计。

（二）过度拟合问题：讨论与反思

早在 20 世纪 40 年代，Kaplan 就提出要加强社会科学中的预测，他认为社会行为较之微观尺度上的自然现象更具备可预测性，“人类之所以和原子或者分子不同，在一定程度上表现为人类行为可以被人为制造的规则所预测”（Kaplan，1940）。然而，百年来社会科学研究者们却并未在社会预测方面获得实质性进展。甚至以实证科学主义自居的量化研究者过多地将精力放在描述数据和证伪理论上，普遍不擅长（或者说是无法实现）社会预测（陈云松等，2020），导致传统理论驱动的实证社会科学研究经常因可复制性差、泛化能力弱、预测准确性低和无法为现实问题提供解决方案等问题而为人所诟病（Ward et al., 2010; Watts, 2017; Yarkoni et al., 2017）。实质上，量化研究存在上述问题的一个关键因素就在于传统方法无法很好地解决过度拟合问题（Babyak, 2014; McNeish, 2015）。

2016 年 9 月，Lever、Krzywinski 和 Altman 三人在 Nature 杂志上联合发表了一篇题为《模型选择与过度拟合》（Model selection and overfitting）的文章，专门讨论科学研究中的过度拟合问题。作者提出，可以参考模型在样本上的输出值与真实值之差（偏差 Bias）以及模型每一次输出结果与模型输出期望值之差（方差 Variance）来评估模型的总体拟合优度。两类偏误的数值直接影响模型总误差值（Error）的大小（Error

= Bias + Variance）（如图 1a 所示）。偏差和方差的大小皆与模型复杂度直接相关，太简单的模型可能具有高偏差和低方差（欠拟合），相反，过于复杂的模型通常具有低偏差和高方差（过度拟合）。过度拟合和欠拟合是回归和分类中的常见问题。如图 1b 所示，根据观测数据集的散点图分布可以有线性拟合、二项式拟合和多项式拟合等多种思路。其中，多项式方案（虚线）在观察数据集中的拟合效果最佳，但模型参数受到极端值噪声的严重影响，存在过度拟合风险。相比之下，线性方案最简单，但对观测样本中的拟合效果较差，存在欠拟合风险，也难以提升模型总体的泛化能力。如果我们的目标是减少总误差值，那么可以选择复杂程度介于线性方案和多项式方案之间的二项式模型。该情况也可以应用于分类问题上，如图 1c 所示，复杂的决策边界可能会完美地将训练集中的各个类分开，但由于它受噪声的影响很大，经常会对新情况产生错误归类。在回归和分类问题中，过度拟合的模型在观测数据上可能表现良好，但对于新数据集而言可能表现非常差，与预期目标背道而驰（Lever et al., 2016）。

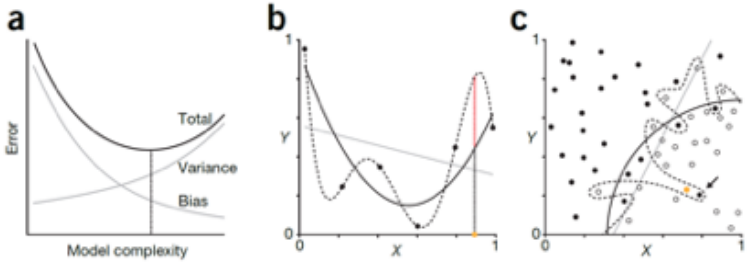

图 1 回归和分类中的过度拟合问题（Lever et al., 2016）

以社会科学领域经典的 OLS 回归模型为例。其基础公式为：

$$Y_i = \beta_0 + \beta_1 X_{1a} + \beta_2 X_{2a} + \cdots + \beta_i X_{ia} + \varepsilon_i \quad (1)$$

式中，β_0 为截距项（也称常数项），β_i 为第 i 个预测变量的回归系数，Y_i 为第 i 项被试在目标变量上的观测值，X_{ia} 为第 i 项被试在第 a 个预测变量上的观测值，ε_i 为残差项。OLS 模型是量化研究者最常使用的回归系数估计方法，能够通过最小化预测值与观测值之间的误差来估计回归模型中的参数，并针对观测样本提供线性无偏估计（McNeish, 2015）。但是越来越多的研究者发现，由于内生性问题的存在，OLS 方法对回归系数的估计值实际上是有偏的（陈云松等，2010；胡安宁，2012），并且非常容易导致模型发生过度拟合和泛化能力弱问题（Hawkins, 2004），即基于观测样本拟合得到的回归模型在应用于同一总体的其他样本集或预测未来数据时拟合效果较差。OLS 方法本质上是通过控制偏差项数值来调整模型总体误差。但随着模型中预测变量的增多，模型的方差项数值会因此而增大，导致过度拟合现象发生，“过拟合现象会导致模型在高估回归系数的同时低估其标准误，容易导致模型中部分无关联的冗余变量被发现存在显著的预测作用，模型得到的结果可能仅适用于当前样本而无法推广到总体”（张沥今等，2020）。

定量社会科学长期专注于建立和测试统计模型，却往往忽略了模型选择的两个重要标准：一是模型应该能够很好地预测样本，而不仅仅是用来拟合建立模型的观测数据；二是应该力求简洁，试图建立简单的模型，以少解释多（Hindman, 2015）。几十年来，社会科学家压倒性地使用 OLS 回归及其衍生方法（如 Logistic 回归、Probit 回归等）来实现这些目标。

但这些模型的预测性能都普遍不佳，原因很简单，有几十个预测变量的回归模型可能拟合度较高（调整 R 方高），但它们既无法捕捉变量之间的复杂机制，又因过度拟合而缺乏稳定应用于其他数据集的能力。有学者通过对近十年（2010-2019）发表于《社会学研究》杂志上 149 篇定量研究论文的分析发现，量化研究者对控制变量的使用呈现出某种“滥用化”趋势。在这些定量研究中，模型控制变量数量在 1-3 个的占比 13.64%；在 4-6 个的占比 47.73%；在 7-9 个的占比 29.55%；10-12 个的占比 6.82%；13 个及以上的占比 2.27%。甚至其中一篇研究的控制变量数量竟高达 21 个（冯帅帅等，2021）。不少量化研究者总是担心自己因无法穷尽所有外生变量（事实上也不可能做到）而得到错误的模型和结论（Antonakis et al., 2010），于是陷入一种“过度控制”（over-control）的失范路径中。而正如我们上文提到的，随着预测变量数量的增加，模型的方差也在同时提升，导致过度拟合问题发生。我们认为，无论从具体方法角度还是因果推断角度考虑，过度拟合问题都是量化研究者绝对不容忽视的一类问题。然而遗憾的是，国内社会科学界对过度拟合问题的关注度并不高。以“过拟合”和“过度拟合”为“主题”设置在中国知网数据库进行检索发现，中国知网学术期刊数据库中合计有 377 篇文献在论文中涉及过度拟合问题，但是其中仅有 1 篇是人文社科领域学者所做的研究。

比较而言，西方学者较早关注到了过度拟合问题且提供了一些应对措施，如 Babyak 给出了回归建模避免过度拟合的策略，包括：收集更多的数据；通过合并等方式减少模型中的预测变量数量；以及最重要的一个建议——在模型中增加收缩项和惩罚项（shrinkage and penalization）。作者认为，通过增加数据体量和减少预测变量数量的方式依旧不够可靠，最终

得到的模型仍然可能过于乐观，而新出现的收缩和惩罚技术才是应对模型过度拟合问题的最佳路径。所谓收缩和惩罚项，即通过算法和统计知识相结合，在模型中添加一个新的组成项，该项是对回归模型拟合新数据集的程度估计。此外，Babyak 还做出大胆预测，“未来几年内，包含收缩项和惩罚项的模型将成为众多统计软件包的标准分析模型。先进的统计知识和计算机技术将带来更多、更复杂的惩罚算法”（Babyak, 2004）。

近来在“人工智能热”的外部背景下，机器学习技术也因其在选元、建模、聚类、大规模计算分析、高维数据处理等方面的独特优势而逐渐进入社会科学研究者视野（Molina et al., 2019）。并且，业已有不少学者发现，除了上述优势外，机器学习方法还是应对量化模型过度拟合问题的有效方法：一方面，相较于传统 OLS 回归建模缩减偏差引入方差的思路，基于正则化（regularization）方法的机器学习建模技术则通过谋求方差和偏差的平衡来削减模型整体误差，从而解决过度拟合问题，提升模型的预测稳定性和精准度（Athey et al., 2016）；另一方面，机器学习的交叉验证（cross validation）逻辑更是公认的克服模型过度拟合问题的行之有效的方法路径（Lever et al., 2016）。

三、机器学习建模方法助力“过度拟合”问题求解

（一）机器学习建模思路：交叉验证

严格意义上讲，机器学习建模更多地体现为一种建模思路和方法体系，其旨在借助一定程序措施来降低模型的“泛化误差”，避免过度拟合问题，提升模型的泛化能力。交叉验证方法是机器学习建模和参数估

计的基本思路和常用方法。所谓交叉验证，顾名思义，就是重复地使用数据。将所获得的样本数据进行标准切分，区分出训练集和测试集（training and test sets）（为了保证模型训练效果，一般为二八比例或三七比例，即 80% 用于训练集和 20% 用于测试集或 70% 用于训练集和 30% 用于测试集）[①]，在此基础上反复地进行训练和测试。一般步骤为：首先用训练集来训练模型，通过调整模型在数据集上的误差不断迭代训练模型，得到对数据集拟合效果良好的模型，然后再用测试集来评估模型优劣（拟合度、预测精准度等）。在此基础上可以得到多组不同的训练集和测试集，某次训练集中的某样本在下次可能成为测试集中的样本，此即所谓“交叉”。测试集数据至少需要满足以下两个条件：第一，规模足够大，可产生具有统计意义结果；第二，能代表整个数据集，即测试集与训练集需要具有相同的特征分布。

依据不同的测试集切分方案，交叉验证主要有以下三类具体方法：第一种是简单交叉验证（两分法）。首先，随机的将样本数据分为两部分（比如：70% 的训练集，30% 的测试集），然后用训练集来训练模型，在测试集上验证模型及参数。接着，再把样本打乱，重新选择训练集和测试集，继续训练数据和检验模型。最后，选择损失函数评估最优的模型和参数。第二种是 K 折交叉验证（K-folder cross validation）。和第一种方法不同，K 折交叉验证会把样本数据随机的分成 K 份，每次随机的选择 K-1 份作

①如果给定的样本数据充足，更好地方式是将数据集拆分成三部分：训练集、验证集（validation set）和测试集。训练集用来训练模型，验证集用于模型的选择，而测试集用于最终对建构模型的评估。

为训练集，剩下的 1 份做测试集。当这一轮完成后，重新随机选择 K-1 份来训练数据。若干轮（小于 K）之后，选择损失函数评估最优的模型和参数。第三种是留一交叉验证（leave-one-out cross validation，简称“留一法”），它是第二种情况的特例，此时 K 等于样本数 N，这样对于 N 个样本，每次选择 N-1 个样本来训练数据，留一个样本来验证模型预测的好坏。此方法主要用于样本量非常少的情况（比如 N 小于 50）。在具体研究中，可以根据样本量（数据）体量可以选择不同的交叉验证方法。

（二）机器学习建模目标：最小化损失

回归或分类建模是机器学习的核心任务。所谓机器建模，简单来说就是将一个机器学习问题转化为数学问题。以常见的“教育 - 收入”话题为例。比如我们通过问卷调查收集了一组关于教育程度（X）和收入状况（Y）的数据 {（x_1，y_1）,（x_2，y_2），…，（x_n，y_n）}。通过文献回顾发现，个体的教育程度与其收入状况存在某种线性关联，其线性公式为：$Y = \beta_0 + \beta X$，其中 β_0 和 β 待定。我们所假设的这个线性公式就称为机器学习的回归建模。而通过数据训练确定待定系数 β_0 和 β 具体数值的过程就是该模型的求解过程。由于有 β_0 和 β 两个未知参数，我们可以借助一次函数的求解方法，简单地通过两组数据来求解系数。但这种方法导致的结果就是，求解得出的 β_0 和 β 两个系数只能很好地解释这两组数据，但对样本总体的其他数据并不适用。显然一次函数的解法并不适宜。因此我们需要找到一组 β_0 和 β，对收集的整个数据集都具有较好的拟合能力。基于此，我们可以定义一个一般化的损失函数来界定模型的拟合偏差值（也是 OLS 回归模型的损失函数）：

$$\text{loss}(\beta)=\sum_1^n(y_i-\beta x_i-\beta_0)^2 \qquad (2)$$

这个目标函数刻画了使用 β_0 和 β 两个系数得到的预测值与真实值之间的距离。机器学习的目标就是让损失函数 loss(β) 越来越小，这就是机器学习的模型优化过程，也被称为训练过程。而一旦我们通过数据训练得出了 β_0 和 β 两系数的值，我们就可以对给定的教育程度 X 进行预测，得到对应的收入状况 Y，这个过程即模型的测试过程。当给定一个损失函数时，我们希望获得一组参数（β_0 和 β）使得该损失函数值达到最小，这叫做机器学习建模的最优化问题（mathematical optimization）。但受训练数据体量的限制——无法穷尽所有数据，我们未必总能找到这样一组函数值来使模型 loss(β) 函数达到最低，也就是说，全局最优问题（global optimum）很难实现，于是寻求局部最优（local optimum）成为机器学习建模的权益选择。两者关系如图 2 所示。

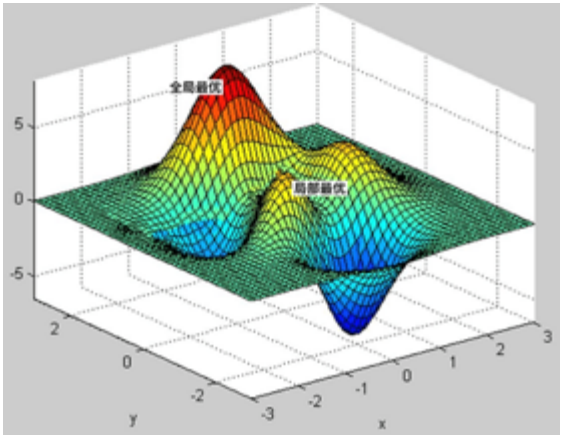

图 2 机器学习建模最优问题

假如存在一组特定参数 $\beta^* = (\beta_1^*, \beta_2^*, \cdots, \beta_n^*)$，对于任意 β 都满足公式（3），则称该参数 β^* 为全局最优值；如果存在一个 $\alpha > 0$，使得所有满足 $|\beta - \beta^*| < \alpha$ 的 β 都满足公式（2），则称该参数 β^* 为局部最优值。局部最优解虽然不一定是全局最优解，但确实是一定空间区域内所有解中的一个解，并且该解的质量一定是比较好的。当现实情况过于复杂，处理的信息量过多时，选择局部最优来代替全局最优是机器学习建模过程中最常采取的策略（丁圣勇等，2018）。尽管上述案例看起来简单，但大部分机器学习算法的核心都是在将问题表达为一个合适的数学公式，然后以数学知识去优化损失函数，从而获得模型的参数值和预测新的数据集。

$$\text{loss}(\beta^*) \leq \text{loss}(\beta) \qquad (3)$$

（三）机器学习建模方法：监督与无监督

经历了几十年的发展，机器学习技术已经形成了比较完整的方法体系。依据其数据集是否已经给出目标特征标签可以将其划分为监督学习（supervised learning）、无监督学习（unsupervised learning）和弱监督学习（weakly supervised learning）三类。每种方法均有自己擅长的领域，监督学习算法擅长分类和回归预测，无监督学习在聚类功能上独树一帜，而弱监督学习则对上述三种功能均有涉猎。其中，监督学习算法是社会科学研究者开展社会预测研究的适宜方法（李航，2012）。

监督学习是机器学习技术中最重要的一类方法，占据了目前机器学习算法的绝大部分。监督学习是指模型训练集的数据是带有“标注”的特征数据，学习算法在进行数据分析时，会利用特征数据进行分析和

建模，在模型训练完成后，再藉由测试集数据对模型进行评估。简单来说，我们在开始训练前就已经知道了输入 Xi 和输出 Yi，监督学习算法的任务是建立起一个将输入准确映射到输出的拟合模型，当给模型输入新值时就能预测出对应的输出结果。由于训练数据中带有标签，监督学习建模相对容易，复杂度低，因而模型拟合效果也往往更好。仍以上文提及的“教育 - 收入”问题为例。假设共计收集到 10000 个样本，每个样本包含 5 个特征值（性别、户口、教育、父辈教育、收入），通过人工标注的方式将性别、户口、教育、父辈教育分别标注为 x1-x4，将收入标注为 y。随机选取其中 7000 个样本为训练集，其余 3000 个样本为测试集。以训练集样本开始训练预测模型，通过一定的方法（如正则化），监督学习算法会自动剔除冗余变量（如剔除 x1 性别变量），筛选出最具预测价值的变量来建构模型（如式中保留 x2-x4 三个变量），然后以测试集样本对模型进行评估，最终得到预测效果最佳（预测准确率最高）的模型。常见的监督学习算法包括正则化回归（regularized regression）、支持向量机（support vector machines）、K- 近邻算法（K-nearest neighbors，KNN）、决策树（decision tree）、随机森林（random forest）等。

同建立在人工标注数据基础上的监督学习算法不同，无监督学习算法通过模型不断地自我认知、自我巩固，最后进行自我归纳来实现其学习过程。由于训练样本的无标注特征，无监督学习输出模型的准确性较难实现评估，可能在实用性上弱于监督学习。但这种独特的方法论可以为社会科学量化研究带来很多启发和灵感，比如可以通过无监督学习算法对传统方法无法处理的高维数据（如文本、图像、音频、视频等非结

构化数据）进行低损耗的降维转换，这个过程不需要使用任何标注。该功能可以拓展传统社会科学研究意义上的实证数据范畴。再比如，我们可以尝试将靠的近的数据归为一类，将离得远的放在不同的类，这样就可以自动挖掘出不同类别，进而帮助研究者发现出一些规律，这就是无监督学的聚类功能。近来无监督学习的价值逐渐得到学术界的重点关注。2015 年，有“深度学习三巨头”之称的 LeCun，Bengio 和 Hinton 在 Nature 杂志撰文指出，无监督学习的建模过程与人类认知世界的过程相类似，“人类和动物的学习过程在很大程度上就是无监督学习的过程：我们是通过观察，而不是通过被告知每个物体的名称来发现世界的结构”，因而“从长远来看，无监督学习将变得越来越重要”（Lecun et al., 2015）。常用的无监督学习方法如聚类算法（clustering algorithm）、网络社区发现（community detection）、潜在语义分析（latent semantic analysis）等。

由于数据标注本身需要很高的成本，因此监督学习算法在很多任务上很难获得全部真值标签这样比较强的监督信息——特别是处理大数据样本时；而无监督学习因为缺失人工标注，在实际应用中的性能往往存在较大局限。针对该两项问题，弱监督学习的概念被提出来。弱监督学习算法不仅可以降低人工标注的工作量和成本，同时也可以引入类人化的监督学习机制，在很大程度上提高无监督学习的性能（Zhou, 2018）。弱监督学习的“弱”是相对于监督学习来说的，同后者不同，弱监督学习的训练数据只有一部分是有人工标注的，其余甚至绝大部分数据都是未标注的原始数据。置言之，弱监督学习的模型训练是间接的，机器学习的信号不会直接指定给模型，而是通过一些必要的引导信息间

接传递给机器学习模型。代表性的弱监督学习算法有半监督学习（semi-supervised learning）、迁移学习（transfer learning）、强化学习（reinforcement learning）等。2016 年，由 DeepMind 开发的“阿尔法狗”程序利用强化学习算法以 4:1 击败世界围棋冠军李世石，强化学习算法得到学界和工业界的青睐。如今，强化学习算法已经在网络游戏、自动驾驶、算法推荐、机器人等多个领域开花结果，谷歌、Facebook、百度、微软等各大科技公司更是将强化学习技术作为其重点发展的技术之一（李德毅，2018）。

（四）监督学习应用举例：正则化回归

上文提到，机器学习的正则化方法可以有效应对传统量化研究中的过度拟合问题，而正则化回归正是监督学习的代表性算法。在理论驱动的传统多元回归建模中，常常存在一个或多个预测变量与目标变量不存在线性关系的情况，违背了奥卡姆的“非必要，不添加”原则，致使模型过度敏感。而对奥卡姆剃刀定律（Occam's Razor）的践行正是机器学习正则化式的重要内驱力。所谓正则化式，就是在典型 OLS 损失函数中添加一个惩罚项（penalty term），目的在于借助算法来剔除式中的无关变量，最终得出一个平均误差和模型复杂度同时较小的模型。其基本原理如下：惩罚项一般是模型复杂度的单调递增函数，而损失函数负责最小化误差，损失函数与惩罚函数呈现为一种张力关系：损失函数越小，模型越复杂，惩罚项的值越大。要使惩罚项也很小，那么模型的复杂程度则必然受到限制，因此就能有效地防止过度拟合现象发生。正则化的具体公式可以表示为：

$$L(\beta)=\sum_1^n(y_i-\beta x_i-\beta_0)^2+\lambda P(\beta)=loss(\beta)+\lambda P(\beta) \quad (4)$$

式中，$L(\beta)$ 是添加惩罚项后的模型损失函数，loss(β) 为一般 OLS 的损失函数，λ 为惩罚项系数（表示惩罚的力度），λ 值越大惩罚的力度越强，约束越紧。特别需要说明的是，当 λ 为 0 时，惩罚项为 0，$L(\beta)$ 等价于一般 OLS 的损失函数 loss(β)。$P(\beta)$ 为惩罚函数，惩罚函数的不同分别对应不同的正则化方法。经常使用的惩罚函数有两种：$L1$ 范数 $\sum_1^n|\beta_i|$ 和 $L2$ 范数 $\sum_1^n\beta_i^2$，分别对应 Lasso 回归（least absoulute shrinkage and selection operator）和岭回归（ridge regression）。岭回归以回归系数的平方和为惩罚函数，可以实现将系数向 0 的方向进行压缩，但是不会把任何一个变量的系数确切地压缩到 0，且容易造成对重要系数的过度压缩（Hesterberg et al., 2008），这在一定程度上限制了岭回归的使用空间。相较而言，Lasso 回归方法可以有效克服岭回归的上述缺点，因此其应用范围更为广泛。Lasso 回归的本质是稀疏性建模（sparsity），它通过惩罚函数让很多自变量的系数转化为 0，祛除大量的冗余变量，只保留与目标变量最相关的预测变量，在简化模型的同时最大程度地保留数据集中重要信息。对于开展社会科学量化研究而言，Lasso 回归方法“能够充当稳定的变量筛选器、建立更具有概化能力和预测能力的模型”，在理论相对缺失的探索性、开创性研究中，“研究者更加需要采用这类方法避免对当前样本的过度解释，探索适用于总体的规律”（张沥今等，2020）。目前，Lasso 回归在临床医学（Demjaha et al., 2017; Omid, 2012）、金融投资（蒋翠侠等，2016）、地方财政（Yan et al., 2020）等领域的预测研究中均拥有不错的应用前景。从操作和应用的角度出发，

随着机器学习技术的日趋成熟，现在不少软件都可以直接调用功能函数实现 Lasso 回归建模，代表工具如 R 语言、Python、Stata15.0 以上等。

四、总结与讨论

随机森林之父Breiman曾在一篇非常有影响力的统计学论文中指出，统计建模有两种文化，一种为依靠直觉和经验选择模型的“数据建模文化”；另一种为完全不考虑模型的可解释性，只选择具有最高预测准确率模型的“算法建模文化”（Breiman, 2001）。经典统计学遵循生成模型的数据建模路径，其核心目标是因果解释，也就是理解一个结果如何与输入相关联。数据建模通常能够产生简单且可解释的模型，但牺牲的是模型的泛化能力。相对而言，算法建模遵循以预测为核心目标，即预测输入变量的未来结果。预测建模的优点在于它在观测样本外也具有良好的预测能力，但可能会产生黑盒结果，即很难拆解输入内容与输出结果之间的真实关联。两种文化分别对应以前向解释为中心（即对当前数据集的解释）的方法和以预测未来（即对新数据集的预测能力）为中心的方法。在撰写该篇文章的时候，Breiman 认为绝大多数统计学家（约98%）都属于前一种文化队列，相比之下，仅有约 2% 的少数统计学家和大多数机器学习研究人员属于后一种文化。“长久以来，统计学领域专家们热衷于使用数据建模来解决社会问题，倾心于数据描述和前向解释，这导致大量粗浅理论和值得质疑的研究结论的产生，使得统计学家无法在更为宽阔的领域实现自身的价值。而算法建模技术已经在统计学之外取得了飞速发展。它既可以被使用在大型复杂的数据集中，也可以

用于小型数据集。并且在小型数据集的处理上，算法模型甚至比数据模型更为准确，能产生更丰富的信息”（Breiman, 2001）。

在某种程度上讲，科学研究的预测问题等价于可复制问题，社会科学研究在结论可复制性问题上的关注度并不算高，这一方面与社会系统本身的复杂性相关，但也部分源自传统模型工具无法有效解决“过度拟合”问题所致。而机器学习建模方法则在工具层面为研究者提供了缓和这一尴尬境地的可能，原因包括：首先，基于机器学习方法产生的模型比传统量化模型的稳健性更高。对异常值的过分敏感和对理论工具的过度倚仗是许多传统模型的常见问题，而机器学习的交叉验证和正则化思路使这两个问题更容易获得诊断和补救。其次，即便抛开模型稳健性优势不谈，机器学习建模能够有效减少模型总误差也会在不同研究者中产生更稳定的结果。通过收缩项精简变量和窄化结果范围可以有效降低因模型过度拟合而产生的预测误差，提高研究结论的可复制性。最后，机器学习的树模型方法、正则化方法和奇异值分解等方法在可视化呈现变量的重要性权重方面明显优于传统模型。机器学习建模将注意力主要集中在那些重要变量上，帮助研究者快速识别和筛选对提升模型预测准确率作用权重排序在顶端的关键变量。

虽然机器学习建模方法具有相当优势，但仍然存在一些阻碍机器学习建模技术大范围应用的因素。首先，技术障碍。目前量化研究者通用的一些统计软件大都无法直接实现机器学习建模功能，机器学习实现路径的软件基础包括 R 语言、Python、Stata 等，它们均需要研究者掌握一定的编程能力。其次，预测能力存在边界。尽管机器学习方法可以最大化降低建构模型的泛化误差，但其推论和预测效果依旧是有边界的，该

边界仅限于训练集数据的样本特性。举个例子，格兰诺维特的“弱连带”（weak tie）假定，其抽象理论的基础为西方世界文化，在中国则并不适用（Bian, 1997），或者说泛化能力大打折扣。从机器学习建模的角度考虑，格兰诺特维是以西方经验数据为训练集训练模型，该模型在同为西方经验的测试集中具有较高预测能力。但是将该模型用于中国经验的测试集中则预测能力不佳。也就是说，机器学习建模方法的泛化能力与训练集数据的代表性直接相关，这并非技术问题，而是数据本身的问题。最后，参数依赖。与传统理论驱动的人工选元与建模不同，机器学习建模诸算法对参数非常敏感，特别是监督学习算法，以监督学习的正则化回归算法为例。如公式（4）所示，正则化回归损失函数 $L(\beta)$ 由典型 OLS 损失函数 $\text{loss}(\beta)$ 和惩罚函数 $\lambda P(\beta)$ 两部分内容构成，其中，参数 λ 的大小直接决定建构模型的复杂度，不同的 λ 取值可能产生不同的结果。λ 过大可能会导致模型自动剔除一些重要的预测变量，而 λ 过小则又会增加模型产生过度拟合问题的概率。目前，交叉验证方法是研究者确定 λ 数值的通用方法（Obuchi et al., 2016），即通过重复训练对比不同 λ 取值下的模型误差大小，选取最小误差下的 λ 值[①]。

伴随大数据时代的到来和计算社会科学的兴起，大规模数据和新型计算工具在为社会科学研究注入新的活力的同时，一种名为“技术恐惧”

[①]幸运地是，由 LASSO 回归的发明人，斯坦福统计学家 Trevor Hastie 领衔开发的 glmnet 包（R 语言）可以有效解决该问题。它的特点是对一系列不同 λ 值进行拟合，每次拟合都用到上一个 λ 值拟合的结果，从而大大提高了运算效率。此外它还包括了并行计算功能，这样就能调动一台计算机的多个核或者多个计算机的运算网络，进一步缩短运算时间。

的氛围也笼罩在部分研究者心头。我们知道，通过正则化方法，监督学习算法（如 Lasso 回归）可以实现关键变量筛选和模型复杂度调整，自主建构简化且具有较佳解释和预测能力的模型，而这一过程的实现几乎不需要理论的介入。那么，在大范围引入机器学习建模技术后，会不会使社会科学面临缺失理论和人文关怀的威胁，沦为技术驱动的数据挖掘游戏？应该说，这种担忧有其合理性存在，但这并不能成为我们拒斥新方法的托辞。一方面，接受新方法不代表全盘否定传统的研究方法，更遑论新方法也有其力有不逮的地方，机器学习的黑盒机制和预测失灵（如著名的谷歌流感趋势预测失效事件（Lazer et al., 2014））也常常遭到批判；另一方面，理论与技术并不是对立的关系，机器学习算法在探索性研究中可以为研究者的理论灵感予以技术支撑，而研究者的理论思考和经验总结恰恰在拆解机器学习“黑盒”为“灰盒”中起到重要作用。机器学习方法在社会科学量化研究中的应用前景可以分为两类：用数据的方法来研究科学（数据驱动 - 无监督学习 - 聚类和隐规则挖掘）和用科学的方法来研究数据（理论驱动 - 监督学习 - 回归和分类问题求解）。这两种应用模式都是机器学习助力社会科学量化研究的重要组成部分，缺一不可，只有把它们有机地整合在一起，才算得上完整。将新的机器学习方法运用于社会科学研究，这应视为机遇，而非威胁，研究者需要持有客观心态，努力确保自己有能力根据实际需要将经典方法和新方法的组合应用到他们的研究中。正如 Breiman 所呼吁那样，“如果我们使用数据的目标是解决问题，那么我们就需要改革创新，逐渐脱离对数据模型的强烈依赖，去接纳更多可能的工具”（Breiman, 2001）。

参考文献

陈云松、范晓光（2010）. 社会学定量分析中的内生性问题测估社会互动的因果效应研究综述 . 社会，(04)，91-117.

陈云松、贺光烨、吴赛尔（2017）. 走出定量社会学双重危机 . 中国社会科学评价，(03)，15-27+125.

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，(03)，94-117+244.

丁圣勇、樊勇兵编著（2018）. 解惑人工智能（页 22）. 北京：人民邮电出版社 .

冯帅帅、罗教讲（2021）. 社会学量化研究控制变量方法的反思与超越 . 深圳社会科学，(06)，95-105+115.

郝龙、李凤翔（2017）. 社会科学大数据计算——大数据时代计算社会科学的核心议题 . 图书馆学研究，(22)，20-29+35.

胡安宁（2012）. 倾向值匹配与因果推论：方法论述评 . 社会学研究，*27*(01)，221-242+246.

黄欣卓（2019）. 数据驱动社会科学研究的方向、路径与方法——关于“大数据与社会科学研究转型”主题的笔谈 . 公共管理学报，16(02)，159-167.

蒋翠侠、刘玉叶、许启发（2016）. 基于 LASSO 分位数回归的对冲基金投资策略研究 . 管理科学学报，(03),107-126.

李德毅主编（2018）. 人工智能导论（页 95、106-107）. 北京：中国科学技术出版社，95.

李航（2012）. 统计学习方法（页 3）. 北京：清华大学出版社 .

张沥今、魏夏琰、陆嘉琦、潘俊豪（2020）. Lasso 回归：从解释到预测 . 心理科学进展，(10)，1777-1791.

Antonakis, J., Bendahan, S. , Jacquart, P. , & Lalive, R. (2010). On making causal claims: A review and recommendations. The Leadership Quarterly, 21(6), 1086-1120.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences of the United States of America, 113(27), 7353-7360.

Babyak M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic medicine, 66(3), 411-421.

Bian, Y. (1997). Bringing strong ties back in: Indirect ties, network bridges, and job searches in China. American Sociological Review, 62(3)，366-385.

Breiman, L. (2001). Statistical modeling: The two cultures. Statistical Science, 16(3), 199-231.

Demjaha, A., Lappin, J. M., Stahl, D et al. (2017). Antipsychotic treatment resistance in first-episode psychosis: prevalence, subtypes and predictors. Psychological medicine, 47(11), 1981-1989.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. Social Science Electronic Publishing, 24, 395-419.

Hawkins D. M. (2004). The problem of overfitting. Journal of chemical information and computer sciences, 44(1), 1-12.

Hesterberg, T., Choi, N., Meier, L., & Fraley, C. (2008). Least Angle and L1 Regression: A Review，Statistics Surveys, 18(2), 61-93.

Hindman, M. (2015). Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. The ANNALS of the American

Academy of Political and Social Science, 659(1), 48-62.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L. , Kleinberg, J. , Margetts, H. , Mullainathan, S. , Salganik, M. J. , Vazire, S. , Vespignani, A. , & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. Nature, 595(7866), 181-188.

Jordan, M. I., Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science (New York, N.Y.), 349(6245), 255-260.

Kaplan, O. (1940). Prediction in the Social Sciences. Philosophy of Science, 7(4), 492-498.

King, G., Keohane, R., & Verba, S. (1995). The Importance of Research Design in Political Science. American Political Science Review, 89(2), 475-481.

Kohannim, O. (2012). Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression， Frontiers in Neuroence, (6), 115.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). Big data. The parable of Google Flu: traps in big data analysis. Science (New York, N.Y.), 343(6176), 1203-1205.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Lever, J., Krzywinski, M., & Altman, N. (2016). Points of Significance: Model selection and overfitting. Nature Methods, (13), 703-704.

McNeish, D. M. (2015). Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences. Multivariate Behavioral Research, 50(5), 471-484.

Molina, M.A., & Garip, F. (2019). Machine Learning for Sociology. Annual

Review of Sociology, 45(1), 27-45.

Montgomery, J. M., Hollenbach, F. M. , & Ward, M. D. (2012). Improving Predictions using Ensemble Bayesian Model Averaging. Political Analysis, (20), 271-291.

Obuchi, T., & Kabashima, Y. (2016). Cross validation in LASSO and its acceleration. Journal of Statistical Mechanics: Theory and Experiment, (5), 1-37.

Shmueli, G. (2010). To Explain Or to Predict? Statistics Science, (25), 289-310.

Ward, M. D., Greenhill, B. D. , & Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. Journal of Peace Research, 47(4), 363-375.

Watts, D. J. (2017). Should social science be more solution-oriented? Nature Human Behaviour, 1(1), 0015.

Wu C, Wang G, Hu S, Liu Y, Mi H, Zhou Y, et al. (2020). A data driven methodology for social science research with left-behind children as a case study. PLoS ONE, 15(11), e0242483.

Yan, D. , Chi, G. , & Lai, K. K. (2020). Financial Distress Prediction and Feature Selection in Multiple Periods by Lassoing Unconstrained Distributed Lag Non-linear Models. Mathematics, (8), 1-27.

Yarkoni, T. , & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspectives on psychological science, 12(6), 1100–1122.

Zhou, Z. (2018). A brief introduction to weakly supervised learning. National Science Review, (5), 44-53.