大数据时代统计学面临的机遇与挑战*

耿 直

内容提要: 大数据给统计学带来了机遇、挑战和紧迫感。本文描述大数据的环境,利用大数据的目的和大数据带来的变革; 介绍国内外有关大数据的研究动向; 探讨大数据包含的信息,大数据的预处理、抽样和分析方法。

关键词: 大数据; 抽样; 数据分析方法论
中图分类号: C829.2　　文献标识码: A　　文章编号: 1002 - 4565(2014)01 - 0005 - 05

Opportunities and Challenges in the Age of Big Data for Statistics

Geng Zhi

Abstract: Big data brings opportunity, challenges and urgency for statistics. This paper describes the environments of big data, the goal of big data and the revolution by big data. And it also introduces the research trends for big data at home and abroad. The information, preprocess, sampling and analysis of big data have been discussed.

Key words: Big Data; Sampling; Methodology of Data Analysis

一、引言

在人类利用观察认知自然的方法论发展的历程中,最初神学、哲学和科学合为一体,巫术、占星术和宗教是哲学和科学的前身。人类旧石器期用神学解释自然,根据蛙鸣预测下雨,用巫术和占星术祈祷、预测和干预自然。中国古代利用阴阳太极图和八卦图作为思辨模型,分析和解释自然和人文社会的现象。古希腊文明孕育了演绎逻辑、归纳逻辑。文艺复兴前后哲学家提出观察和试验的方法,培根(F Bacon, 1620, 新工具) 提出通过观察实验,运用三表法: 存在与具有表、差异表、程度表。穆勒(J S Mill, 1843, 逻辑体系) 提出归纳四法: 求同法、求异法、共变法、剩余法。在统计方法论的发展中,贝叶斯(T Bayes, 1764) 提出逆概率方法,利用观察结果推断事件的概率。高尔顿(F Galton, 1886) 将变量间的相关关系进行了形式化,提出了相关系数,进一步在相关性的基础上提出了回归预测的方法。皮尔逊(K Person, 1900) 提出了拟合优度检验的方法,使得人们能够利用概率度量观测现象与科学假说的拟合程度。在此后的一个多世纪中统计方法有了突飞猛进的发展,被广泛地应用到自然科学、经济金融和人文社会科学的各个领域。在人类利用观察探索自然和社会规律的历程中,从远古时代的观察加臆想,古希腊时代的观察加理性推理,文艺复兴时代的试验加理性推理,直到现代的抽样加统计模型。

当今时代,一方面人们在主动地获取数据。各个科学领域都在大量地获取数据,自然科学领域收集着从宏观的天文数据到微观的基因数据,经济、金融和人文社会科学收集着大量的观察和调查数据。一些人们在通宵达旦地制造和收集数据,他们相信这些数据会对别人有用。也有一些人们脱离了实验室,仅依靠网络数据从事研究。另一方面人们在被动地囤积数据。随着计算机互联网、搜索引擎、电子商务、多种传感器和多媒体技术的发展和广泛使用,各种形式的数据如江河流水般地涌来。当今数据的获取和规模发生了根本的变化,统计学面临着新的机遇和挑战,需要在方法论上有所突破。

本文在第二部分描述大数据的形式和环境,以及利用大数据的目的; 第三部分描述大数据带来的变革; 第四部分介绍国内外有关大数据的研究动向; 第五部分探讨大数据的信息问题; 第六部分介绍大

* 本文获国家自然科学基金项目“因果推断的统计方法”(批准号 11171365) 和“生物统计”(批准号 11331011) 的资助。
本文为第十七次全国统计科学讨论会特邀论文。

数据需要的预处理、抽样和分析方法，特别地介绍了网络图模型对大数据分析的潜在用途；最后一部分是结束语，讨论大数据给统计学带来了机遇、挑战和紧迫感。

二、大数据及其目的

狭义地讲，大数据是一个大样本和高维变量的数据集合。针对样本大的问题，统计学可以采用抽样减少样本量，达到需要的精度。关于维数高的问题，需要变量选择、降维、压缩、分解。但认知高维小样本存在本质的困难。广义地讲，大数据涵盖多学科领域、多源、混合的数据，自然科学、人文社会、经济学、通讯、网络、商业和娱乐等各领域的数据集相互重叠连成了一片数据的海洋。各学科之间数据融合和贯通，学科的边界已重叠和模糊。大数据涉及各种数据类型，包括文本与语言、录像与图像、时空、网络与图形。我认为当代的大数据不仅数据量大，还包括多种类型数据和大量数据项目集的覆盖重叠。

大部分传统的统计方法只适合分析单个计算机存储的数据。而目前大数据的环境包括了[2]：

1. 数据流环境：数据快速不断涌来，现有存储设备和计算能力难以应付这种洪水般的数据流；

2. 磁盘存储环境：数据已不能完全存储在内存中，需要硬盘存储；

3. 分布存储环境：数据分布存储在多个计算机中；

4. 多线条环境：数据存储在一个计算机中，多个处理器共享内存。

大数据的目的是将数据转化为知识（Big Data to Knowledge，BD2K），探索数据的产生机制，进行预测和制定政策[2,6,7]。把信息转变为有用的知识还需漫长的时间[3]。“预测”不同于“制定政策”。一个儿童的鞋子越大，可以预测他掌握的词汇量越多；但是，制定政策强制他穿大鞋子并不能提高他的词汇量。

进一步，大数据有记录保存自然与社会现状的作用。现在有些人收集着大量数据，尽管他们还不清楚如何分析这些数据，但是他们相信需要保存现今社会和经济高速发展的过程，期待着今后分析和解释这段历史。还有些人将百岁老人的血液和其他各种生物标本等存放在冰箱里，他们认为当今的技术还不足以测试和分析这些资源，期待着今后更先进的测试技术。大数据就如同自然和社会的血液那样记录着社会的现状和发展过程。

17 世纪望远镜和显微镜的发明使人类看到了以前从来没有看到过的宇宙空间和微生物，扩大了人类对自然的认识。大数据就像“望眼镜”和“显微镜”那样，使得人们能够通过数据观察分析丰富的自然、经济、社会的现象。借助互联网数据，可以及时了解疾病疫情、科学动态、社会动态。谷歌借助频繁检索词条能及时判断流感从哪传播，哪些人可能感染了流感[6]。大数据将形成自然和人文社会的历史长河，不但能用于探索当代的科学问题，将来可以用于研究人们食用转基因食品对子孙后代的影响等追踪研究问题，为未来留下现今的历史资料。

三、大数据带来的变革

大数据给我们的时代带来了变革。目前，人们习惯于根据“研究问题”来驱动“收集数据”。今后，大数据到处可得，人们将会用“数据”驱动“研究问题”。就像我们出远门前常常查询目的地的天气、交通和宾馆那样，未来人们在研究和决策前将会通过查询数据做决定。目前已经有科学家开始使用软件搜索和汇总已发表论文中的成果。古希腊文明时代哲学家是百科全书式的人物，文艺复兴以来学科的分裂和细化产生了单一学科的专家。大数据时代，将可能再次出现百科全书式的人物。大数据将减弱专家在各个领域的影响，甚至导致专家的消亡[6]。例如目前已经有很多统计学家、物理学家和计算机专家凭着他们对数据的敏锐和处理能力进入了生命科学领域。假若我们有了成千上万本中文书和它们的阿拉伯语译本，即使我们不懂阿拉伯语，我们也能采用匹配文本的方法将中文翻译成阿拉伯语。谷歌机器翻译团队并不会说他们翻译出的语言[6]。大数据中包含有各种不同目的的数据集，综合利用它们可以做出原来目的之外的意外成果。例如，将医院病历数据与信用卡消费数据结合，我们能发现食品与健康的相关关系，指导人们进行健康饮食。假若再加上手机和 GPS 等数据，还能随时对人们进行体检，指导健身，减少猝死，帮助医生诊断疾病等，应用大数据可以设想的用途不计其数。

四、大数据的研究动向

美国科学院全国研究理事会的“大数据分析委

员会[2]提出了大数据的挑战问题: 处理高度分布的数据资源,追踪数据的来源,核实数据,处理样本偏倚和异质性,处理不同格式和结构的数据,开发并行和分布式算法,数据的完整性、安全性、一体化和共享,开发大数据的可视化方法和可扩展算法,处理实时分析和决策制定。美国国家卫生研究院(NIH) [7]提出将数据转换为知识(BD2K) 的使命,设立生物医学大数据研究中心基金。

在我国,科技部组织召开了两次香山会议: 2012 年 5 月"大数据科学与工程",2013 年 5 月"数据科学与大数据的科学原理及发展前景"; 设立了大数据的"973"专项研究计划。国家自然科学基金委 2013 年 3 月组织了双清论坛"大数据技术与应用中的挑战性科学问题",国家自然科学基金委的数理学部、信息学部、管理学部都计划设立大数据的重大项目和重点项目群,国家社科基金计划设立大数据的重大项目。另外,业界、咨询公司和投资商都在寻找大数据的机会。

大数据分析的方法研究超出了单一学科领域,需多学科联合。统计学家需要关注计算机资源和实时决策。计算机学家需要了解统计推断和算法的复杂性。另外,利用大数据还需要相关领域专家的专业知识。

五、大数据的信息和问题

大数据是多源异质的、覆盖不同范围的数据。为了融合各种数据,需要对数据来源、数据的获取方式和数据描述进行形式化,以支撑数据分析。大数据来自多种渠道,存在抽样偏倚、随机的和非随机的误差、无意的和有意的错误。数据收集的准则与数据分析和决策的准则不相符合,有些数据不是原始数据,而是推断的结果(如填补的缺失数据) ,数据的循环使用导致偏差和噪音被放大。数据量大不一定有用的信息多,大量的含偏差数据甚至会破坏信息。应意识到分析大数据也许会得到虚假知识,而自己却不知情。在大数据环境下,收集数据的人也许不清楚未来使用数据的人要做什么; 使用数据建模的人也许不清楚数据是如何得到的; 使用模型的人也许不知道模型是从什么数据得出来的。因此,难免人们会根据自己的意图过分地解释模型,超出了原始数据所包含的信息范围。

获取的数据也可能存在选择偏倚,如医院就诊的病人和使用互联网的人不能代表研究总体。大数据难免存在不响应和缺失数据,有些数据是随机缺失的、非随机缺失的,因为敏感问题或隐私问题而缺失的。不同研究收集不同的、有重叠变量的数据集。例如: 经济、保险、社会、医学、生物、化学等研究的问题不同,收集数据的变量集合不同(X, Y, Z) 、(X, Y, U, V) 等,但是它们也许有共同感兴趣的变量交集。

另外,来自观察的数据和来自试验的数据具有不同的信息,不同信息导致不同的认知范围。数据本身含有的信息是有边界的,决定了数据分析解释的范围。模型只是数据信息的精练,不能向外延展数据的信息。利用模型进行超出数据信息之外的推断,需要额外的假定,而这些假定不能被数据证伪。

六、大数据的处理、抽样与分析

(一) 数据的预处理

大数据的预处理包括数据清洗、不完全数据填补、数据纠偏与矫正。利用随机抽样数据矫正杂乱的、非标准的数据源。统计机构的数据是经过严格抽样设计获取的,具有总体的代表性和系统误差小的优势,但是数据获取和更新的周期长,尽管调查项目有代表性,但难以无所不包。而互联网数据的获取速度快、量大、项目繁细,但是难以避免数据获取的偏倚性。将统计机构的数据作为金标准和框架对互联网数据进行矫正,将互联网数据作为补充资源对统计机构的数据进行实时更新,也许是解决问题的一个思路。研究利用多源数据的重叠关系整合多数据库资源的方法,多种专题(panels) 的数据可以相互联合,实现单一专题数据不能完成的目标。

(二) 大数据环境的抽样

大数据的抽样方法有待研究[2]。"样本"不必使用所有"数据",不管锅有多大,只要充分搅匀,品尝一小勺就知道其滋味。针对大数据流环境,需要探索从源源不断的数据流中抽取足以满足统计目的和精度的样本。需要研究新的适应性、序贯性和动态的抽样方法。根据已获得的样本逐步调整感兴趣的调查项目和抽样对象,使得最近频繁出现的"热门"数据,也是感兴趣的数据进入样本。建立数据流的缓冲区,记录新发生数据的频数,动态调整不在样本中的数据进入样本的概率。对于罕见案例,如果采用简单随机抽样将会抽到很少的案例和过多的非案例数据。需要研究大数据的案例抽样方法

(Case-based sampling)。探索基于事件的抽样方法(Event-based sampling)。设置信号强度门槛值,仅抽样超过门槛值的数据。利用其他各种抽样技术,例如:捕获—再捕获,不等概率抽样,将注意力放到总体中难以观测到的部分。在大数据环境中采用非随机抽样方法,如滚雪球方法,从种子开始逐步扩大样本。研究对社会关系网络和图的抽样方法,从随机种子出发,不断加入新种子,了解网络性质和结构。需要研究发现稀疏信号的方法和压缩感知方法。成组检测是发现稀疏信号的一个特别方法。例如美国 1943 年对新兵验血检查梅毒感染时,由于梅毒是罕见疾病,采用了将一组人群的血液混合在一起进行检测的高效快捷方法。

(三)大数据的分析与整合

针对大数据的高维问题,需要研究降维和分解的方法。探讨压缩大数据的方法,直接对压缩的数据核进行传输、运算和操作。除了常规的统计分析方法,包括高维矩阵、降维方法、变量选择之外,需要研究大数据的实时分析、数据流算法(data stream computing)。不用保存数据,仅扫描一遍数据的数据流算法,考虑计算机内存和外存的数据传送问题、分布数据和并行计算的方法。如何无信息损失或无统计信息损失地分解大数据集,独立并行地在分布计算机环境进行推断,各个计算机的中间计算结果能相互联系沟通,构造全局统计结果。研究多个数据资源的融合算法。研究利用数据流寻找模型变化时间点的动态变化模型。

针对多种不同数据库的环境,利用关系数据库技术,根据关键字(例如,身份证)将很多小数据库连接成一个大数据库。另一方面,能无信息损失地将大数据库拆分为多个小数据库。组合多数据库的不同数据集合,可以做出有创意的东西。丹麦有一个手机用户的数据库,共 358403 人。另一个记录了癌症患者的数据库,有 10729 名中枢神经系统患者的信息。将两个数据库结合,研究手机与癌症之间的关系。发现使用手机和癌症之间不存在任何关系,其结果发表在 2011 年的《英国医学杂志》[6]。

在大数据环境,很多数据集不再有标识个体的关键字,传统的关系数据库连接方法不再适用,需要探讨利用数据库之间的重叠项目来结合不同的数据库,利用变量间的条件独立性整合多个不同变量集的数据为一个完整变量集的大数据库的方法。探索不必经过整合多数据库,直接利用局部数据进行推断和各推断结果传播的方法。另一方面,利用统计性质无信息损失地分解和压缩大数据。

在多源和多专题的数据库环境,各个数据集的获取条件不同,项目不同又有所重叠。在这种情况下,一种分析方法是分别利用各个数据集得到各自的统计结论,然后整合来自这些数据集的统计结论,如荟萃分析方法。我们曾提出“中间变量悖论”,指出统计结论不具备传递性[7]。例如,变量 A 对变量 B 有正作用,并且变量 B 对变量 C 有正作用,但是可能变量 A 对变量 C 有负作用。为了避免“中间变量悖论”的现象发生,可以先整合数据,再利用整合的数据进行推断。我们提出了判断已有的各种条件数据集是否能识别所有变量联合分布的算法[5]。例如,有 5 个数据库,包含的变量的模式为 [D,E,F|A,B,C,G], [A,D,G|B,C], [D,E|F], [B|A,C,D,G], [C|D,E],一个字母表示一个变量。[D,E|F] 表示在给定变量 F 条件下获得的变量 D 和 E 的数据。根据我们的算法可以判断由这些条件数据库可以识别和估计所有变量 [A,B,C,D,E,F,G] 的联合分布。

(四)网络图模型

网络图模型用图的结构描述高维变量之间的相互关系,包括无向图概率模型、贝叶斯网络、因果网络等[8]。网络图模型是处理和分析高维大数据和多源数据库的有效工具,目前已经有丰富的图模型的软件系统,例如 MSBN, BN Toolbox, WinBUGS, Hugin, Tetrad, MIM, CoCo 等。无向图模型利用有或无一条无方向边来描述变量之间的关联关系和条件独立性,可以将高维变量的统计推断问题(例如参数估计和假设检验)分解为低维变量的统计推断问题。贝叶斯网络是一个有向无环图,用于计算大网络中信息的收集和传播。在一个由众多变量作为结点的大网络中,当收集到一部分变量的信息后,不用计算高维联合概率,而是采用网络传播信息流的方法有效地计算目标变量的后验概率。Pearl(2011 年图灵奖获得者)提出因果网络,采用有向图刻画变量间的因果关系,利用数据学习网络结构,发现产生数据的机制和因果关系网络[8]。

网络图模型可以用于分解大数据集合,处理多源数据库,利用局部数据,进行并行计算。网络图模型还可以引入隐变量简化复杂的关联关系。利用关

联网络图进行基于关联关系的预测,例如,朴素贝叶斯分类器和贝叶斯网络分类器。利用因果网络图可以进行基于因果关系的预测和政策制定。

我们提出“盲人摸象”方法,利用多个不完全数据库学习整体网络结构的算法[10] [11]。首先分别利用各个数据库学习各自的局部网络结构,然后将这些局部结构相互交流配合,最终整合一个全局的网络结构。当因果关系不能完全根据数据确定时,我们提出采用主动学习的方法,抓住主要变量进行干预试验,确定整个网络的因果关系,达到“壹引起纲,万目皆张”的作用[4]。我们提出利用因果关系制定干预政策的“寻根问题 + 顺誉摸瓜”的方法[12]。这个方法不必构造高维变量的完整因果网络,而是从一个目标结点出发,逐步进行局部变量选择和局部网络结构学习,最终确定并能区分该目标节点的原因与结果。

七、结束语

一个新生事物的出现将必定导致传统观念和技术的革命。数码照相机的出现导致传统相片胶卷和影像业的已近消亡。如果大数据包含了所有父亲和儿子的身高数据,只要计算给定的父亲身高下所有儿子的平均身高就可以预测其儿子身高了。模型不再重要,当年统计学最得意的回归预测方法将被淘汰。大数据的到来将对传统的统计方法进行考验。统计学会不会象科学哲学那样,只佩戴着历史的光环,而不再主导和引领人们分析和利用大数据资源。现在其他学科和行业涌入大数据的热潮,如果统计学不抓紧参与的话,将面临着被边缘化的危险。

现今统计学的目标是通过获取数据和分析数据发现真理(总体的参数和性质),统计方法和理论对数据有过高的要求。而大数据充满了各种随机的、非随机的误差和偏倚,不能满足这些苛刻的要求。按照波普的科学划界准则,只要我们能从大数据中提炼出具有可证伪的结论,那么这个结论还是科学的,可以用于知识积累。这些可证伪的大数据结论可作为进一步科学研究的假说,以数据驱动研究。

我们在看到大数据给统计学带来了机遇的同时,也应该看到现在的统计方法普遍只适用于全部数据放在单个计算机内存的环境,分布式大数据和数据流的环境给统计学带来了挑战。统计学家不应该固守传统数据的环境,必须积极学习新生事物,适应新的大数据环境,扩展统计学的应用领域,创造出迎合大数据的新统计方法,“机遇”与“挑战”并存。

参考文献

[1] Chen H, Geng Z, Jia J. Criteria for surrogate end points [J]. J Royal Statist Soc Ser,2007, B 69:919 – 932.

[2] Committee on the Analysis of Massive Data et al. (2013) Frontiers in Massive Data Analysis [J]. National Academies Press, Washington. http://www.nap.edu/catalog.php? record_id = 18374.

[3] Deng K, Geng Z, Liu J. Association Pattern Discovery via Theme Dictionary Models [J]. To appear in J Royal Statist Soc B,2013.

[4] He Y, Geng Z. Active learning of causal networks with intervention experiments and optimal designs [J]. J Machine Learning Research, 2008,9: 2523 – 2547.

[5] Jia J, Geng Z, Wang M. Identifiability and estimation of probabilities from multiple databases with incomplete data and sampling selection [J]. Lecture Notes in Computer Sciences, 2006, 4109: 792 – 798.

[6] 维克托·迈尔 – 舍恩伯格,肯尼思·库克耶. 大数据时代—生活、工作与思维的大变革 [M]. 盛杨燕,周涛 译,杭州: 浙江人民出版社.

[7] NIH Big Data to Knowledge (2013). http://bd2k.nih.gov/index.html#sthash. Yu5HxjcM. dpbs.

[8] Pearl J. Causality 2nd ed Cambridge University Press,2009.

[9] 纳特·西尔弗. 信号与噪声 [M]. 胡晓姣,张新,朱辰辰 译,北京: 中信出版社.

[10] Xie X, Geng Z. A recursive method for structural learning of directed acyclic graphs [J]. J Machine Learning Research,2009, 9: 459 – 483.

[11] Xie X, Geng Z, Zhao Q. Decomposition of structural learning about directed acyclic graphs [J]. Artificial Intelligence, 2006, 170: 422 – 439.

[12] Yin J, Zhou Y, Wang C, He P, Zheng C, Geng Z. Partial orientation and local structural learning of causal networks for prediction. Challenges in Causality Volume 1: Causation and prediction challenge. Ed. by I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J. Pellet, P. Spirtes and A. Statnikov, 2009: 93 – 105.

作者简介

耿直,男,1956 年生,江苏徐州人,1989 年毕业于日本九州大学,获理学博士学位,现为北京大学数学科学学院教授,中国现场统计研究会理事长,中国统计学会副会长。研究方向为数理统计学、因果推断。

(责任编辑: 程 晞)