DOI:10.3969/j.issn.1009-6868.2016.02.002
网络出版地址:http://www.cnki.net/kcms/detail/34.1228.TN.20160302.1129.002.html

试论大数据之“大”

A Commentary on the “Big” of Big Data

李廉/LI Lian
(合肥工业大学 计算机与信息学院,安徽 合肥 230009)
(School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

中图分类号:TP393 文献标志码:A 文章编号:1009-6868 (2016) 02-0007-004

摘要: 认为大数据提供了一种全新的认知世界的角度和方法。与熟知的数学和大部分物理学的基本认知规律不同,大数据分析原则上是一种基于观察和归纳的经验主义认知,这种方法曾一度被现代实证主义的研究模式边缘化。随着近年来大数据产生与分析的技术进步,这一古老方法正在重新焕发活力,并赋予大数据新的内容和形式。在这个意义上,给出了关于大数据4V的新解释。同时通过一个NP问题的例子,探讨了大数据对于复杂问题解决的新方法和新思路。

关键词: 大数据;观察归纳;概率近似正确;数据分布;数据清洗;数据价值;例证法

Abstract: Big data provides a brand-new angle and method of perceiving the world. Like mathematics and physics, big data analysis is, in principle, a methodology based on observation and empirical induction, which has been marginalized in recent times by positivism in research models. As techniques for big data creation and analysis have developed, this methodology has blossomed. We give a new explanation of the “four Vs” of big data: state the four Vs here. We also discuss an example of an NP problem to explore new methods for solving complex.

Key words: big data; observation and induction; probability approximately correct; data distribution; data cleaning; data value; exemplification method

1 大数据的应用目的

毫无疑问,对于大数据的分析与处理,目的是要获取知识,或者说认知结论。那么,通过大数据来获取知识,与大数据时代之前获取知识有什么不同吗?为此,我们需要回顾人类直接从自然界获取知识的两种手段:观察和实验。

早期人们获取知识的手段是观察,通过对于自然现象的仔细观察,得到关于自然规律的认知。由于观察本身没有干预自然的运行,因此可能会受到众多因素的干扰而影响认知的质量,甚至得到不正确的知识。16世纪之后,由伽利略等逐步开创了现代实证主义研究的手段,这种研究需要预设因果关系,然后在实验室里进行现象重建。由于在实验条件下,干扰因素被抑制到最小,因此可以准确重现现象之间的因果。实验与观察的区别是:实验需要预先假定一种或者多种因果现象,然后在实验室设计适当的实验来重现这些现象,从而证实因果关系。实验并不特别依赖研究人员的直观经验,而且具有很强的说服力。观察是需要在众多的现象之间,找出其中的因果关系。这里面并没有什么统一的方法和标准,因此通过观察得到结论需要直观和经验,同时说服力往往也不够。在实证主义的研究体系建立之后,观察研究就让位于实验,除了少数的学科(例如宇宙学),在绝大多数自然学科中,实验成为形成结论的标准手段,任何结论必须在实验室里面被验证,仅仅在自然界被观察到是不够的。究其原因,还是因为历史上由于观察手段的不足,难以获得大量数据,而建立在小数据基础上的观察,往往是不准确的,得到的结论也缺乏说服力。例如通过观察,人们最容易得到的结论是地球中心论,这种学说统治了科学界1 500多年。只是到了开普勒、哥白尼时代,随着观察数据的增加,才能够颠覆以前的结论,重新建立新的学说。这说明:观察研究这种人类最基本的研究手段,其结论的可靠性依赖于是否有足够的观察数据,当数据多到一定程度时,所获取的结论才具有可靠性。因此一个重要的问题出现了:对于一个具体的观察对象,数据量达到多大时,我们才能采信所获取的结论呢?

既然过去是受限于数据的不足,使得人们研究自然问题主要依赖于实证主义的实验方法。那么现在随着信息技术的发展,获取数据的能力有了极大提高,进入了大数据时代。我们是否可以重新回到先辈那里,采用观察的方法来研究问题,获取知

收稿时间:2016-01-17
网络出版时间:2016-03-02
基金项目:自然科学基金(61370219);广东省佛山市创新团队项目(2015IT100095)

识？这个不是可能不可能的问题，而是已经在我们身边发生的事实。在人文科学、社会科学、自然科学等领域已经开始采用大数据来进行研究，产生新的知识，这些新知识极大地丰富了我们对于自然和社会的认知，有许多成果是依赖试验方法无法想象的，其中最典型的例子可能是图像识别和语音分析，在基本无法通过实验来重构现象的人文社科领域更是如此。通过观察设备（传感器）作用于各种自然现象、社会活动和人类行为，产生了大量的数据，分析和处理这些数据就是对这些观察结果的归纳和提炼；因此通过大数据来认知各种自然的、社会的和人文的规律，是传统意义上对于观察研究的新提升和新表现。人们研究科学的手段又重新回到了观察这个最原始和最基本的手段，但是这一次的回归是螺旋式上升，比起张衡和托勒密时代的观察完全不在一个层面上。从古代依靠人的感官来观察现象，到现在依靠传感器来观察现象，数据的密度、广度、准确性和一致性已经不能同日而语了，因此观察这种研究手段在信息时代换发了新的生命力，成为新时代的科学研究方法。

2 大数据的定量化

大数据是与观察研究密不可分的，大数据分析和处理的目标是获取知识，得到结论。那么怎样从大数据得到的结论呢？在小数据时代，这需要经验和直观。在大数据时代，需要应用计算机来进行分析和处理。一般来说，大数据分析是一种归纳的方法，因此必然具备归纳方法的普遍特点，即通过大数据获取的结论具有某种不确定性，这就是数据分析理论中常说的概率近似正确（PAC）[1]。确切地说，一个结论概率近似正确，是指该结论能够以$1-\delta$的概率获取，并且具有误差ε（类似于机器学习里说的泛化误差）。也就是说：我们通过大数据来获取知识，不能保证每次都能够正确获取，而且获取的知识也不能保证绝对正确。δ和ε这两个数，反映了使用大数据获取知识的能力和精度。这是所有归纳分析的共同特点，也是观察研究的固有性质。这一点既可以说是优点，又可以说是缺陷。优点是这样可以保证我们至少获得一个接近真理的结论；缺点是我们不能期待获取绝对正确的结论。如文献[2]中所说：“当我们掌握了大量新型数据时，精确性就不那么重要了，我们同样可以掌握事情的发展趋势。大数据不仅让我们不再期待精确性，也让我们无法实现精确性。然而，除了一开始会与我们的直觉相矛盾之外，接受数据的不精确和不完美，我们反而能够更好地进行预测，也能够更好地理解这个世界。”

但是问题到此远没有结束，反而是刚刚开始。和古代的科学家不同，在大数据时代，我们需要回答这样一个问题：给定任意的δ和ε，为了在大于$1-\delta$的概率下得到一个误差小于ε的结论，我们需要多少数据？如果能够回答这个问题，哪怕是在某种程度上回答了这一问题，我们就超越了古代科学家凭经验和直观做出结论的限制，真正把获取结论的过程建立在客观和科学的基础上，这样得到的结论自然也就有了很强的说服力。

为了更加仔细考察从大数据获取的知识的过程，从中得到方法论的一些结果，我们需要明确一些概念。

第1个概念是样本和分布。从观察现象得到的数据并从中来获取知识，首先需要解决的问题是得到的数据不可能是所有的数据，我们能够得到的数据永远是客观上整体数据的一部分。显而易见，只有明确知道样例数据与整体数据之间满足的分布假设，从样例来获取知识才具有可靠性和准确性。其中最受关注的就是样例集合与整体数据之间具有何种分布状态，同分布自然是理想状态，但是也已经发展了一些方法来讨论非同分布的情况[3-4]。

第2个概念是数据的清洗。观察是现象的记录，并且从记录的数据来获取结论。数据都是具有属性的，如果属性与期望的结论之间没有可关联的关系，那么数据只是一堆随机的噪声而已。在小数据时代，我们主要靠直觉和经验来筛选属性和处理数据，使得从处理后的数据能够有效地得到结论。在现代大数据分析和处理过程中，发展了一些自动或者半自动的方法来进行处理。

第3个概念是获取结论的成本。从计算机科学的角度，是指获取结论所花费的时间复杂度和数据空间复杂度，主要是时间复杂度。

综上所述，在大数据背景下获取结论，与数学和大部分物理学的结论形式不同，采用了概率近似正确的概念，并由此建立结论的获取方法和标准。实际上，由于观察得到的数据总是局部的和不完整的，所以通过观察得到的结论原则上都是PAC形式。

现在我们可以讨论一个有意义的问题：预设一个目标结论以后，需要多少数据量才能以PAC的方式得到该结论。这个问题无疑是大数据研究中最重要的内容之一。在小数据时代，对于这个问题并没有特别关注，因为通过数据来获取结论是借助直观和经验的，数据量的多少对于能够得到结论没有直接的联系，一个聪明人只要少数的几个例子就可以“猜”到结论，而对于一般的人来说，再多的例子也无法从中得到结论。但是在大数据时代，由于是通过设计算法，借助计算机进行数据分析，因此数据量的多少自然会对于结论的产生和结论的正确性具有直接的关系。由于大数据的研究才仅仅起步，对于这个问题目前上没有一般的结果。但是在附加一些不太苛刻的条件之后，却有一个出乎意料的结果，这就是Blumer等在1989年得到的一个定理。

定理1（Blumer定理）[5]：设D是实例的集合，S是样例的集合，H是目标

函数，A是算法，如果：

（1）S与D具有相同的分布；

（2）H是一个二分类函数；

（3）H在算法A的假设空间中。

那么，对于任意给定的δ和ε，当数据量N满足

$$N \geqslant \frac{1}{\varepsilon}\left(4\log_2\frac{2}{\delta}+8VC(\aleph)\log_2\frac{13}{\varepsilon}\right) \quad (1)$$

可以在期望$1-\delta$内，得到函数G，并且G与H的误差不超过ε，即以PAC的模式得到函数G。其中VC($\aleph$)是算法A的假设函数空间$\aleph$的VC维数。

我们经常说大数据有4个V，即体量（Volume）、高速（Velocity）、多态（Variety）和价值（Value）。这些V反映了大数据的特点，但是究竟达到什么程度才叫做大数据，需要有一个量化的讨论，否则大数据就仅仅是一个笼统的概念。

结合前面的讨论和定理，我们尝试给出一种大数据的量化的解释。首先要指出的是：数据量大不大是依据所要得到的结论性质而言。对于一个工厂的产品检验来说，可能几百个抽样（观察）数据就足够了，但是对于暗物质的探测，可能几个P的数据量也未必够用[7]。这说明谈论数据量之大小，脱离了目标是无意义的。

定理1指出：在给定目标（包括预设的结论形式和精度，即δ和ε）的前提下，当数据量达到一定程度后，就可以按照PAC模式得到结论。因此我们可以把Blumer定理中的N的倒数$1/N$定义为数据的价值密度，这就给出了4个V中Value的量化定义。在数据平等的前提下，每一个数据相对于期望结论与相应算法，它的价值就是$1/N$。同样的数据对于不同的期望结论和算法，其价值是不同的。同时根据该定理，可以定义N为解决问题所需要的最小数据体量，即Volume。当数据量达到N时，就可以称为关于期望结论和相应算法的大数据。由于这个数量的巨大，因此如何存储和处理海量数据是重要的技术问题。对于另外两个V：Velocity是指需要有快速存储技术和计算技术来接纳和处理高速涌入的数据，但是也可以看作是最小数据体量与问题解决时间要求的比值，这个值决定了数据处理的最低速度；Variety是指数据的来源和类型很多，对于问题解决而言，这种多态性取决于数据清洗的质量。

一般来说，数据的多态性越丰富，越是会有利于数据的整理和表现，也越会容易得到结论，对机器学习的语言来说，越容易保证目标函数在假设集合中。当然，数据的多态性会增加数据获取和整理的难度，因此需要在数据处理的成本和效率之间加以折中[8-10]。

3 1个NP复杂类的例子

上面已经讨论了如何通过大数据来获取结论，以及获取结论的精确性和可靠性问题。在这一节，我们继续通过1个例子来说明这个问题。

一个NP问题是指一台非确定图灵机在多项式时间可以解决的问题。NP问题能否具有确定的多项式算法是一个长期以来未能解决的重要问题。现在我们通过大数据的思维方式来探讨此类问题，寻求新的解决问题思路。

定理2：对于任意的NP语言类L，以及给定的n、δ和ε，则存在一个算法A，当随机抽取的样例个数超过了

$$N=\frac{1}{\varepsilon}\left(4\log_2\frac{2}{\delta}+f^2(n)\log_2\frac{13}{\varepsilon}\right)$$

时，可以期望$1-\delta$获取一个确定的函数，该函数对每一个长度等于n的x，计算$x \in L$？误差不超过ε。并且N多项式（实际上是平方）依赖于n，$1/\delta$和$1/\varepsilon$。

这个定理只是一个理论上的结果，因为即使当n=100，δ=0.05，ε=0.01时，需要的样例个数也达到了8 000万这样的数量级。对于这么多的样例，需要进行标注，即一个个注明它们是否属于L，本身就是一项十分费力的事情。但是该定理却表现了通过大数据分析获取结论一些规律。首先该结果表明了通过一些例子的分析，就可以得到一般性的结论（具有一定的误差）。对于非确定语言L而言，不需要去构造相应的图灵机，只需要计算一定数量的样例，同样可以某种概率得到一个判断函数H，在误差ε的范围内判断是否$x \in L$？大数据给我们带来的一个重要方法论正是在这个意义上的，通过对大量的观察数据的分析和处理，可以得到原来只有实验验证和逻辑推理才能得到的结论。这种模式在古代就存在，但是后来被更先进的实证主义的研究方法所取代，而大数据的出现重新召回了它的灵魂。

通过例子来证明问题，这个方法在80年代就被洪加威等研究过[11]，称为例证法。在小数据时代，例证法需要经过仔细挑选的特殊例子，在大数据时代，可以通过大量的数据来取代这个苛刻的条件，因此大数据的出现将例证法推到了几乎可以在所有领域应用的地步。这对于过去只靠实验和逻辑证明问题而言自然是开创了一个新时代。

4 结束语

大数据提供了认识世界的新方法和新角度。有别于我们习惯的实验验证和逻辑推理方法，大数据定义了通过观察和样例获取结论的模式，这种模式古已有之，而且是人类研究自然的最古老的方法。大数据的出现使得这一方法重新焕发活力，并且赋予了新的内容和形式。由于大数据本质上是通过观察来获取结论，因此和所有采用观察方法研究问题（无论是否采用大数据分析）具有相通之处，所获取的结论具有某种不确定。在当前讨论的大数据分析方法中，这种不确定性主要表现在两个方面：一个是获取结论的可能性，一个是结论本身的可靠性。同时，获取结论的不确定性可以在某些条件下任意逼近

确定性。正如舍恩伯格所说：这种不确定性不是表示大数据分析不如物理学和数学，而是说明大数据提供了一种新的认知世界的模式。

大数据分析并不排斥传统的物理学和数学的研究模式，相反，大数据分析建立的关联关系可以为因果关系和逻辑关系的研究提供佐证和启示。

参考文献

[1] MITCHELL T. Machine Learning [M]. 曾华军, 译. 北京: 机械工业出版社, 2008
[2] SCHONBERNER V. Big Data：A Revolution that Will Transform How We Live, Work and Think [M]. 周涛, 译. 杭州: 浙江人民出版社, 2013
[3] FAKOOR R , LADHAK F , NAZI A , et al. Using Deep Learning to Enhance Cancer Diagnosis and Classification[C]// Proceedings of the 30 th International Conference on Machine Learning. USA: ICML, 2013：211–218
[4] WANG A, AN N, YANG J, et al. Alterovitz, Incremental Wrapper Based Gene Selection with Markov Blanket[C]//ASE BioMedCom Conference. USA. USA: ASE, 2014：106–108
[5] BLUMER A, EHRENFEUCHT A, HAUSSLER D, et al. Learnability and the Vapnik–Cherbonenkis Dimension [J]. Journal of the ACM, 1989: 36(4): 929–965
[6] 罗军舟. AMS大数据处理的挑战[R]. 合肥: 中国计算机大会, 2015
[7] 周志华, 李武军, 张利军. CCF2014–2015中国计算机科学技术发展报告[M].北京: 机械工业出版社, 2015
[8] TOPOL E. The Creative Destruction of Medicine [M]. 张南, 等译. 北京: 电子工业出版社, 2014
[9] CHO K. A Brief Summary of the Panel Discussion at DL Workshop of ICML[EB/OL]. [2015–07–13]. http://deeplearning.net/2015/07/13/a–brief–summary–of–the–panel–discussion–at–dl–workshop–icml–2015
[10] 洪加威. 能用例证法来证明几何定理吗?[J]. 中国科学A辑, 1986(3): 234–242
[11] LASZLO BARABASI A. Bursts: The Hidden Pattern Behind Everything We Do [M]. 马慧, 译. 北京: 人民出版社, 2012

作者简介

李廉，合肥工业大学计算机与信息学院教授，中国计算机学会理论计算机科学专业委员会主任，教育部高等学校大学计算机课程教学指导委员会主任；主要研究方向为机器学习、计算机网络、无线传感器网络等；获国家教学成果二等奖1项，安徽省教学成果特等奖1项。

综合信息

2015年中国发明专利申请量首次突破百万件

国家知识产权局发布的数据显示：2015年，国家知识产权局共受理发明专利申请110.2万件，同比增长18.7%，连续5年位居世界首位。

随着中国科技水平发展，专利创新越来越重要。2015年中国专利数量有显著增长，专利申请量超过200万件，发明专利申请受理量首次超过100万件。

其中，中国发明专利授权26.3万件，比2014年增长了10万件，同比增长61.9%。而同为专利大国的美国，2015年专利申请数量却少有地迎来了授权专利数量下降。美国商业专利数据库IFICLAIMS发布的2015年专利统计数据显示：全年授权专利数约为29.8万件。虽然同比下降不到1%，但这是自2007年以来授权专利数量首次下跌。

美国专利在十年间维持了小幅的上涨，申请量年均30万～40万件。中国专利申请量一直呈增长态势，在2005年以前，数量上已经超越美国。兰德公司发布的中国的专利与创新报告指出：中国十年来专利爆发性增长，主要是专利激励政策和市场力量推动的结果。在政策指导下，国家鼓励个人和企业创业。

激烈的市场竞争也促进企业进行专利技术研发。以手机行业为例，近十年间，中国手机品牌崛起，加入以往被国外手机厂商所垄断的市场，手机品牌数量大幅度增长。对于专利技术的投入能够帮助这些品牌在市场中占据一席之地。

（转载自《中国信息产业网》）

通信行业支出将再次增长 设备厂商应密切关注需求变化

自Ovum的最新研究报告称：传统的电信行业正在萎缩。2015年全球电信运营商的收入预计为1.78万亿美元，相较2014年下降5%，2012—2014年这3年间全球CSP收入一直处于持平状态。

对于设备厂商来说，一些全新的客户正在从OTT风潮中出现。包括所有这些供应商在内，行业资本支出预计将会逐渐上升，从2014年的4 050亿美元增至2020年的4 660亿美元。

Ovum表示：全球电信运营商的资本支出仍旧受到严格限制，2014年这一数字为3 390亿美元，CSP资本支出在2015—2017年将出现下滑，之后到2020年将再次出现增长，而这主要是受到早期5G支出的推动。

从总体来看：ICP的资本支出正在稳步增长，到2020年将达到1 100亿美元。这一数字与所有固定电信运营商2020年的网络资本支出差不多。许多ICP自身拥有强大的内部技术研发队伍，并会通过ODM厂商来制造定制化的设备。

虽然ICP代表着一种机会，但企业和政府客户也正在建设日益复杂的网络，他们也需要帮助。设备厂商需要调整他们的解决方案，并与互补企业进行合作，同时大力投入销售和营销，才能赢得市场。简单地将那些针对电信运营商的技术转售给这些市场是无法赢得客户认可的。

（转载自《中国信息产业网》）