大数据预测研究及相关问题

■ 吕本富* 陈 健* 中国科学院大学管理学院 100190

摘 要：大数据的应用核心就是大数据预测。大数据预测完全依赖大数据来源，因此具有“全样非抽样、效率非精确、相关非因果”的特征。按照预测的精细程度，大数据预测的层级分为三级，依次递进：圈子范围的大小、圈子内有哪些族群、族群中每个成员的个性特征。能否在三个层次获得准确的预测结果，关键在于前台数据和后台数据、宏观数据和微观数据、共性数据和个性之间的关联分析。对于任何大数据的预测结果，必须设计一个验证流程，以防止出现不必要的错误。

关键词：大数据预测 大数据特征 预测层次 数据关联 预测验证

DOI:10.11842/chips.2014.01.010

2012 年 3 月 19 日，奥巴马政府宣布美国投资 2 亿美元启动“大数据研究与开发计划（Big Data Research and Development Initiative）”，旨在提高从大型复杂数字数据中抽取知识与观点的能力，以帮助解决国家在科学与工程中最紧迫的诸多挑战问题，增强国家安全，实现教育与学习方式的转变，大数据因此也被誉为“未来的新石油”。挖掘“石油”的方法有哪些？预测是大数据应用的核心，也是挖掘的价值所在。寻找基于大数据的预测方法，在数据采集、数据存储、数据分析以及数据可视化能够准确应用，必须清晰大数据的预测特征和预测范围。

大数据的更大用途在于根据建立的模型预测未来某一事件的发生，并可据此进行人为干预，使其向着理想的方向发展。决策行为将日益基于数据分析做出，而不是像过去更多凭借经验和直觉。更多地基于事实与数据做出决策，这样的思维方式，将推动一些习惯于靠“差不多”运行的社会发生巨大变革。

一、大数据的定义

有关大数据 (big data) 的定义，维基百科认为大数据或称巨量资料、海量资料、大资料，无法透过人工在合理时间内达到撷取、管理、处理、并整理成为人类所能解读的资讯。学术界和产业界的理解差异很大，并不统一。本文的论述以著名产业机构的定义为基础。

知名咨询机构伽特纳（Gartner）指出，大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。从数据的类别上看，大数据指的是无法使用传统流程或工具处理或分析的信息。它定义了那些超出正常处理范围和大小、迫使用户采用非传统处理方法的数据集。

麦肯锡在其报告《大数据：下一轮创新、竞争、增效的利器》（Big data: The next frontier for innovation, competition, and productivity）中指出：大数据指的是大小超出常规的数据库工具获取、存储、

* 吕本富，中国科学院大学管理学院教授，博士生导师，主要研究方向：网络经济、管理智慧。陈健，中国科学院大学管理学院 博士生。

管理和分析能力的数据集。但它同时强调，并不是说一定要超过特定 TB 值的数据集才能算是大数据。

从预测的角度看，大数据和网络空间密切相关，网络上每一笔搜索，网站上每一笔交易、每一笔输入都是数据，透过计算机做筛选、整理、分析，所得出的结果不仅仅得到处理现实业务简单、客观的结论，更能用于帮助企业经营决策，收集起来的资料还可以被规划，引导开发更大的消费力量。

二、大数据预测案例——谷歌流感趋势

2009 年全球首次出现甲型 H1N1 流感，在短短几周之内迅速传播开来，引起了全球的恐慌，公共卫生机构面临巨大压力，如何预防这种疾病的传染。预防的核心是预测病情的蔓延程度，现实的情况是人们可能患病多日、实在忍不住才会去医院，即使医生在发现新型流感病例时，同时告知美国疾病控制与预防中心（CDC），然后 CDC 汇总统计，整体上大约需要两周时间。对于一种飞速传播的疾病而言，信息滞后两周将会带来非常严重的后果，能否提前或者同时对疫情进行预测呢？

碰巧的是，在甲型 H1N1 流感爆发的几周前，谷歌的工程师们在《自然》杂志上发表了论文，通过谷歌累计的海量搜索数据，可以预测冬季流感的传播。在互联网普及率比较高的地区，当人们遇到问题时，网络搜索已经成为习惯。谷歌保留了多年来所有的搜索记录，而且每天都会收到来自全球超过 30 亿条的搜索指令，谷歌的数据分析师通过人们在网上的搜索记录就可以来完成各种预测。就流感这个具体问题，谷歌用几十亿条检索记录，处理了 4.5 亿个不同的数字模型，构造出一个流感预测指数。结果证明，这个预测指数与官方数据的相关性高达 97%。和 CDC 流感播报一样，可以判断流感的趋势和流感发生的地区，但是比 CDC 的播报可以提前两周，有力地协助卫生当局控制流感疫情。

总之，2009 年甲型 H1N1 流感爆发的时候，与滞后的官方数据相比，谷歌的流感趋势是一个更有效、更及时的指示标。公共卫生机构的官员获得了非常及时、有价值的数据信息。谷歌并不懂医学，也不知道流感传播的原理，但是以事物相关性为基础，以大数据为样本，其预测精准性与传统方式不相上下，而且其超前性是传统方式所无法比拟的。

三、大数据预测的特征

分析大数据预测依赖数据来源，因此数据源的特征也决定了大数据的预测特征。

1. 实样而非抽样

在小数据时代，由于缺乏获取全体样本的手段，人们发明了“随机调研数据”的方法。理论上，抽取样本越随机，就越能代表整体样本。但问题是获取一个随机样本代价极高，而且很费时。人口调查就是典型一例，即使一个大国都做不到每年都发布一次人口调查，因为随机调研实在是太耗时耗力。但有了云计算和数据库以后，获取足够大的样本数据乃至全体数据，就变得非常容易。谷歌可以提供谷歌流感趋势的原因就在于它几乎覆盖 7 成以上的北美搜索市场，已经完全没有必要去抽样调查这些数据，只需要对大数据记录仓库进行挖掘和分析。

但是这些大数据样本也有缺陷，实际样本不等于全体样本，依然存在系统性偏差的可能。所以存在一个数据规模的阈值问题。数据少于这个阈值，问题解决不了 达到这个阈值，就可以解决以前束手无策的大问题 ;而数据规模超过这个阈值，对解决问题也没有更多的帮助。我们把这类问题称为“预言性数据分析问题”，即在做大数据处理之前，可以预言，当数据量到达多大规模时，该问题的解可以达到何种满意程度。如何确定阈值？当前的学术界还没有一个完整的解决方案。

2. 效率而非精确

过去使用抽样的方法，就需要在具体运算上非常精确，因为所谓“差之毫厘便失之千里”。设想一下，在一个总样本为 1 亿人口中随机抽取 1000 人，如果在 1000 人上的运算出现错误的话，那么放大到 1 亿中偏差将会很大。但全样本时，有多少偏差就是多少偏差而不会被放大。谷歌的人工智能专家诺维格写道 :大数据基础上的简单算法比小数据基础上的复杂算法更加有效。数据分析的目的并非就是数据分析，而是有多种决策用途，故而时效性也非常重要。

精确的计算是以时间消耗为代价的，在小数据时代，追求精确是为了避免放大的偏差不得已而为之。在

大数据时代，快速获得一个大概的轮廓和发展脉络，就要比严格的精确性要重要得多。但是，在需要依赖大数据进行个性化决策时，张冠李戴是个很大忌讳，精确性就变得非常重要。所以在效率和精确之间存在一个平衡点，这是大数据预测中一个棘手问题。

3. 相关而非因果

舍恩伯格说：大数据时代只需要知道是什么，而无需知道为什么。大数据研究不同于传统的逻辑推理研究，需要对数量巨大的数据做统计性的搜索、比较、聚类、分类等分析归纳，因此继承了统计科学的一些特点。统计学关注数据的相关性或称关联性。所谓“相关性”是指两个或两个以上变量的取值之间存在某种规律性。“相关分析”的目的就是找出数据集里隐藏的相互关系网（关联网），一般用支持度、可信度、兴趣度等参数反映相关性。难道大家都喜欢购买 A 和 B，就一定等于你买了 A 之后的果就是买 B 吗？未必，但的确需要承认，相关性很高——或者说，概率很大。知道喜欢 A 的人很可能喜欢 B 但却不知道其中的原因。

亚马逊的推荐算法非常有名，它能够根据消费记录来告诉用户可能会喜欢什么，这些消费记录可能是别人的，也可能是该用户历史的记录。但它不能说出喜欢的原因。如果把这种推荐算法用于亚马逊的物流和仓储布局，仅仅了解相关性远远不够，必须“知其然，还知其所以然”。否则将带来额外的损失。这也是相关性预测和因果性预测的分界线。

四、大数据预测的层次和关联分析

大数据预测有很多行业应用，以企业方面的预测最为成熟。很多领先的企业已经采用一系列的管理流程、技术手段去挖掘这些数据所带来的价值，从大量客户的交互数据、网站访问的行为中去辨识客户访问数据的模式，并从中获取更为精确的客户洞察力，以制定精确的行动纲领，去为服务的对象设置更好的产品和服务，从而能够获得更高的业务收益。

大数据的预测可以形象分为三个层级：第一层级：谁在圈子内？第二层级：属于圈子内哪个族群？第三层级：族群中个体的画像，尤其是支付能力。

第一层级：针对一个具体的应用，依据性别、收入、地域、年龄等特点，圈定相近的人群。比如，在电子商务网站内，预测“什么地方的人买东西最疯狂”或是预测“什么型号手机最好卖”；麦当劳、肯德基以及苹果公司等旗舰专卖店的位置的精准选址；针对这个全体如何进一步打磨广告、市场营销以及产品。总之，数据科学家可以确定某个定价范围或产品是否会挤占来自其它定价或是产品的销量，由此，可以优化你的定价策略和产品线。

第二层级：在确定的圈子之内，把商品或人群分成不同的族群，然后分析族群和商品之间的关系。比如，通过对消费者的购物行为，特别是购物篮的分析，沃尔玛发现了“啤酒与尿布”之间关联的经典商业案例；Target 分辨出哪一类顾客更有可能是怀孕的女性；销售漏斗里的哪类顾客最有可能在哪个水平上被转化。总之，营销数据和网站日志数据以及交易数据都关联起来，以确定市场推广活动背后的 ROI（投资回报率）。

第三层级：确定圈子内人群的个性特征，由此提供个性化的定价、产品和服务。在电影《点球成金》中，大家可能对奥克兰队寻找棒球手的方式印象深刻——他们运用复杂的统计工具来深度分析比赛进程中运动员的各项指标，从而计算出哪位球员的市值被低估。航空公司和通信公司则利用大数据向客户提供体贴入微的服务、沃尔玛利用数据分析结果来调整库存和个性化价格、联邦快递可以优化每天不同的递送路线等也属于第三层级的范畴。

在大数据集合中，能否在这三个层级做出准确预测，严重依赖下面三类数据的辨析和关联分析。

1. 前台（行为）数据和后台（结果）数据

前台数据指访问量、浏览量、点击流及站内搜索等反应用户行为的数据，而后台数据更侧重商业数据，比如交易量、转化率（ROI）、终身价值 (LTV，Life time Value)。传统的预测分析中，既有行为数据的分析，也有商业（结果）数据的分析，但把行为数据和商业数据联系起来，进行综合预测比较少。

传统企业的数据大多数来自于交易型数据，而交易数据只是结果，处于用户消费决策漏斗的最底部。而交易前的各种浏览、搜索、比较等用户行为数据都处于漏斗的顶部，所以其数量远远超过交易数据。“用户行

为信息”（User Behavior Information）非常丰富，包括用户在网站上发生的所有行为，如搜索、浏览、打分、点评、加入购物筐、取出购物筐、加入期待列表（Wish List）、购买、使用减价券和退货等；包括在第三方网站上的相关行为，如比价、看相关评测、参与讨论、社交媒体上的交流、与好友互动等。和门店通常能收集到的购买、退货、折扣、返券等和最终交易相关的信息相比，电子商务的突出特点就是可以收集到大量客户在购买前的行为信息，而不是像门店收集到的是交易信息。

2. 宏观（整体）数据和微观（细节）数据

信息的价值是什么？信息给它的拥有者带来了什么？更直观的感受、更精细的判断、更准确的预测。大数据能挖掘出不同细节的信息。

哈耶克在《知识在社会中的利用》一文中已经把知识分成两类：一类是科学知识，即被组织起来的知识由专家所掌握，在理论和书籍中可以得到；一类是特定时间和地点的知识，为处于当时和当地的人所拥有。哈耶克所讲的知识就是信息（数据），也可以把它们这样分作两类：经济理论作为信息是经济现象中规律性的总结，是一种普遍的趋势或状态，多用于对宏观经济和经济整体状况的描述；特定时间和地点的知识是由每个人所掌握的可以利用的独一无二的信息，是信息不对称的根源，基于这种信息做出的才是个性化的，才是有价值的。但是在各类的经济管理中，这两类信息经常被搞混，有时拿经济理论作特定时间和地点的知识用，有时拿特定时间和地点的知识作经济理论用，错误地配置了社会的资源以至影响了经济运行的效率。

这两类信息的数据描述，就是数据颗粒度的大小。粗粒度信息描述的经济整体状况，而细粒度信息描述了微观细节，其微观干预能力更强，应用价值更高。如何区分粗粒度信息和细粒度信息？各省市的身高排名是粗粒度信息，每个人实名的身高就是是细粒度信息；张三的九型人格类型是粗粒度信息，张三每次在某些特定情境中的行为记录是细粒度信息；一家媒体 / 一个品牌 / 一个品牌官微的影响力是粗粒度信息，这家机构每次发出的信息到达了谁、这些人产生了什么反应是细粒度信息。

3. 共性（历时）数据和个性（场景）数据

在大数据预测中，个性数据和共性数据存在着巨大的差别，特别是在关联预测和推荐算法中。

有些关联预测只需要考察历史数据，通过历史数据来预测用户未来的需求；有些预测和场景有关，通过历史数据预测远远不够，还需要情景数据、情绪数据和社交数据。而基于这些预测产生的关联推荐算法，就存在不同维度的排序，和业务场景相关的推荐具有相当的技巧性（Tricky）。

对于书、手机、家电这些东西，亚马逊称之为硬系（Hard Line）产品，或者成为“标品”（但也不一定），通过历史数据（时间序列）预测是比较准的，甚至可以预测到相关的产品属性的需求。但是，对于服装这样的产品，亚马逊称之为软系（Soft Line）产品，亚马逊探讨十多年都没有特别好的办法。因为这类商品受到的干扰因素太多，比如：用户的对颜色款式的喜好，穿上去合不合身，爱人朋友喜不喜欢…… 甚至买的人多了反而会卖不好，这类的东西太容易变，所以根本没法预测好。需要情景和职业人士（Stock/Vender Manager）的经验相结合，才能“预测某品牌的某种颜色的衣服或鞋子”。

在电子商务中，用户买一个东西没有退货，有很大的概率相信用户喜欢这个商品。对于音乐和视频，用户听了一首歌或是看了一段视频，不能武断地觉得用户是喜欢这首歌和这个视频的。因此推荐音乐视频与推荐商品的场景完全不一样。推荐算法在不同的业务场景下的实现难度也完全不一样。

推荐算法也有对应的两种：一种是共性化推荐，结果就是推荐流行、普适的东西，也许会是用户已知的东西。比如，外地人到了北京，如果想找个饭馆，推荐烤鸭总是一个不错的选择；如果找个地方游玩，推荐天安门、故宫、天坛，受欢迎的程度比较高。需要注意的是：在网络空间中这些共性化的东西不是被水军刷的；另一种是个性化推荐，需要分析用户的个体喜好和情景改变。比如，就餐饮来说，需要考虑对方口味随年龄和环境的改变，甚至需要逆用户口味，帮客户发掘新鲜点。对于喜欢吃辣的人，总是被推荐川菜和湘菜，时间长了会觉得烦，频繁地推荐烤鸭，也降低了推荐水准。

五、大数据预测的流程和验证

由于大数据具有 4V 特点 数据体量巨大 (Volume)、数据种类繁多 (Variety) 流动速度快 (Velocity)、价值密度低 (Value)，需要制定合理的数据分析流程，能够提升预测的效率。在应用大数据预测的结论之前，需要验证流程，以免出现不必要的错误。

1. 数据分析流程

大数据的“ 4V ”特征表明其不仅仅是数据海量，对于大数据的分析将更加复杂、更追求速度、更注重实效。数据量呈指数增长的同时，隐藏在海量数据的有用信息却没有相应比例增长，反而使我们获取有用信息的难度加大。以视频为例，连续的监控过程，可能有用的数据仅有一两秒。数据科学家必须借助预测分析软件来评估他们的分析模型和规则，预测分析软件通过整合统计分析和机器学习算法发挥作用。

统计与分析主要利用分布式数据库，或者分布式计算集群来对存储于其内的海量数据进行普通的分析和分类汇总等，以满足大多数常见的分析需求，在这方面，一些实时性需求会用到 EMC 的 GreenPlum、Oracle 的 Exadata，以及基于 MySQL 的列式存储 Infobright 等，而一些批处理，或者基于半结构化数据的需求可以使用 Hadoop。统计与分析这部分的主要特点和挑战是分析涉及的数据量大，其对系统资源，特别是 I/O 会有极大的占用。

IBM SPSS 和 SAS 是两个数据科学家常用的分析软件。R 项目则是一个非常流行的开源工具。如果数据量大到“ 大数据 ”的程度，那么还需要一些专门的大数据处理平台如 Hadoop 或数据库分析机如 Oracle Exadata。

2. 结果验证流程

大数据预测结果可能带来和原来截然不同类型的决策，非常容易引起争议。解决方案就是设计验证流程，利用对照实验，测试各种假设和分析结果，以指导投资决策和运营变革。

亚马逊网站上的某段页面文字只是碰巧出现的吗？其实，亚马逊有严格的验证流程，整个网站的布局、字体大小、颜色、按钮以及其他所有的设计，其实都是在多次审慎测试后的最优结果。销售实物商品的企业也有决策的验证流程，麦当劳的部分门店安装了搜集运营数据的装置，用于跟踪客户互动、店内客流和预订模式。研究人员可以对菜单变化、餐厅设计以及培训等是如何对劳动生产力和销售额的影响进行建模。

对比性实验可以帮助管理者将因果关系与单纯的相关性区分开来，从而减少结果的可变性和改善财务表现和产品性能。完善的实验可以有多种形式。例如，主要在线企业是持续的测试者。在某些场合，它们将网页的固定部分用于开展实验，以找出提高用户参与或促进销售的动因。事实上，恰如其分的验证流程才能使大数据预测更上一层楼。

参考文献：

[1] 李国杰．大数据研究 :未来科技及经济社会发展的重大战略领域 :大数据的研究现状与科学思考 [J]．中国科学院院刊 ,2012,27(6):647-657.

[2] Weiss R, Zgorski L. Obama Administration Unveils 'Big Data' Initiative: Announces $200 Million In New R&D Investments[R/OL].[2012-03-29]. Washington: Office of Science and Technology Policy, Executive Office of the President, White House, http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_ 2.pdf.

[3] Tom Kalil. Big Data is a Big Deal, March 29, 2012. Available at http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal.

[4] 人大经济论坛，传统分析与大数据分析的对比 [DB/OL]. (2012) http://bbs.pinggu.org/forum.php?mod=viewthread&tid=2140833&page=1.

[5] 甘晓，李国杰．大数据成为信息科技新关注点 [J]．中国科学报，2012.

[6] 马帅，李建欣，胡春明．大数据科学与工程的挑战与思考 [J]．中国计算机学会通讯，2012,8(9):22-30.

[7] 涂兰敬．大数据与海量数据的区别 [J]．网络与信息，2011,25(12):37-38.

The Predication Based on Big Data and Related Issues

Lv Benfu, Chen Jian
School of Management, University of Chinese Academy of Sciences, 100190

Abstract: The core application of big data is forecasts based on big data. It is accuracy entirely dependent on big data sources, and therefore has the feature of "full but non-part, efficiency but non-precise, relevant but non-causal". According to forecasts fineness, big data predicted l accuracy is divided into three levels, followed by progressive: the size of the groups, the size of sub- groups, the personality characteristics of each member within the sub-groups. The key to get the accurate predictions is correlation analysis between the behavioral data and outcome data, the macro data and micro data, and the common and personality data. For any predictions of big data, you must design a verification process to prevent unnecessary errors.

Keywords: Big Data Forecast, Big Data Characteristics, Forecast Levels, Correlation Analysis, Forecast Verification

（责任编辑：何岸波，张志华；责任译审：龚　宇）