doi: 10.12012/T03-19

大数据、机器学习与统计学: 挑战与机遇

洪永淼, 汪寿阳

(1. 中国科学院数学与系统科学研究院, 北京 100190; 2. 中国科学院大学经济与管理学院, 北京 100190)

摘 要 随着数字经济时代的来临,基于互联网、移动互联网以及人工智能技术的经济活动每时每刻产生了海量大数据,这些海量大数据又反过来驱动各种经济活动. 大数据来源不一,形式多样,种类繁杂,既有结构化数据,也有非结构化数据,如文本、图像、音频、视频等,即使是结构化数据,也有新型数据,如函数数据、区间数据与符号数据等. 大数据大多拥有巨大的样本容量,也有潜在解释变量维数超过样本容量的高维大数据. 大数据的产生以及基于大数据的机器学习的广泛使用,对统计学产生了深刻影响. 本文从大数据的特点和机器学习的本质出发,讨论了大数据和机器学习对统计建模与统计推断的挑战与机遇,包括由抽样推断总体分布性质、充分性原则、数据归约、变量选择、模型设定、样本外预测、因果分析等重要方面,同时也探讨了机器学习的理论与方法论基础以及统计学和机器学习的交叉融合.

关键词 人工神经网络; 大数据; 维数灾难; 数据科学; LASSO; 机器学习; 统计学习; 数理统计学; 模型多样性; 模型不确定性; 非参数分析; 统计显著性; 充分性原则; 因果关系

Big Data, Machine Learning and Statistics: Challenges and Opportunities

HONG Yongmiao, WANG Shouyang

(1. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China;
2. School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract In the era of digital economy, economic activities based on Internet, mobile

收稿日期: 2020-04-15
基金项目: 国家自然科学基金 "计量建模与经济政策研究" 基础科学中心项目 (71988101); 国家自然科学基金 "经济科学发展战略研究" 项目 (71940004)
Supported by National Natural Science Foundation of China (NSFC) Fundamental Scientific Center Project "Econometric Modelling and Economic Policy Studies" (71988101); National Natural Science Foundation of China (NSFC) Project "Studies on Development Strategies for Economic Science in China" (71940004)
作者简介: 洪永淼, 世界计量经济学会会士、发展中国家科学院院士, 中国科学院数学与系统科学研究院特聘研究员, 研究方向: 计量经济学、时间序列分析、金融计量学、统计学, E-mail: ymhong@amss.ac.cn; 汪寿阳, 发展中国家科学院院士、国际系统与控制科学院院士, 中国科学院数学与系统科学研究院特聘研究员, 研究方向: 金融系统工程、经济分析与预测, E-mail: sywang@amss.ac.cn.
致谢: 感谢陈丽纯、李四光、刘婧媛、钟威、钟锃光的帮助与建议.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Internet and artificial intelligence are generating massive data, which have promoted economic growth in return. Big data have various sources and forms, including structured and unstructured data. Unstructured data can be text, images, audio and video, and new structured data can be functional data, interval data, symbolic data, etc. Most Big data has a huge volume, and some is Tall Big data, whose dimensionality of potential explanatory variables is larger than its sample size. The rise of Big data together with machine learning, a main computer-based automatic analytic tool for Big data, has profound implications on statistical science. Based on the characteristics of Big data and the nature of machine learning, the paper discusses the challenges and opportunities brought by Big data and machine learning to statistical modeling and inference, including sample inference, sufficiency principle, data reduction, variable selection, model specification, out-of-sample prediction and causality. We also explore the theory and methodology foundation of machine learning and its integration with statistics.

Keywords artificial neural network; Big data; curse of dimensionality; data science; LASSO; machine learning; statistical learning; mathematical statistics; model variety; model uncertainty; nonparametric analysis; statistical significance; sufficiency principle; causality

1 导言

统计学是一门关于数据分析的方法论科学，为自然科学和社会科学的实证研究和经验分析提供严谨的分析方法和工具. 随着互联网与移动互联网技术及其应用的快速发展，大数据 (Big data) 和用于大数据分析的机器学习 (machine learning) 正在对统计科学产生深刻的影响. 与传统数据相比，大数据体量巨大，来源不一，种类繁多，有结构化、半结构化、非结构化等各种形式，大多数是实时或近乎实时生成和记录的数据. 一种观点认为，大数据是全样本与几乎接近全样本，因此统计学的随机抽样理论，特别是以随机样本推断总体分布性质的统计方法不再适用. 同时，也有观点认为，大数据特别是高频乃至实时数据的出现以及机器学习的应用，使得基于数据的系统特征与变量之间相关性的精准预测成为可能，因此在实际应用中，只需要相关性，不需要因果关系. 那么，大数据是否改变了统计科学的理论基础？比如，随机抽样推断、充分性原则、数据归约、样本外预测、因果分析等统计方法，是否将会改变，甚至有些统计学的基本原理是否将不再适用？另外，大数据给统计建模与统计推断的理论与应用带来了哪些挑战与机遇？作为大数据分析的重要工具，机器学习与统计建模的主要区别是什么？机器学习与统计推断有什么联系与共同点？众所周知，基于大数据的机器学习常常能够提供较为精准的样本外预测，但在大多数情况下，它就像一个“黑箱”，很难甚至无法给予直观解释. 那么，统计学能否为机器学习提供有意义的理论解释呢？机器学习与统计学是否可以结合起来？如果可以，这种交叉融合对统计科学的未来发展将产生什么影响？本文试图回答这些重要问题，并提供一些探索性的解决思路. 在第二节，我们简要讨论统计建模与统计推断的习惯做法，指出传统统计建模与统计推断的基本假设和基本思想. 在第三节，我们将讨论大数据特别是经济大数据的主要来源和主要特点. 在第四节，

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

我们将讨论机器学习的本质以及几种重要的机器学习方法. 第五节将探讨大数据与机器学习对统计建模与统计推断的影响, 特别是对统计科学所带来的挑战与机遇, 同时也探讨在大数据背景下如何将机器学习和统计学有机结合起来, 开辟统计科学和计量经济学研究的新领域与新方向. 第六节是结论.

我们得出以下主要结论:

1) 大数据没有改变统计学通过随机抽样推断总体分布特征的基本思想. 许多基本统计方法, 包括充分性原则、数据归约、因果推断等, 依然适合于大数据分析, 其中有些统计方法, 如充分性原则与数据归约, 其重要性甚至因为大数据的出现而大大增强. 当然, 这些统计方法在大数据条件下需要创新与发展.

2) 大数据提供了很多传统数据所没有的信息, 大大拓展了统计学研究的领域边界. 例如非结构化文本数据 (text data) 使得构建一些重要社会经济心理变量成为可能, 包括测度投资者情绪、居民幸福感、经济政策不确定性等, 而高频甚至实时数据使得实时预测和高频统计建模与统计推断成为可能.

3) 由于样本容量巨大, 大数据预计将改变基于统计显著性来选择统计模型重要变量的习惯做法. 特别地, 抽样数据变异性对统计建模与统计推断产生了巨大影响, 研究范式也将从参数估计不确定性转变为模型选择不确定性; 这同时也对统计建模与统计推断提出新的挑战, 包括数据生成过程的同质性与平稳性以及统计模型唯一性等基本假设的适用性问题.

4) 机器学习的兴起得益于大数据的产生以及计算能力的爆炸式发展. 机器学习与统计推断有很多共同之处, 包括在数据生成过程的随机性假设和由抽样推断总体分布性质等基本思想. 与统计建模与统计推断一样, 机器学习也存在并且特别注重样本偏差问题.

5) 与统计学的参数建模方法相比, 绝大多数机器学习方法不对数据与变量之间的关系给予具体的模型假设或限制, 而是根据目标函数通过算法直接学习、探索数据的系统特征和变量之间的统计关系, 使目标函数最优化. 机器学习的本质是一个数学优化问题与实现该优化问题的计算机算法问题, 它比统计学的参数建模更普遍、更灵活, 包括对重要解释变量的选择与测度.

6) 与机器学习一样, 统计学的非参数分析 (nonparametric analysis) 也是不用假设任何具体模型形式而能够一致估计刻画数据生成过程的未知函数 (如概率密度函数或回归函数). 很多重要的机器学习方法, 如决策树、随机森林、k 最近邻法 (k-NN)、人工神经网络、深度学习等, 其实就是统计学的非参数方法. 这些非参数方法的统计性质, 特别是其对未知函数的一致性估计的性质, 能够从理论上解释与帮助理解为什么一些机器学习方法拥有精准的样本外预测能力. 但是, 机器学习不完全等同于统计学的非参数分析方法, 例如, 机器学习在处理高维解释变量时具有更大的灵活性, 而非参数分析则存在众所周知的 "维数灾难 (curse of dimensionality)" 问题.

7) 在大数据背景下, 机器学习与统计推断的有机结合有望为统计科学与数据科学提供一些新的发展方向, 特别是在统计学习这一新兴的交叉领域, 包括变量降维、稳健推断、精准预测、因果识别等重要方面.

2 实证研究与统计分析

统计科学为现代科学的实证研究奠定了坚实的方法论基础, 提供了重要的方法与工具, 其应

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

用包括以非实验观测数据为基础的经济学与其他社会科学. 统计推断的基本思想是假设所研究的系统是服从某一概率法则的随机过程, 现实观测数据是从这个随机过程产生的, 而这个随机过程称为数据生成过程 (data generating process, DGP). 统计实证分析的主要目的是通过对观测数据进行统计建模, 推断出 DGP 的概率法则或其重要特征, 然后运用于各种实际应用中, 如解释经验典型特征事实、检验经济理论与经济假说、预测未来变化趋势、评估公共政策等. 详细讨论可参见文献洪永淼 (2007).

在统计建模与统计推断中, 一般假设 DGP 的概率法则可由唯一的数学概率模型来刻画, 模型通常将因变量与一些解释变量或预测变量联系起来. 同时, 假设该数学模型的函数形式已知, 但包含低维的未知参数. 这是一种参数建模 (parametric modeling) 方法, 在统计学中应用最为广泛. 统计推断的主要目的是用观测数据估计模型的未知参数值, 将经济理论或经济假说转化为统计参数假设, 然后进行参数假设检验, 并对实证结果提供经济解释. 在统计实证研究中, 常见做法是基于一个预设的显著性水平 (如 5%) 判断一个参数估计值或参数假设在统计学上是否显著, 特别是使用检验统计量的 P 值来判定参数估计值或参数假设的统计显著性. 如果具有统计显著性, 则相应的解释变量将视为一个重要决定因素, 并留在统计模型中. 如果一个具有统计显著性的解释变量没有被包含在统计模型中, 则称该变量为遗漏变量, 且模型误设. 模型误设还有其他原因, 如函数形式错误、忽视结构变化或异质性等. 通常会通过样本内诊断检验或拟合优度来判断设定模型是否足以描述观测数据或者刻画 DGP 的概率法则.

在实际应用中, 常用的标准统计模型包括经典线性回归模型、Probit 或 Logit 离散选择模型、生存分析或久期分析中的比例风险模型 (Cox (1972)) 等. 作为模型的重要输入, 经济观测数据一般指在现实条件下所观测到的数据, 这些数据不是在可控实验条件下产生的. 非实验性是经济学乃至社会科学的最显著特征. 大多数实际观测数据的样本容量通常不太大. 观测数据以及相关的统计模型可能也存在各种缺陷或不尽如意的特征, 如随机扰动项的条件异方差与自相关、删失数据、截断数据、变量误差、遗漏观测值、内生性、维数灾难、弱工具变量、不可观测的虚拟事实、部分识别、甚至数据操纵与数据造假等, 充分考虑这些数据缺陷或特征有助于改进统计推断. 许多年来, 统计学和计量经济学的实证研究一直沿用上述统计建模与统计推断过程.

我们发现, 这些做法直接或间接地基于至少六个关键假设:

假设 1 随机性. DGP 是一个随机过程.

假设 2 模型唯一性. DGP 的概率法则由唯一的数学概率模型来刻画.

假设 3 模型正确设定. 概率模型设定是正确的, 即存在唯一的未知参数值, 使得概率模型与 DGP 的概率法则相吻合.

假设 4 抽样推断总体. 使用包含 DGP 信息的样本数据来推断总体分布特征, 特别是 DGP 的概率法则, 这是基本的统计推断方法, 也导致概率论成为推断统计学的理论基础.

假设 5 代表性样本. 描述观测数据的随机样本不存在样本选择偏差, 而观测数据的样本容量通常不会太大.

假设 6 统计显著性. 基于统计推断, 尤其是使用统计检验量的 P 值, 在预设的显著性水平 (如 5%) 上判断解释变量或预测变量是否重要, 并据此提供逻辑解释.

接下来, 我们将讨论大数据特别是经济大数据的主要特征和机器学习的本质, 以及它们给统

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

计建模与统计推断的理论与应用所带来的重要影响、挑战与机遇. 作为一种基于计算机算法的优化分析工具, 机器学习是分析大数据不可或缺的重要方法.

3 大数据的主要特征

大数据的产生得益于信息技术的快速发展, 尤其是互联网与移动互联网技术的广泛应用. 互联网设备与传感器的指数增长是产生与收集海量大数据的主要原因. 大数据的来源很多, 包括计算机商业交易平台、移动电话、社交媒体、网站信息、搜索数据、传感器与卫星图像、交通数据等. 在金融市场、各种线下线上商品交易平台, 扫描器与电子支付系统记录了逐笔交易数据. GPS 和北斗传感器记录了地球上各种重要的气候环境数据与物理数据, 如中国主要大城市 PM 2.5 的观测值、全球大城市的夜间灯光亮度数据. 望远镜与射电望远镜全天候观测太空, 实时记录了各种天文物理数据流. 各类企业和政府网站也提供了有用的信息, 特别是互联网巨头, 即所谓的大型科技 (Big Tech) 公司, 如中国的百度、阿里巴巴、腾讯、京东, 美国的谷歌、亚马逊、脸书、苹果等. 在数字经济时代, 海量经济大数据的产生得益于基于计算机的互联网与移动互联网的各种经济活动与商业交易, 而且大数据作为一种新的生产要素, 反过来进一步推动经济发展. 无人驾驶的发展就是大数据应用的一个典型案例. 截至 2019 年底, 中国互联网与移动互联网用户人数超过 9 亿人, 远远超过美国与欧盟网民人数的总和. 现在已出现了一个新的 GDP 概念, 即数据生成总值 (gross data product), 用于测度每个国家或地区的数据资源总量及其利用程度.

大数据具有以下四大特征, 即所谓的 "4V" 特征:

1) 海量性 (volume). 从各种渠道收集的信息, 包括商业交易数据、社交媒体数据、传感器数据以及机器对机器数据等, 在过去, 如何存储如此大规模的数据是一个技术难题, 但新技术 (如 Hadoop) 的快速发展已经减轻了存储负担.

2) 高速性 (velocity). 大数据以前所未有的速度产生与传播, 必须及时存储与处理. RFID 电子标签、传感器、智能停车收费系统实现了实时或近乎实时处理海量数据的需求. 在许多情况下, 大数据可能会以聚类方式产生, 即数据产生的速度并不均匀, 而是随着时间的推移出现周期性波动. 比如, 股市交易有明显的周期模式, 通常开盘和收盘时成交量较大, 午间成交量较小. 基于事件触发的日常周期性峰值数据在加载管理上难度很大, 更不用说非结构化数据了.

3) 多样性 (variety). 大数据形式多样, 既有传统结构化数字型数据, 也有非结构化的文本文档、邮件、图片、视频、音频、股票行情数据等. 非结构化数据提供了传统数据所没有的非常丰富的新信息, 这已成为大数据的一个最重要的特征. 结构化数字型数据也有新型数据, 如函数数据、区间数据和符号数据 (symbolic data) 等.

4) 真实性 (veracity). 与传统数据相比, 大数据一般体量庞大, 但很多大数据信息密度低, 噪声大. 此外, 也可能存在遗漏数据和操纵数据, 导致信息失真, 因此有必要进行数据清洗与处理.

大数据的海量性具有双重含义. 一方面, 大数据拥有非常大的样本容量. 许多大数据的样本容量可能是数万甚至是数百万的观测值. 如果大数据的样本容量很大且远大于解释变量或预测变量的维数, 那么这种大数据称为 "高大数据 (Tall Big data)". 庞大的样本容量意味着可以从大数据尤其是非结构化数据中获取很多新的信息, 从而改进对 DGP 的统计推断. 通常, 由于计算机容量与计算速度的限制, 只有一小部分高大数据用于可行性统计分析 (如 Engle and Rusell

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

(1998), Engle (2000)). 另一方面, 大数据的海量性不一定是指样本容量非常大. 它也可能是指在给定时间内从不同维度对 DGP 的大量描述. 换句话说, 大数据拥有一个高维的潜在解释变量或预测变量的集合. 比如, 利用谷歌搜索中国一些城市的旅游趋势. 这为探索重要解释变量提供了巨大的可能性与灵活性. 当潜在解释变量或预测变量的维数超过样本容量时, 这将给统计建模与统计推断造成巨大挑战, 这在统计学上称为 "维数灾难", 而具有此特征的大数据则称为 "胖大数据 (Fat Big data)". 对于高维解释变量的集合, 许多解释变量可能对因变量没有影响, 也有可能很多解释变量之间存在多重共线性. 因此, 有必要发展各种可行的变量选择方法, 这本质上是一种变量降维 (dimension reduction) 或数据归约 (data reduction).

大数据的高速性指的是能够在高频甚至实时条件下记录或收集数据. 这使得及时的数据分析与预测成为可能. 比如, 在经济统计学中, 可以构建高频宏观经济变量, 以便及时了解宏观经济变化趋势, 提升经济政策干预的时效性. 经济统计学的现行做法只能获取居民消费指数 (consumer price index, CPI) 和生产者物价指数 (producer price index, PPI) 等月度时间序列数据. 然而, 基于互联网信息和人工智能工具, 完全可以构建 CPI 和 PPI 的日度数据, 甚至抽样频率可以更高. 在时间序列分析中, 高频数据的可获得性可以避免依时间加总 (temporal aggregation) 而导致的信息缺失. 例如, 比起使用每日收盘股票价格数据, 我们可以用股价的日内 (intraday) 数据甚至逐笔交易数据来估计金融资产的每日波动率. 日内时间序列数据包含了当日价格变动范围, 比当日收盘价数据拥有更多的波动信息. 再如, 可以利用点过程的时间序列数据来研究不同资产或不同市场间的 Granger (1969) 因果关系或时间维度上的领先滞后关系. 高频数据也使时变结构研究成为可能. 如果模型参数随时间缓慢改变, 我们可能需要更高频的观测值来推断任意时间点的参数值.

大数据的多样性指的是数据种类繁多、形式多样, 有结构化、半结构化与非结构化数据, 而结构化数据也包括一些新型数据, 如函数数据、区间数据乃至符号数据等, 同时可能结合了不同的抽样频率. 长期以来, 统计学主要关注传统结构化数据. 当今的数据拥有各种来源, 也可能有不同的物理存储地址, 导致不同系统间各种数据的连接、匹配、清洗、转换变得困难. 如何将不同来源、不同结构、不同形式、不同频率的各种数据汇聚一起, 这是一个巨大挑战. 从统计学角度看, 大数据将比传统数据提供更多有价值的信息, 因此可以用来发展更高效的统计推断方法与工具. 特别是, 社交媒体 (如微博和脸书) 数据越来越受关注, 这些信息通常是非结构化或半结构化的数据, 很难甚至无法从传统数据中获取. 将非结构化数据与传统结构化数据相结合, 可以更好推断 DGP 的本质特征.

大数据的真实性是指大数据存在大量噪声, 包括虚假信息和失真数据. 因此, 如何去伪存真、有效概括并提取大数据的有用信息显得非常重要. 统计分析的本质是有效地从数据中提取有价值的真实信息. 虽然很多经典统计方法很有用, 如主成分分析和聚类分析, 但也需要发展概括、提取大数据中有用信息的新方法与新工具. 由于大数据具有容量大、维度高与信息密度低等特点, 统计学的充分性原则在大数据分析方面可发挥巨大作用, 尤其在数据归约与变量降维方面, 因此我们迫切需要发展基于计算机算法的有效的数据归约方法.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4 机器学习及其本质

与统计学一样, 机器学习也是一种重要的大数据分析工具. 在大数据时代, 统计学和机器学习已经成为新兴的数据科学的最重要分析方法. 机器学习由于大数据和云计算的出现而得到迅速发展与广泛应用, 但是机器学习不能替代统计分析. 例如, 尽管机器学习在改善样本外预测和模式识别 (如面部识别) 方面非常有用, 但统计学在推断分析、维数约简、因果识别和结果解释等方面可以发挥很大作用. 机器学习与统计学是互补的, 两者的交叉融合可以为统计科学与数据科学提供新方法与新工具.

"机器学习" 这一术语是由人工智能开拓者之一 Arthur Samuel 于 1959 年提出来的. 机器学习是计算机科学的一个重要领域, 尤其是人工智能的一个重要组成部分. 机器学习利用数学、人工智能工具赋予计算机系统自动 "学习" 数据、"识别" 模式、并做出预测或决策的能力, 无须明确的人工编程. 它是从人工智能的模式识别研究和机器学习理论中演变而来的, 主要探索能够自己有效学习数据并做出预测的算法研究与算法构建. 机器学习可以分为三个主要类别: 监督学习 (supervised learning)、无监督学习 (unsupervised learning) 和强化学习 (reinforcement learning).

监督学习基于训练数据 (包含输入和输出) 来构建算法. 训练数据包含一组训练样例, 每个训练样例拥有一个或多个输入与输出, 称为监督信号. 通过对目标函数的迭代优化, 监督学习算法探索出一个函数, 可用于预测新输入 (非训练数据) 所对应的输出. 优化目标函数能够使算法准确计算出新输入所对应的输出预测值. 监督学习算法包括分类和回归. 当输出只能取一个有限值集时, 可用分类算法; 当输出可取一定范围内的任意数值时, 可用回归算法.

无监督学习在只包含输入的训练数据中寻找结构, 如数据点的分组或聚类. 无监督学习算法不回应反馈, 而是识别训练数据的共性特征, 并基于每个新数据 (非训练数据) 所呈现或缺失的这种共性特征作出判断. 无监督学习主要应用于统计学概率密度函数估计, 也可用于涉及数据特征总结与解释的其他领域. 聚类分析是一种重要的无监督学习方法. 它将一个观测数据划分为多个子集 (称为簇, clusters), 使得同一簇的观测数据在一个或多个预设准则上具有相似性, 但是不同簇的观测数据不具有相似性. 不同的聚类方法对数据结构做出不同的准则假设, 一般由某种相似性度量准则所定义, 通过内部紧密度 (同一簇中数据的相似度) 和分离度 (簇间差异) 进行评估.

强化学习是研究算法如何在动态环境中执行任务 (如无人驾驶) 以实现累计奖励的最大化. 由于强化学习的一般性, 许多学科也对该领域有所研究, 如博弈论、控制论、运筹学、信息论、仿真优化、多智能体系统、群集智能、统计学与遗传算法等. 在机器学习中, 动态环境一般表现为马尔可夫决策过程 (Markov decision process). 许多强化学习算法使用动态规划技术. 强化学习算法可用于自动驾驶或与人类博弈比赛.

从本质上说, 机器学习是数学优化问题与算法优化问题. 机器学习与数学优化联系紧密, 数学优化为该领域提供了理论、方法与应用. 同时, 机器学习与计算统计学密切相关, 常常交叉重叠, 注重利用快速有效的计算机算法进行预测. 在机器学习领域, 许多学习问题可表述为最小化某个预设的损失函数. 为了避免过度拟合 (overfitting) 现象, 其最终目的通常转化为基于未知数据的预测误差最小化问题. 具体地说, 机器学习基于训练数据, 学习与挖掘训练数据的系统特征和变量之间的统计关系 (如相关性), 以预测新的未知数据. 为了得到精准预测的算法, 一般将现

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

有数据分为两个子集 —— 训练数据 (training data) 和测试数据 (test data). 训练数据用以学习与挖掘数据的系统特征以及变量之间的统计关系, 然后利用这些系统特征与统计关系预测未知数据的行为. 为了保证精准预测, 必须避免对训练数据的过度拟合. "过度拟合" 现象是指挖掘只存在于训练数据但不会出现于未知数据的特征与统计关系, 而这些特征与统计关系可以改进训练数据的样本内拟合, 但无助于样本外预测. 因此, 对预测效果的评价需要基于另一部分数据, 即测试数据. 此外, 为了进一步避免过度拟合, 通常还引入一个惩罚项, 对算法的复杂程度给予相应的惩罚, 即算法的复杂程度越高, 惩罚越重. 因此, 机器学习就是从训练数据中寻找一个优化算法, 使预测测试数据的损失函数加上惩罚项最小化, 以达到最优样本外预测效果. 常见的机器学习方法包括决策树、随机森林、k 最近邻法、支持向量机、人工神经网络、深度学习等. 现在, 分别简单介绍如下:

决策树 (decision tree). 决策树学习将决策树作为预测方法, 体现了从一些特征变量 (如解释变量) 的观测值 (在分支中体现) 到目标变量 (在叶子中体现) 的目标值的整个预测过程. 决策树学习是统计学、数据挖掘和机器学习的一种预测方法. 若目标变量取一组离散值, 则决策树称为分类树, 其中, 叶子代表类标签, 分支代表产生这些类标签的功能连词. 若目标变量取连续值 (通常是实数), 则决策树称为回归树. 在决策分析中, 决策树可具体形象地描绘决策和决策过程. 在数据挖掘中, 决策树对数据进行描述, 但是所得分类树可用作决策的输入.

随机森林 (random forest). 对大数据特别是胖大数据而言, 由于存在很多潜在的解释变量或预测变量, 解释变量可能存在着不同程度的多重共线性, 使得对样本数据的 "微扰 (perturbation)" 可能导致最优预测模型 (不同解释变量的组合) 的大幅变动, 这称为模型不确定性 (model uncertainty). 为了获得稳健预测, Breiman (2004) 于提出了随机森林方法. 基于原始观测数据, 通过重复抽样产生一系列新的随机数据, 每个数据培植一棵决策树, 然后对所产生的一系列决策树的预测值进行平均, 这种预测方法称为随机森林.

k 最近邻法 (k-nearest neighbor). 这个方法根据一些特征变量 (如解释变量) 的取值, 选择 k 个取值最靠近某个预定值的特征变量观测值, 然后将对应于这 k 个取值最邻近预定值的因变量观测值进行平均, 作为对因变量的一个预测. 这个方法称为 k 最近邻法.

支持向量机 (support vector machine, SVM). 这是一种用于分类和回归的监督学习方法. 若给定一组训练样例, 每个样例标记为属于两个类别中的一类, 则 SVM 训练算法可预测新样例属于哪个类别. SVM 训练算法是一个非概率的二元线性分类器. 除了实现线性分类, SVM 也可以进行高效的非线性分类, 将其输入隐式映射到高维特征空间中.

人工神经网络 (artificial neural network, ANN). 这是一个计算机算法系统, 其部分灵感源自构成动物大脑的生物神经网络, 通过考察训练数据的样例 "学习" 如何执行任务. 人工神经网络由大量称为 "人工神经元 (neurons)" 的单元或节点相互连接而成, 大致模仿生物大脑中的神经元系统. 如同生物大脑中的突触, 每个连接都可以将一个人工神经元的 "信号" 传递到另一个人工神经元. 接收到信号的人工神经元可以处理该信息, 然后将信息传递给其他与之关联的人工神经元. 人工神经元之间的连接信号通常是一个实数, 人工神经元一般具有一个根据学习所得而调整的权重, 可提高或降低连接中的信号强度. 人工神经元可能具有一个阈值, 只有当汇总加权信号超过该阈值时才会发送信号. 这样, 每个人工神经元的输出由其所有输入的权重总和的某个非线

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

性函数 (称为激活函数, activation function) 计算而得. 通常, 人工神经元聚集成一个或几个隐藏层 (hidden layers). 不同的隐藏层可以对其输入执行不同类型 (即不同的激活函数) 的转换. 信号可能在多次遍历图层后从最初输入层传递到最后输出层. 人工神经网络方法的最初目标是以与人类大脑相同或类似的方式解决问题, 但随着时间的推移, 人们将目光转移到执行特定任务上, 从而偏离了生物学. 目前, 人工神经网络已有各种应用, 如计算机视觉、语音识别、机器翻译、社交网络过滤、下棋游戏、电子游戏、医学诊断等.

深度学习 (deep learning). 如果人工神经网络包含多个隐藏层, 则称为深度学习方法. 深度学习试图模拟人类大脑将光和声处理成视觉和听觉的方式. 计算机视觉和语音识别就是深度学习的一些成功应用.

5 大数据、机器学习与统计学的关系

数据描述是数据分析的起点, 这一点在大数据时代由于不同种类、不同形式特别是非结构化数据的出现而显得更为重要. 事实上, 鉴于大数据的多样性, 尤其是文本、图表、音频、视频等非结构化数据, 必须开发新的方法与工具来记录、存储、整理、清洗、描述、表现、分析、概括与解释大数据. 很多大数据特别是非结构化大数据的获得与分析, 都必须使用人工智能技术, 一个著名例子是爬虫的应用. 美国劳工统计局原来依靠人工操作的调查问卷答案分类工作, 现在已有 85% 被深度学习替代, 而且深度学习的准确率高于人工. 又如, 大数据可视化作为大数据一种直观表现形式, 在实际应用中越来越受欢迎. 商业智能就是大数据在现代商业中的一个重要应用, 它通过应用各种人工智能的技术与方法来提取、概括、表现大数据的重要信息, 从而改善商业决策的科学性与提升企业管理的精细化水平.

由于大数据的 "4V" 特征, 大数据分析需要使用来自不同领域的方法与工具, 包括数学、计算机科学、统计学、数据科学等学科. 大数据分析的主要目的是从传统数据中发现不易察觉的模式、趋势、异象 (anomalies)、关联、因果效应以及其他特征等各种有价值的信息. 目前, 广泛使用的大数据分析方法与工具主要是机器学习和统计方法, 尤其是计算统计学工具. 在本节, 我们将论证大数据和机器学习并没有改变统计建模与统计推断的一些基本思想, 如抽样推断总体分布性质、充分性原则与数据归约、因果推断、预测等. 因此, 现代统计学在大数据分析方面仍然将发挥基础性的关键作用. 但是, 大数据的复杂性和机器学习的广泛应用的确给统计科学提出了一些重要挑战, 这些挑战有望为推动现代统计学的发展提供各种机遇, 尤其是创新统计理论、方法与工具等方面.

5.1 非结构化数据与文本回归分析

从统计学角度看, 相比传统数据, 大数据特别是非结构化数据将带给我们更多的有价值的信息, 这些信息可用于发展新的统计方法与工具. 比如, 在互联网时代, 社交媒体 (如微博和脸书) 数据经常反映了社会公众或社会群体对每个时期重要事件的看法, 而这些重要事件常常对社会经济造成很大影响, 因此受到越来越多的关注 (参见 Shiller (2019)). 社交媒体数据通常以非结构化或半结构化形式呈现, 但通过爬虫等技术抓取信息, 可用于构建新的解释变量或预测变量, 如消费者幸福感指数、投资者情绪指数、经济政策不确定性指数、经济政策变化指数、社会舆情指数

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

等 (参见 Baker and Wurgler (2007), Baker, Bloom and Davis (2016), Chan and Zhong (2018)). 这些从文本数据构建的重要变量包含传统数据所没有的信息, 可通过统计回归模型等方法, 分析与测度它们对社会经济金融市场的影响, 这就是所谓的文本回归 (textual regression) 分析.

除了基于社交媒体非结构化数据构建经济心理指数之外, 我们还可以通过大数据与人工智能方法, 构造高频宏观经济时间序列指数, 如 CPI 和 PPI 的每日时间序列数据. 这将有助于我们及时预测宏观经济的变化趋势, 包括实时预测 (nowcasting); 参见文献 Giannone, Reichlin and Small (2008), Bok et al. (2017). 目前, 绝大部分宏观经济指标最高频数据是月度数据, 像国内生产总值 (GDP) 这样重要的宏观经济变量还没有月度数据. 大数据的出现和人工智能技术的使用可以显著提高宏观经济数据的测度频率.

5.2 抽样推断原则

大数据并不意味着可以获取 DGP 的总体分布的完全信息. 曾经有一种观点认为, 大数据提供了总体分布的完全信息或近乎完全的信息, 因此在大数据时代, 海量数据将使推断统计学变得价值有限甚至毫无价值. 这种情形只有在统计模型是唯一正确设定而且不变的假设条件下才可能发生. 众所周知, 推论统计学的基本思想是从随机样本推断总体分布特征, 而所推断出来的总体分布特征, 也适合于刻画从同一总体分布产生的其他随机样本. 假设某一参数统计模型是正确设定, 则当样本容量非常大时, 确实可以不必担心参数估计量的抽样可变性 (sampling variability), 即参数估计不确定性将可以忽略不计. 尽管当大数据的样本容量很大时, 模型参数估计结果的抽样可变性也因此变得没有以前那么重要, 但是通过随机样本推断总体分布特征的统计思想仍未改变, 取而代之的很可能是模型选择不确定性. 模型选择不确定性可能是因为大数据中存在大量解释变量, 而这些解释变量具有不同程度的多重共线性, 或者是因为 DGP 具有异质性或时变性, 或者是因为模型误设. 因此, 当对数据进行 "微扰" 时, 即增加或减少一小部分数据, 基于预定统计准则的最优统计模型将会显著改变. 我们知道, 机器学习的主要目的, 是基于对训练数据的 "学习" 经验, 预测未知样本的行为或表现. 其假设前提是从训练数据中 "学习" 到的一些系统特征与统计关系 (如相关性、异象), 会在未知数据中再次出现, 不管未知数据是截面数据或时间序列数据. 换言之, 机器学习就是从训练数据中挖掘出可以泛化到未知数据的系统特征, 并根据这些共同系统特征进行预测. 如果我们将这些共同系统特征定义为 DGP 的总体特征, 那么机器学习这种样本外预测方法, 无论是基于截面数据还是时间序列数据, 均遵循类似从样本推断总体特征的基本统计思想. 之所以需要测试与验证的主要原因是基于训练样本的 "学习" 经验可能会存在过度拟合现象, 因而不能刻画样本外的系统特征. 过度拟合可能是由样本选择偏差、异质性、时变性、甚至模型误设所导致. 例如, 在预测当前新冠肺炎疫情未来发展趋势时, 需要考虑可能的新冠肺炎病毒变异性, 即结构变化. 因此, 机器学习也可视为遵从抽样推断总体分布性质的统计思想, 至少是一种广义的抽样推断的统计方法, 同时, 由于拥有海量大数据, "总体" 的概念可以更一般化, 即允许具有异质性或时变性的 DGP, 当然不同异质主体或不同时期的 DGP 仍然需要假设具有一些共同的系统特征.

机器学习早在 20 世纪 50 年代就已经提出来, 但是它的快速发展与广泛应用发生在从 20 世纪 90 年代开始的大数据时代. 海量大数据的收集、存储、处理与分析必须依赖人工智能方法, 而

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

海量大数据的可获得性为机器学习探索与学习数据之间可能存在的复杂关系 (如非线性关系) 提供了丰富的素材. 作为大数据的一种重要分析方法, 机器学习与统计学密切相关, 两者拥有一些共同点. 机器学习是一种设计、推导复杂算法的数学方法, 通过学习训练数据所包含的历史关系与系统特征, 利用计算机算法自动得出最佳预测. 与统计学一样, 机器学习也假设 DGP 是一个随机过程, 而且其结构或概率法则是未知的. 算法的核心目标是泛化从训练数据中所 "学习" 到的经验, 即外推预测, 其本质是从训练样本推断未知样本的总体特征. 所谓泛化 (generalization) 指的是机器以学习训练数据的经验为基础, 对一个未知的新样本进行精准预测. 一般假设训练样本来自一个未知的概率分布, 机器学习需要从训练数据中学习未知概率分布的系统特征, 以便对新样本做出准确预测. 对未知新样本能够做出准确预测的重要前提是训练数据和测试数据的 DGP 或概率法则保持不变, 这与统计推断通过抽样推断总体分布性质的基本思路是一致的. 两者最主要的区别在于机器学习的预测不用统计模型而直接基于计算机算法, 而统计预测一般是基于某个参数模型, 其函数形式假设已知, 但包含一个未知的低维参数向量. 如果数据容量不大, 参数模型可能很有用, 但如果数据非常多, 模型可以拓展为一般化的数据算法, 这样更有可能捕捉大数据中变量之间的各种复杂关系.

均方误差特别是其平方偏差 - 方差分解就是测度泛化误差的一种常用统计准则. 为了实现最佳泛化, 算法的复杂性必须匹配 DGP 的复杂性. 一方面, 若 DGP 比算法结构更复杂, 则算法拟合数据的能力较弱. 另一方面, 如果算法复杂性增高, 则训练数据的拟合误差将减小. 然而, 若算法过于复杂, 则会导致过度拟合且泛化误差增大. 概率理论可以为测度和约束泛化误差提供一个有效方法. 这是机器学习和统计推断共同的概率论基础. 事实上, 贝叶斯统计学也是机器学习的一个重要理论方法.

在实际应用中, 机器学习可能会遇到各种样本偏差问题. 比如, 一个只基于现有客户训练数据的机器学习算法并没有体现新客户的信息, 因此可能无法预测新客户群的需求. 这就是统计学中著名的样本选择偏差问题, 其原因是不同客户群可能存在潜在的异质差别. 另一种可能性是时变性即结构变化所导致的样本偏差. 针对样本偏差问题, 可以使用统计学的 Holdout 方法和 k 折交叉验证法 (k-fold cross-validation) 等来验证机器学习算法. Holdout 方法将数据分为训练集和测试集, 这是最常用的验证方法; 而 k 折交叉验证法则是随机地将数据分为 k 组子集, 其中 $k-1$ 组用于训练算法, 剩下一组用于测试训练算法的预测能力.

在统计分析中, 由于传统数据一般样本容量较小, 通常采用样本内模型检验法, 如模型拟合优度或模型设定检验. 然而, 如果采用样本内统计准则, 那么模型过度拟合的可能性将一直存在. 比如, 当解释变量个数增加时, 线性回归模型的 R 平方总会越来越大, 即使这些解释变量与因变量毫不相关. 一般来说, 增加模型复杂性可以提高拟合优度, 甚至在很多情况下最终总会通过样本内检验. 关于统计模型通常最终能够通过基于样本内残差的模型检验的讨论, 可参见文献 Breiman (2001). 更严重的是, 样本内统计建模与统计推断, 如果多次重复使用同一个样本数据, 有可能会导致所谓的数据窥视偏差 (data snooping bias), 其原因是同一个样本数据的多次重复使用可能导致统计显著性水平控制不当 (参见 Lo and MacKinlay (1990), White (2000)). 由于大数据样本容量通常较大, 因此可以使用样本外模型检验方法或交叉验证方法, 作为一个一般化的模型评估准则 (如 Varian (2014)). 样本外模型评估很重要, 因为误设模型一般不能很好地预测未来样本

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

或其他未知样本. 即使一个统计模型对训练数据而言设定正确, 但如果存在结构变化, 该模型对未来样本的预测可能不准, 或者如果训练数据和测试数据之间存在显著异质性, 该模型对其他样本的预测效果也可能不好. 样本外模型评估还可以有效降低数据窥视偏差. 总之, 样本外模型验证比样本内模型检验更严格更科学, 同时更适用于样本容量大的数据.

由于科技进步、偏好改变、政策变化和制度改革, DGP 可能会随着时间而改变. Lucas (1976) 指出, 理性的经济主体将正确预测政策变化的影响, 并相应调整他们的经济行为. 当 DGP 随时间而改变时, 只有最近的数据信息与 DGP 的现状密切相关; 遥远的旧数据则与 DGP 的现状越来越不相关, 对推断 DGP 的当下行为用处不大. 同样地, 由于经济主体之间存在异质性, 训练数据的经济主体可能无法代表测试数据的经济主体. 因此, 现有样本不能提供关于未来 DGP 或现有样本未涵盖的经济主体的信息. 实际上, 如果 DGP 随时间而改变, 任何样本在给定时间内, 无论信息多么丰富, 都无法包含未来总体的所有信息. 所以, 任何时间序列数据在给定时间内只能提供一个动态时变随机过程的信息子集, 而不是全样本信息. 因此, 在推断 DGP 的总体分布特征时, 统计抽样理论依旧有用, 而且适用于更一般的存在时变性或异质性的情况.

5.3 统计显著性与经济显著性

由于大数据的样本容量大, 我们可以探索大数据中可能存在的非线性、时变性、异质性等复杂结构, 这是机器学习能够比参数统计模型预测更精准的一个主要原因. 另一方面, 样本容量大也可能给统计建模与统计推断的习惯做法带来挑战. 比如, 对于样本容量不是很大的传统数据, 如果一个解释变量的参数估计量的 P 值根据预设显著性水平 (一般为 5%) 具有统计显著性, 那么通常认为该解释变量是重要变量. 现在假设有一个样本容量为 100 万的大数据, 模型的大部分解释变量可能都达到 5% 的显著性水平, 都具有统计显著性. 众所周知, 无论真实参数值多小 (只要不等于零), 随着样本容量不断增大, 统计显著性检验最终将会变为显著. 那么, 对于 100 万的样本容量, 恰好达到 5% 显著性水平的参数估计量意味着什么呢? 显然, 对于如此大的样本容量, 该参数值可能会非常接近 (但不等于) 零, 因此相应的解释变量可能在经济学上并不重要. 换句话说, 当样本容量非常大时, 具有统计显著性并不意味着具有现实重要性或经济重要性. 因此, 大数据的大样本容量使得传统的统计显著性检验变得不再适合 (Abadie et al. (2014, 2017)). 同时, 这也产生了一个新的问题: 当样本容量达到 100 万这么大时, 如何衡量解释变量的经济重要性呢? 我们需要合适的方法来判定解释变量的经济重要性, 而不是仅仅评估其统计显著性. 机器学习领域已提出各种判断特征重要性 (feature importance) 的方法, 其中所谓特征其实就是解释变量. 这些方法很多都不依赖于具体参数模型 (即 model-free). 参见 Liu, Zhong and Li (2015).

为了说明与模型无关的变量选择方法的重要性, 我们举一个简单例子. 假设因变量与某个解释变量真实的函数关系是非线性关系, 但是我们设定一个线性回归模型, 即模型误设. 很有可能这个解释变量的 t- 检验统计量在样本容量很大时也不具有统计显著性, 则依据线性回归模型的检验结果应该将该解释变量扔掉. 显然, 这将会导致所谓的遗漏变量问题.

上述分析表明, 当样本容量很大时, 只关注一个参数统计模型中的解释变量的统计显著性, 其实际意义并不太大. 更有意义的是关注模型选择, 特别是当存在高维潜在解释变量时, 可以通过比较不同的模型以显著提高拟合优度或预测精度, 这里所谓不同的模型既可以是指拥有不同的解

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

释变量集合 (参见 Breiman (2001)), 也可以是指不同的函数形式, 或者两者的混合. 换句话说, 对于大数据特别是胖大数据而言, 模型选择可能比解释变量的统计显著性更有助于改进对数据的拟合或预测效果. 与此同时, 高维解释变量的集合可能存在多重共线性或近似多重共线性, 根据某一统计准则 (如均方误差), 不同的解释变量集合可能会导致相同或相似的预测或拟合. 如果对数据进行 "微扰", 即增加或减少一小部分数据点, 便会导致最佳模型的显著改变. 这里, 抽样可变性导致最优模型的显著改变, 这称为模型不确定性 (model uncertainty). 因此, 在大数据时代, 我们可以预计, 统计分析将从参数估计不确定性过渡到模型选择不确定性或模型不确定本身.

5.4 模型多样性与模型不确定性

对于一个庞大数据, 高维解释变量的集合有很大的可能性存在多重共线性. 因此, 基于某一统计准则 (如均方误差), 不同的统计模型有可能呈现相似甚至相同的统计表现, 这称为模型多样性 (model multiplicity), 即不同模型的统计表现近似甚至相同 (参见 Breiman (2001)). 模型多样性可能与统计学关于 DGP 的模型唯一性假设并不矛盾. 一种情形是, 存在 DGP 的唯一模型设定, 但受限于数据证据和统计工具, 无法挑选出正确的模型, 所有统计模型都是对 DGP 的近似, 误设模型从不同方面刻画了 DGP 的特征, 但根据某个统计准则, 这些误设模型的表现近似甚至相同. 在经济学, 也可能同时存在多个经济模型能够解释同一经济现象, 有些模型甚至还会互相矛盾, 这称为模型模糊性 (model ambiguity). Hansen and Sargent (2001), Hansen et al. (2006) 研究了模型不确定性对经济主体的决策行为的影响. 当然, 也存在另外一种可能性, 即生成数据的 DGP 并不能用唯一模型设定来刻画. 举一个统计学的著名例子 —— 污染数据, 这些数据是由两个或两个以上不同的概率分布所生成的随机数的集合, 需要用一个混合概率分布来刻画. 在经济学中, 经济主体在不同状态下可能有不同的经济行为. 在这种情况下, 需要用一系列模型的 "组合" 来描述整个经济的运行, 其中每一个模型描述某个状态下的经济行为, 而这些模型的 "组合" 可由某种概率法则 (如马尔可夫链转移概率) 决定. 统计学和计量经济学一个著名的 "组合" 模型就是马尔可夫链转移模型 (参见 Hamilton (1989)).

基于同一统计准则, 对数据的 "微扰" 可能会导致最优统计模型的显著改变, 这种模型不确定性在实际应用中并不罕见, 与模型多样性密切相关. 另一方面, DGP 也可能会出现结构变化. 时间序列数据的每个时间段存在一个最佳预测模型, 但因为结构变化, 最佳预测模型会随着时间而改变, 这称为模型不稳定性 (model instability).

模型不确定性与模型不稳定性使得稳健统计分析变得格外重要. 在模型不确定性和模型不稳定性条件下进行统计建模与统计推断是大数据统计分析的一个新方向, 已经取得一些进展. 一般而言, 如果数据杂糅或者不同状态下存在不同经济行为, 那么模型平均 (model averaging) 或模型组合可能是最佳预测方法. 在预测领域 (如 Hansen (2007)), 已提出了用各种模型平均法或预测组合法来提高预测的稳健性和准确性, 这种想法至少可追溯到 Bates and Granger (1969) 的预测组合方法. 在机器学习领域, 为了克服模型不确定性带来的影响, Breiman (2004) 提出了随机森林方法, 通过计算机重复抽取产生一系列相关性不太强的随机样本, 对每个样本训练一棵决策树, 然后对所有决策树预测取平均以获取稳健预测.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

5.5 充分性原则、数据归约与维数约简

样本容量大并不是胖大数据的最重要特征. 对时间序列数据而言, 大数据的时间维度信息总是受到时间长短的限制 (当然, 实时或近乎实时的记录可以提供高频观测值). 然而, 如果大数据包含高维潜在解释变量的信息, 关于 DGP 的横截面信息就非常丰富. 当解释变量的数目多于样本容量时, 从统计学维数灾难的角度看, 胖大数据事实上是一个 "小样本". 因此, 需要发展新的统计降维方法以选择重要解释变量, 这其实是一种数据归约 (data reduction) 方法. 数据归约本质上是统计学充分性原则的一种方法, 为高维参数统计模型的有效推断提供了强大的分析工具. 统计分析就是寻找最有效的手段 (模型、方法、工具等) 从数据中总结、提取有价值的信息, 而充分性原则是从样本数据中总结信息的一个统计学基本原则. 充分统计量在统计推断中能够完全总结样本数据中所有的关于未知模型参数信息的低维统计量. 鉴于大数据的样本容量大、潜在解释变量的维度高以及信息密度低等特点, 统计充分性原则在大数据分析中将发挥十分重要的作用. 我们需要创新分析大数据的数据归约方法, 其中最重要的一种方法是变量降维 (dimension reduction), 特别是在胖大数据条件下的变量选择. 这种降维方法可视为机器学习方法在高维统计建模分析中的应用, 属于 "统计学习" (statistical learning) 的交叉领域.

在 "统计学习" 这一新兴的交叉领域, Tibshirani (1996) 提出 LASSO 方法, 可以在一个高维线性回归模型框架中挑选出重要解释变量并排除众多不相关的协变量. 简单地说, LASSO 方法的目标函数是最小化高维线性回归模型的残差平方和, 加上一个对高维回归模型维度的惩罚项. 这个惩罚项是所有回归系数的绝对和. 给定稀疏性 (sparsity) 假设, 即假设所有潜在解释变量中只有少数未知变量的系数不为零时, LASSO 方法及其拓展 (如 Fan and Li (2001), Zou (2006)) 能够在样本容量趋于无穷大时正确识别那些系数不为零的解释变量. 因此, LASSO 方法可视为在一个高维线性回归模型框架下统计推断和机器学习相结合的一种重要的变量选择方法. 从统计学的充分性原则看, 这本质上是一种数据归约. LASSO 方法在统计学与计量经济学领域拥有广泛的应用前景. 例如, 在 2SLS 和 GMM 估计中, 选择有效的工具变量一直是一个难点 (参见 Belloni et al. (2012)). 因此, 可以使用类似 LASSO 的方法从大量潜在工具变量中挑选出重要工具变量, 以改进 2SLS 和 GMM 估计效率. 又如, 高维方差 - 协方差的降维估计, 也可以通过拓展 LASSO 方法得以实现 (参见 Cui et al. (2020)). 事实上, 变量选择问题还可以拓展到高维非线性回归模型和高维非参数回归模型.

5.6 机器学习与非参数建模

如前文所言, 机器学习不用参数统计模型, 而是直接基于数据构建算法. 这些算法从训练数据中学习系统模式, 并基于这些系统模式进行预测. 许多情况下, 机器学习算法可以得到精准的样本外预测. 然而, 这些算法就像黑箱一样, 很难甚至无法解释为什么能够得到比较精准的样本外预测. 使用基于测试数据的泛化准则, 可以解释其中一部分原因, 但不能解释全部. 事实上, 机器学习算法类似于统计学的非参数分析方法. 不少重要的机器学习方法, 如决策树和随机森林, 最早是由统计学家首先提出来的. 与参数统计建模方法不同, 非参数方法不对 DGP 的结构或总体分布假设任何具体的函数形式, 而是让数据告诉合适的函数形式. 非参数方法关注对数据的拟合优度, 如最小化残差平方和, 同时也顾及拟合函数的平滑性 (如二阶连续可导), 最终通过选择

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

一个平滑参数 (smoothing parameter) 使均方误差中的方差和平方偏差达到均衡, 这样便可一致估计关于 DGP 的未知函数, 如回归函数或概率分布函数. 许多机器学习方法具有很强的非参数方法的特征, 加上使用基于测试数据的泛化准则, 非参数分析可以从理论上解释为什么很多机器学习方法在大数据条件下能够取得较好的预测效果. 例如, Lai (1977) 通过推导 k 最近邻法 (k-NN) 均方误差中的方差和平方偏差的收敛速度, 证明当整数 k 随着样本容量 n 的增加而增加, 但增加速度比 n 慢时, k 最近邻法可以一致估计未知回归函数. Breiman (2004) 证明, 假设 DGP 存在唯一的未知概率分布, 而数据由独立分布的随机样本遵循未知概率分布生成, 那么如果决策树的节点数量随着样本容量的增加而增加, 但其增加的速度比样本容量慢, 则决策树可以一致估计 DGP 的未知概率函数. Biau, Devroye and Lugosi (2008), Scornet, Biau and Vert (2015) 证明了随机森林可以一致估计未知回归函数. White (1989, 1992) 则严格证明了人工神经网络估计的一致性, 前提是假设隐藏层的数量随着样本容量的增加而增加. 人工神经网络是模仿人类认知过程的一个非参数模型, 如果其复杂性随样本容量的增加而增加, 最终可以一致估计出未知回归函数. 实际上, 就变量选择而言, 许多机器学习算法比典型的非参数方法更灵活. 对于非参数分析, 由于臭名昭著的 "维数灾难" 问题, 需要事先给定解释变量, 而且这些解释变量的维度不能太大, 否则在实际中无法应用. 相比之下, 机器学习经常面对大数据中高维的潜在解释变量, 其维度很大甚至超过数据的样本容量, 机器学习可以通过合适算法快速 "穷尽" 所有合适的解释变量子集, 为最佳预测挑选出一个低维的重要解释变量集合. 这是机器学习比非参数方法更有优势的一点.

统计建模与机器学习的交叉融合是大数据分析的一个重要发展趋势. 一方面, 没有机器学习, 无法想象如何分析海量大数据. 另一方面, 大数据是我们能够 "教" 机器而不用直接为它们编程的主要原因之一. 大数据的可获得性使得训练机器 "学习" 模式成为可能. 相对于参数统计模型, 机器学习算法的难点之一是缺乏可解释性, 这是因为机器学习方法直接基于数据构建算法而非用参数建模. 相反地, 统计推断大多采用参数建模. 严格地说, 一个统计参数模型只能刻画数据与 DGP 的一些总体特征, 但通常并非全部总体特征 (除非模型正确设定). 因此, 统计参数模型所刻画的证据其实是模型证据 (model evidence), 与直接基于数据的机器学习所刻画的证据存在一定差别. 由于其灵活性与一般性, 机器学习所刻画的证据将比较接近数据原有的证据, 即数据证据 (data evidence). 模型证据与数据证据之间的差别, 对我们在解释统计推断特别是参数假设检验的实证结果时, 非常重要. 例如, 使用一个 p 阶线性自回归模型验证金融市场有效性假说时, 如果我们基于观测数据发现所有自回归系数均为零, 这并不意味着市场有效性原假说是正确的, 因为线性自回归模型只是众多预测金融市场方法中的一种, 很有可能收益率数据存在可预测的成分, 但是需要使用非线性模型. 由于机器学习与非参数方法一样, 并不依赖某一个特定的统计模型, 因此机器学习发现的证据将比较接近数据证据, 从而避免参数统计模型的缺点.

5.7　相关性与因果关系

曾经有一种观点, 认为大数据分析只需要相关性, 不需要因果关系. 之所以产生这个论断, 一个主要原因是在大数据条件下, 有很多实时或高频数据, 而基于实时或高频数据的预测主要是依靠相关性, 而不是因果关系. 然而, 很多情况下, 经济因果关系在高频或实时条件下可能还无法充分显示出来, 所以不需要因果关系的论断是不对的, 至少不适用于经济学. 在许多实际应用中, 机

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

器学习方法, 如决策树、随机森林、人工神经网络、深度学习等, 基于数据的系统特征与统计关系 (如相关性) 确实可以进行精准的样本外预测. 然而, 经济研究的主要目的是推断经济系统中经济变量之间的因果关系, 揭示经济运行规律. 比如, 在信用风险管理中, 大数据分析可以帮助查明信用风险的根本原因, 尽早发现可能的欺诈行为以防止金融机构遭受损失, 这些都需要分析大数据背后的因果关系. 在大数据时代, 经济因果关系依旧是经济学家与计量经济学家在经济学实证研究中的主要目的. 信息技术, 尤其是互联网、移动互联网与人工智能, 从根本上改变了人类的生产方式与生活方式, 但它们没有改变经济学因果推断的目的. 在过去 20 年, 计量经济学诞生了一门新兴学科, 即政策评估计量经济学 (econometrics of program evaluation), 研究非实验条件下经济因果效应的识别与测量. 所谓因果关系是指在所有其他变量 (如控制变量 Z) 不变的条件下, 改变一个变量 (如政策干预 X) 是否会导致另一个变量 (如经济结果 Y) 的改变. 如果有, 则称存在从 X 到 Y 的因果关系. 在实验科学中, 要识别因果关系或检验一个政策干预的效应, 可以将实验主体随机分为两组, 一组是实验组, 接受实验干预, 另一组是控制组, 不接受实验干预, 其他条件或变量则保持不变. 干预效应是两组在同等条件下的结果之差. 在计量经济学中, 当评估政策效应时, 由于经济系统的非实验性特点, 往往无法进行控制实验, 尤其是无法确保实验组与控制组满足 "同等条件" 假设. 统计学和计量经济学关于政策评估的基本思想是, 在同等条件下, 比较实施了该政策的观测结果与假设没有实施该政策的虚拟事实. 在已实施某个政策的现实情况下假设这个政策没有实施, 显然是一种虚拟假设, 该虚拟假设下的经济结果常称为虚拟事实 (counterfactuals). 由于虚拟情况不会真正发生, 故需要对虚拟事实进行估计, 这实质上是一种预测. 这可以借助一个统计模型来估计, 也可以通过机器学习来预测. 鉴于机器学习精准的预测能力, 机器学习有望精准估计虚拟事实, 从而精确识别与测度经济因果关系. 换句话说, 虽然机器学习不能直接揭示因果关系, 但它可以通过准确估计虚拟事实帮助精确识别与测度因果关系. 关于因果推断, 可参见 Pearl (2009), Varian (2016).

5.8 新型数据建模

除了非结构化数据 (如文本、图像、音频、视频数据等), 大数据包括很多新型的结构化数据. 例如, 函数数据就是一种新型数据, 而大家比较熟悉的面板数据 (参见 Hsiao (2014)) 是函数数据的一个特例. 函数数据的例子还有很多, 如一天内温度是时间的函数; 每个交易日从开盘到收盘, 股票价格是时间的函数; 从 1 岁到 15 岁, 女孩每月测量的身高是时间的函数. 另一种新型数据是区间数据 (interval-valued data), 即某个变量取值的范围. 相对点数据 (point-valued data) 来说, 区间数据包含更多关于变量的水平和变化范围的信息. 区间数据在现实生活中并不少见, 如病人每天的最高血压与最低血压、每天天气的最高温与最低温、每天股票的最高价与最低价、金融资产的买卖差价等, 均构成区间数据. 也可以通过结合多个原始数据得到区间数据, 如某行业男性员工与女性员工的平均工资、农村家庭与城镇家庭的平均收入. 区间数据是符号数据 (symbolic data) 的一个特例, 符号数据是更一般化的数据形式.

新型数据比传统点数据包含更多信息. 很多情况下, 人们一般是将这些新型数据转换为点数据, 然后使用传统的计量经济学模型与方法进行分析. 但是, 将新型数据转换为点数据, 通常伴随着信息损失. 因此, 直接对这些新型数据进行建模比先将它们转化为传统点数据再建模更有价

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

值. 新型数据需要新的统计模型与统计方法. 在这方面, 统计学和计量经济学已产生了一些原创性成果, 如函数数据分析 (functional data analysis) 和区间数据建模. 关于函数数据分析, 可参见文献 Horváth and Kokoszka (2012), 而关于区间数据建模, 可参见 Han et al. (2018), Sun et al. (2018).

6 总结

本文讨论了大数据与机器学习给统计科学的理论与应用带来的影响、挑战和机遇. 首先, 尽管大数据正在改变基于统计显著性的统计建模和统计推断的传统做法, 但大数据并没有改变从随机抽样推断总体分布特征的统计思想. 重要的统计学原则, 如抽样推断、充分性原则、数据归约、变量选择、因果推断、样本外预测等基本统计思想, 在大数据分析上依旧适用, 一些统计学方法如充分性原则甚至因为大数据的出现而变得更加重要, 但其具体的方法与表现形式需要有所创新. 其次, 大数据允许放松统计建模的一些基本假设, 如模型唯一性、正确设定与平稳性, 从而扩大了统计建模与统计推断的应用范围. 再次, 大数据, 尤其是非结构化数据, 带来了很多传统数据不具备的有价值的信息, 大大拓展了实证研究的范围与边界. 最后, 新型数据也催生了新的统计模型与方法.

机器学习是伴随大数据和云计算的产生而广泛兴起的大数据分析方法. 它是计算机自动算法, 通过学习训练数据的系统特征与统计关系而对未知样本进行预测, 这与统计学由抽样推断总体的思路一致. 机器学习与数理统计学拥有相同的随机概率基础, 但它不假设 DGP 的结构或概率分布满足具体的函数或模型形式, 而是通过计算机算法从训练数据中学习数据的系统特征与变量之间的统计关系, 实现样本外预测与分类. 机器学习算法通常以精准的样本外预测著称, 但它们经常就像黑箱一样, 很难甚至无法解释. 然而, 很多重要的机器学习方法, 如决策树、随机森林、k 最近邻法、人工神经网络以及深度学习, 与非参数分析的基本思想一致或非常类似. 因此, 可以从非参数方法的视角、从统计理论上说明为什么机器学习方法在大数据和使用泛化准则条件下可以获得精准的样本外预测. 机器学习与统计建模相结合催生了一个新的交叉领域, 即统计学习. 比如, 统计学习中的 LASSO 方法及其拓展就是一种强大的变量选择方法, 它可以在一个高维线性回归模型框架内, 正确挑选出重要的解释变量, 并排除大多数不相关的变量. 统计学和计量经济学中存在很多高维建模与数据归约难题, 这些难题有望通过借鉴、应用与创新机器学习的方法加以解决.

参　考　文　献

洪永淼, (2007). 计量经济学的地位、作用和局限 [J]. 经济研究, (5): 139–153.
Hong Y M, (2007). The Status, Roles and Limitations of Econometrics[J]. Economic Research Journal, (5): 139–153.

Abadie A, Athey S, Imbens G W, Wooldridge J M, (2014). Finite Population Causal Standard Errors[R]. Working Paper, National Bureau of Economic Research.

Abadie A, Athey S, Imbens G W, Wooldridge J M, (2017). When Should You Adjust Standard Errors for Clustering[R]. Working Paper, National Bureau of Economic Research.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Baker M, Wurgler J, (2007). Investor Sentiment in the Stock Market[J]. Journal of Economic Perspectives, 21(2): 129–152.

Baker S R, Bloom N, Davis S J, (2016). Measuring Economic Policy Uncertainty[J]. Quarterly Journal of Economics, 131(4): 1593–1636.

Bates J M, Granger C W, (1969). The Combination of Forecasts[J]. Journal of Operational Research Society, 20(4): 451–468.

Belloni A, Chen D, Chernozhukov V, Hansen C, (2012). Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain[J]. Econometrica, 80(6): 2369–2429

Biau G, Devroye L, Lugosi G, (2008). Consistency of Random Forests and Other Averaging Classifiers[J]. Journal of Machine Learning Research, 9: 2015–2033.

Bok B, Caratelli D, Giannone D, Sbordone A M, Tambalotti A, (2017). Macroeconomic Nowcasting and Forecasting with Big Data[R]. Staff Repots 830, Federal Reserve Bank of New York.

Breiman L, (2001). Statistical Modeling: The Two Cultures[J]. Statistical Science, 16(3): 199–231.

Breiman L, (2004). Consistency for a Simple Model of Random Forests[R]. Technical Report 670, Statistical Department, University of California at Berkeley.

Chan J T, Zhong W, (2018). Reading China: Predicting Policy Change with Machine Learning[R]. AEI Working Paper 998561, American Enterprise Institute.

Cox D R, (1972). Regression Models and Life-tables[J]. Journal of Royal Statistical Society: Series B (Methodological), 34(2): 187–202.

Cui L, Hong Y, Li Y, Wang J, (2020). Large Positive Definite Covariance Estimation for High Frequency Data via Sparse and Low-rank Matrix Decomposition[R]. Working Paper, City University of Hong Kong.

Engle R F, (2000). The Econometrics of Ultra-High-Frequency Data[J]. Econometrica, 68(1): 1–22.

Engle R F, Russell J R, (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data[J]. Econometrica, 66(5): 1127–1162.

Fan J, Li R, (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties[J]. Journal of American Statistical Association, 96(456): 1348–1360.

Giannone D, Reichlin L, Small D, (2008). Nowcasting: The Real-time Informational Content of Macroeconomic Data[J]. Journal of Monetary Economics, 55(4): 665–676.

Granger C W J, (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods[J]. Econometrica, 37(3): 424–438.

Hamilton J D, (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle[J]. Econometrica, 57: 357–384.

Han A, Hong Y, Wang S, (2018). Autoregressive Conditional Interval Models for Time Series Data[R]. Working Paper, Department of Economics, Cornell University.

Hansen B E, (2007). Least Squares Model Averaging[J]. Econometrica, 75(4): 1175–1189.

Hansen L P, Sargent T J, (2001). Robust Control and Model Uncertainty[J]. American Economic Review, 91(2): 60–66.

Hansen L P, Sargent T J, Turmuhambetova G, Williams N, (2006). Robust Control and Model Misspecification[J]. Journal of Economic Theory, 128(1): 45–90.

Horváth L, Kokoszka P, (2012). Inference for Functional Data with Applications[M]. Berlin: Springer Science & Business Media.

Hsiao C, (2014). Analysis of Panel Data[M]. Cambridge: Cambridge University Press.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Lai S L, (1977). Large Sample Properties of k-Nearest Neighbor Procedures[D]. Los Angeles: University of California.

Liu J, Zhong W, Li R, (2015). A Selective Overview of Feature Screening for Ultrahigh-dimensional Data[J]. Science China Mathematics, 58(10): 1–22.

Lo A W, MacKinlay A C, (1990). Data-snooping Biases in Tests of Financial Asset Pricing Models[J]. Review of Financial Studies, 3(3): 431–467.

Lucas R E, (1976). Econometric Policy Evaluation: A Critique[J]. Carnegie-Rochester Conference Series on Public Policy, 1(1): 19–46.

Pearl J, (2009). Causality: Models, Reasoning and Inference[M]. Cambridge: Cambridge University Press.

Samuel A L, (1959). Some Studies in Machine Learning Using the Game of Checkers[J]. IBM Journal of Research and Development, 3(3): 210–229.

Scornet E, Biau G, Vert J P, (2015). Consistency of Random Forests[J]. Annals of Statistics, 43(4): 1716–1741.

Shiller R J, (2019). Narrative Economics: How Stories Go Viral and Drive Major Economic Events[M]. Princeton: Princeton University Press.

Sun Y, Han A, Hong Y, Wang S, (2018). Threshold Autoregressive Models for Interval-valued Time Series Data[J]. Journal of Econometrics, 206(2): 414–446.

Tibshirani R, (1996). Regression Shrinkage and Selection via the Lasso[J]. Journal of Royal Statistical Society: Series B (Methodological), 58(1): 267–288.

Varian H R, (2014). Big Data: New Tricks for Econometrics[J]. Journal of Economic Perspectives, 28(2): 3–28.

Varian H R, (2016). Causal Inference in Economics and Marketing[J]. Proceedings of National Academy of Sciences, 113(27): 7310–7315.

White H, (1989). Some Asymptotic Results for Learning in Single Hidden-layer Feedforward Network Models[J]. Journal of American Statistical Association, 84(408): 1003–1013.

White H, (1992). Artificial Neural Networks: Approximation and Learning Theory[M]. Oxford: Blackwell Publishers.

White H, (2000). A Reality Check for Data Snooping[J]. Econometrica, 68(5): 1097–1126.

Zou H, (2006). The Adaptive Lasso and Its Oracle Properties[J]. Journal of American Statistical Association, 101(476): 1418–1429.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net