Chapter 5

第5章 大数据分析

学习目标

- 了解大数据分析的基本分类和步骤。
- 了解预测分析的作用。
- 了解大数据分析的应用。
- 掌握几种常用的大数据分析工具。

随着大数据时代的来临，大数据分析也应运而生。大数据分析是指对规模巨大的数据集进行分析。

5.1 大数据分析概述

传统的数据分析通过数据抽样，并不断改进抽样方法以提高样本的精确性，从而对整体数据进行推算，并竭力挖掘数据之间的因果关系；而大数据分析的对象是全体数据，不存在因采样的不合理而导致的预测结果的偏差。传统数据分析的算法比较复杂，通常是用多个变量的方程追求数据之间的精确关系；而大数据分析则使用简单的算法实现规律性的分析。传统的数据分析关注的是“为什么”的因果关系思维方式，而大数据分析关注的是“是什么”的相关性关系，即从海量数据中分析出人类不易感知的关联性。传统数据分析追求的是精确性，即探寻问题的最终答案，而大数据分析是基于海量数据进行分析而得出的结果，该结果一般都是一种供决策参考的指向性意见。

下面通过表5-1说明传统的数据分析和大数据分析的区别。

表5-1 传统数据分析和大数据分析的区别

对比项目	传统数据分析	大数据分析
分析对象	部分数据的采用	全部数据
分析类型	结构化数据	结构化、半/非结构化数据

续表

对比项目	传统数据分析	大数据分析
精确性	必须接收精确、规范化的数据	可以是非精确、非规范化、不完整的数据
分析算法	对算法的要求较高	算法简单有效
分析结果	注重因果关系	更注重相关性，而非因果关系

5.2 大数据分析基础

5.2.1 大数据分析基本分类

本节主要讲述数据挖掘分析领域中最常用的四种数据分析方法：描述型分析、诊断型分析、预测型分析和指令型分析，如图 5-1 所示。

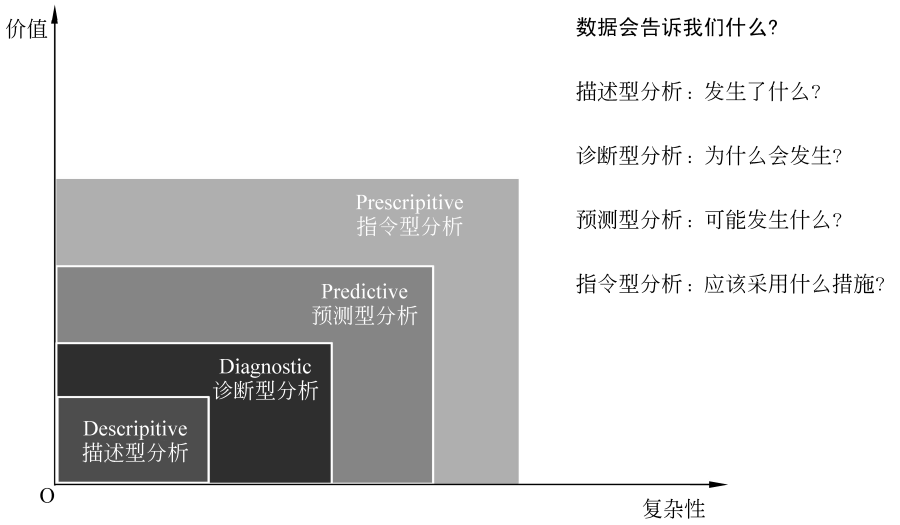

图 5-1 四种大数据分析方法

1. 描述型分析

描述型分析是最常见的分析方法。在业务中，这种方法向数据分析师提供了重要指标和业务的衡量方法，例如每月的营收和损失账单。数据分析师可以通过这些账单获取大量的用户数据。了解用户的地理信息，这就是描述型分析的方法之一。利用可视化工具能够有效地增强描述型分析所提供的信息。

2. 诊断型分析

描述型数据分析的下一步就是诊断型数据分析。通过评估描述型数据，诊断分析工具

能够让数据分析师更深入地分析数据，获取到数据的核心。

被良好设计的工具按照时间序列进行数据读入、特征过滤和获取数据等功能，以便更好地分析数据。

3. 预测型分析

预测型分析主要用于预测事件未来发生的可能性，预测一个可量化的值或者预估事情发生的时间点。

预测模型通常使用各种可变数据实现预测。数据成员的多样化与预测结果密切相关。

在充满不确定性的环境下，预测能够帮助人们做出更好的决定。预测模型也是很多领域正在使用的重要方法。

4. 指令型分析

预测型分析的下一步就是指令型分析。指令模型基于对“发生了什么”“为什么会发生”和“可能发生什么”的分析帮助用户决定应该采取什么措施。通常情况下，指令型分析不是单独使用的方法，而是在前面的所有方法都完成之后最后需要完成的分析方法。

例如，交通规划分析考量了每条路线的距离、每条路线的行驶速度以及目前的交通管制等因素，以帮助驾驶员选择最佳的回家路线。

最后需要说明一点，每一种分析方法都对业务分析具有很大的帮助，它们都应用在数据分析的各个方面。

5.2.2 大数据分析步骤

大数据分析步骤归结起来有以下 6 个基本方面。

1. Analytic Visualizations（可视化分析）

不论是数据分析专家还是普通用户，数据可视化都是数据分析工具最基本的要求。可视化可以直观地展示数据，让数据自己说话，让观众听到结果。

2. Data Mining Algorithms（数据挖掘算法）

可视化是给人看的，数据挖掘则是给机器看的。集群、分割、孤立点分析还有其他算法让人们可以深入数据内部挖掘价值。这些算法不仅要处理大数据的量，也要处理大数据的速度。

3. Predictive Analytic Capabilities（预测性分析能力）

数据挖掘可以让分析员更好地理解数据，而预测性分析则可以让分析员根据可视化分析和数据挖掘的结果做出一些具有预测性的判断。

4. Semantic Engines（语义引擎）

由于非结构化数据的多样性给数据分析带来了新的挑战，因此需要一系列工具解析、提取、分析数据。语义引擎需要能够从“文档”中智能地提取信息。

5. Data Quality and Data Management(数据质量和数据管理)

数据质量和数据管理是一些管理方面的最佳实践。通过标准化的流程和工具对数据进行处理可以保证得到一个预先定义好的高质量的分析结果。

6. Data Storage and Data Warehouse(数据存储和数据仓库)

数据仓库是为便于多维分析和多角度展示数据而按特定模式进行存储所建立起来的关系数据库。在商业智能系统的设计中,数据仓库的构建是关键,是商业智能系统的基础,其承担了对业务系统数据进行整合的任务,为商业智能系统提供了数据抽取、转换和加载(ETL)功能,并按主题对数据进行查询和访问,为联机数据分析和数据挖掘提供了数据平台。

5.2.3 异步分析

异步分析遵循捕获、存储、分析的流程,在这个过程中,数据由传感器、网页服务器、销售终端、移动设备等获取,之后再存储到相应的设备上,最后再进行分析。由于这些类型的分析都是通过传统的关系数据库管理系统(RDBMS)进行的,因此数据形式都需要转换或者转型为 RDBMS 能够使用的结构类型,例如行或者列的形式,并且需要和其他数据相连续。

处理的过程称为提取、转移、加载或者 ETL。首先将数据从源系统中提取并处理,然后将数据进行标准化处理,最后将数据发送至相应的数据仓储进行进一步分析。在传统数据库环境中,这种 ETL 步骤相对直接,因为分析的对象往往是人们熟知的金融报告、销售或者市场报表、企业资源规划等。然而在大数据环境下,ETL 可能会变得相对复杂,因此转型过程在不同类型的数据源之间的处理方式是不同的。

当分析开始的时候,数据首先从数据仓储中被抽取出来,然后被放进 RDBMS 中以产生需要的报告或者相应的商业智能应用。在大数据分析的环节中,裸数据以及经过转换的数据大多会被保存下来,因为其可能在后面还需要被再次转换。

5.3 大数据预测分析

5.3.1 什么是预测分析

预测分析是一种统计或数据挖掘解决方案,它可以在结构化和非结构化数据中使用,以确定未来结果的算法和技术,可用于预测、优化、预报和模拟等许多用途。大数据时代下,预测分析已在商业和社会中得到了广泛应用。随着越来越多的数据被记录和整理,未来的预测分析必定会成为所有领域的关键技术。

5.3.2 预测分析的作用

预测分析和假设情况分析可以帮助用户评审和权衡潜在决策的影响力,可以用来分析

历史模式和概率，以预测未来的业绩并采取预防措施，其主要作用如下。

1. 决策管理

决策管理是用来优化并自动化业务决策的一种卓有成效的成熟方法，它通过预测分析让组织能够在制定决策以前有所行动，以便预测哪些行动在将来最有可能获得成功，以优化成果并解决特定的业务问题。决策管理包括管理自动化决策设计和部署的各个方面，供组织管理其与用户、员工和供应商之间的交互。从本质上讲，决策管理使优化的决策成为了企业业务流程的一部分。由于闭环系统会不断将有价值的反馈纳入决策的制定过程中，所以对于希望对变化的环境做出即时反应并最大化每个决策的组织来说，决策管理是非常理想的方法。

当今世界中竞争的挑战之一是组织如何在决策制定过程中更好地利用数据。可以用于企业以及由企业生成的数据量非常大且正以惊人的速度增长。与此同时，基于此数据制定决策的时间却非常短，且有日益缩短的趋势。虽然业务经理可以利用大量报告和仪表板监控业务环境，但是使用此信息指导业务流程和用户互动的关键步骤通常是人工参与的，因此不能及时响应变化的环境，希望获得竞争优势的组织必须寻找更好的方式。

决策管理使用决策流程框架和分析优化并自动化决策，通常专注于大批量决策并使用基于规则和分析模型的应用程序实现决策。对于传统上使用历史数据和静态信息作为业务决策基础的组织来说，这是一个突破性的进展。

2. 滚动预测

预测是指定期更新对未来绩效的当前观点，以反映新的或变化中的信息的过程，是基于分析当前数据和历史数据以决定未来趋势的过程。为应对这一需求，许多企业正在逐步采用滚动预测方法。

大企业采用 7×24 小时的业务运营影响造就了一个持续而又瞬息万变的环境，风险、波动和不确定性持续不断，并且任何经济动荡都具有近乎实时的深远影响。

毫无疑问，对于这种变化，感受最深的是 CFO(财务总监)和财务部门。虽然业务战略、产品定位、运营时间和产品线改进的决策可能是在财务部门外部做出的，但制定这些决策的基础是财务团队使用绩效报告和预测提供的关键数据与分析。具有前瞻性的财务团队意识到传统的战略预测无法完成这一任务，因此他们会迅速采用更加动态的、滚动的和基于驱动因子的方法。在这种环境中，预测变为一个极其重要的管理过程。为了抓住正确的机遇，满足投资者的要求以及在风险出现时对其进行识别，关键的一点就是深入了解潜在的未来发展，管理不能再依赖于传统的管理工具。在应对过程中，越来越多的企业已经或者正准备从使用静态预测模型转型到使用滚动时间范围的预测模型。

采用滚动预测的公司往往有更高的预测精度、更快的循环时间、更好的业务参与度和更多明智的决策制定。滚动预测可以对业务绩效进行前瞻性预测；为未来计划周期提供基线；捕获变化带来的长期影响。与静态年度预测相比，滚动预测能够在察觉到业务决策制定的

时间点后定期更新，并减轻财务团队的行政负担。

3. 预测分析与自适应管理

稳定、持续变化的工业时代已经远去，现在的时代是一个不可预测、非持续变化的信息时代，未来还将变得更加无法预测，员工将需要具备更多的技能，创新的步伐将进一步加快，产品价格将会更低，顾客将具有更多的发言权。

为了应对这些变化，CFO 需要一个能让各级经理快速做出明智决策的系统，他们必须将年度计划周期替换为更加常规的业务审核，通过滚动预测提供支持，让经理能够看到趋势和模式，以在竞争对手行动之前取得突破，在产品与市场方面做出更明智的决策。具体来说，CFO 需要通过持续计划周期进行管理，让滚动预测成为主要的管理工具，每天和每周都报告关键指标。同时需要注意，使用滚动预测可以改进短期可见性，并将预测作为管理手段，而不是度量方法。

5.3.3 数据具有内在预测性

大部分数据的堆积并不是为了预测，但预测分析系统能够从这些庞大的数据中学会预测未来的能力，正如人们可以从自己的经历中汲取经验和教训一样。

数据最激动人心的不是其数量，而是其增长速度。人们会敬畏数据的庞大数量，因为有一点永远不会变，那就是今天的数据必然比昨天的数据多。规模是相对的，而不是绝对的。数据规模并不重要，重要的是膨胀速度。

世上万物均有关联，只不过有些事物之间是间接关系，这在数据中也有反映。

① 你的购买行为与你的消费历史、在线习惯、支付方式以及社会交往人群相关。数据能从这些因素中预测出你的行为。

② 你的身体健康状况与生活习惯和环境有关，因此数据能通过住宅小区以及家庭规模等信息预测你的健康状态。

③ 你对工作的满意程度与你的工资水平、表现评定以及升职情况相关，而数据则能反映这些事实。

④ 经济行为与人类情感相关，数据也能反映这种关系。

数据科学家通过预测分析系统不断地从数据堆中找到规律。如果将这些数据整合在一起，那么尽管你不知道自己将从这些数据里发现什么，但至少你能通过观测和解读数据语言发现某些内在的联系。数据效应就是这么简单。

预测通常是从小处入手。预测分析是从预测变量开始的，这是对个人单一值的评测。近期性就是一个常见的变量，表示某人最近一次购物、最近一次犯罪或最近一次发病等距离现在的时间，近期值越接近现在，观察对象再次采取行动的概率就越高。许多模型的应用都是从近期表现最积极的人群开始的，无论是试图建立联系、开展犯罪调查还是进行医疗诊断。

与此相似，频率可以描述某人做出相同行为的次数，它也是常见且富有成效的指标。如

果有人此前经常做某事，那么他再次做这件事的概率就会很高。实际上，预测就是根据人的过去行为预见其未来行为。因此，预测分析模型不仅要依靠那些枯燥的基本人口数据，例如住址、性别等，也要涵盖近期性、频率、购买行为、经济行为以及产品使用习惯之类的行为预测变量。这些行为通常是最有价值的，因为要预测的就是未来是否还会出现这些行为的概率，这就是通过行为预测行为的过程。正如哲学家萨特所言："人的自我由其行为决定。"

预测分析系统会综合考虑数十项甚至数百项预测变量，需要把个人的全部已知数据都输入系统，然后等待系统给出预测结果。

5.4 大数据分析应用

5.4.1 大数据分析的主要应用行业

大数据分析的发展应用不仅有助于加速智慧城市与智慧生活科技的实现，而且如果将其应用于制造与服务产业，则不但能有效控制营运成本，还可以洞察市场趋势，提前掌握用户的需求，还有机会透过跨产业的大数据分析结果，以发展智慧联网、智慧自动化、智慧生活、智慧城市等新兴科技服务业，进而重塑产业形貌，创造我国产业转型的崭新契机。

例如，面对全球人口结构的转变，预防医学、健康照护、个体化医疗需求的增加，如果医疗产业与可穿戴技术能结合彼此的专长，则可以运用大数据分析技术助力新药、医疗器材的开发，为健康管理与诊治方式带来改变。

我国在智慧终端装置，包括各式可穿戴式装置、智慧联网装置的制造优势，可以说是发展大数据分析的最有力后盾。

运用大数据分析也可以加速提升物流及资讯流的流通便利性，尤其是在极端气候变迁与复合性灾害日趋严重的大趋势下，城市治理需要面对的环境变化将变得更为多变，对大数据分析技术的需求自然也就应运而生。

大数据分析技术对制造产业的帮助更是显而易见，尤其是产品开发与组装成本可以因此大幅降低，营运成本也会因此降低。由于大数据分析技术也可应用于个人分析，因此对于运营范围涉及全球的服务业而言，其可以因此增加更多的产值。

综上所述，可以把大数据分析应用归纳为以下 9 方面，这些方面都是大数据在分析应用上的关键领域。

1. 理解客户、满足客户服务需求

大数据的应用目前在此领域是最广为人知的，其重点是如何应用大数据更好地了解客户及其爱好和行为。企业非常喜欢搜集社交方面的数据、浏览器的日志，以分析出文本和传感器的数据，从而更加全面地了解客户。一般情况下，企业会建立出数据模型进行预测。例如美国的著名零售商 Target 就是通过大数据分析得到有价值的信息，并精准地预测客户在什么时候想要生育小孩。另外，通过大数据的应用，电信服务商可以更好地预测出流失的客

户，超市可以更加精准地预测哪个产品会热销，汽车保险行业可以了解客户的需求和驾驶水平。

2. 业务流程优化

大数据可以更多地帮助业务流程进行优化，企业可以通过利用社交媒体、网络搜索以及天气预报挖掘出有价值的数据，其中大数据应用得最广泛的就是供应链以及配送路线的优化。在这两个方面，地理定位和无线电频率的识别可以帮助追踪货物和送货车，并利用实时交通路线数据制定更加合理的路线。人力资源部门也可以通过大数据分析进行改进，其中包括人才招聘的优化。

3. 大数据正在改善生活

大数据不仅应用于企业和政府部门，同样也适用个人。人们可以利用可穿戴装备（如智能手表或智能手环）生成最新的健康数据，可以根据热量的消耗以及睡眠模式进行健康状态追踪，还可以利用大数据分析扩展交友范围。大多数时候，交友网站就是利用大数据应用工具帮助人们匹配合适的社交伙伴的。

4. 提高医疗和研发技术

大数据分析应用的计算能力能够在几分钟内解码整个 DNA，并且制定出最新的治疗方案，同时可以更好地理解和预测疾病。就好像人们手上的智能手表可以产生健康数据一样，大数据同样可以帮助病人进行更好的治疗。大数据技术目前已经在医院用来监视早产婴儿和患病婴儿的情况，通过记录和分析婴儿的心跳数据，医生可以针对婴儿可能出现的不适症状做出预测，以帮助医生更好地开展治疗。

5. 提高体育成绩

现在，很多运动员在训练的时候都会应用大数据分析技术。例如 IBM Slam Tracker 工具可以使用视频分析技术追踪足球或棒球比赛中每个球员的表现，而运动器材中的传感器则可以获得比赛数据，从而分析和判断如何改进战术。很多精英运动队还会追踪运动员在赛场之外的活动，它们通过使用智能技术追踪运动员的营养状况、睡眠以及社交活动，从而监控其情感状况。

6. 优化机器和设备性能

大数据分析还可以让机器和设备在应用上更加智能化和自主化。例如，大数据工具曾经就被 Google 用来研发自动驾驶汽车。Toyota 的普瑞就配有摄像机、GPS 以及传感器，其在道路上能够安全地行驶，不需要人类的干预。大数据工具还可以应用于优化智能电话。

7. 改善安全和执法

大数据现在已经广泛应用到安全执法的过程中。例如美国安全局利用大数据打击恐怖主义，企业应用大数据技术防御网络攻击，警察应用大数据工具捕捉罪犯，银行信用卡公司应用大数据工具监测欺诈性交易。

8. 改善城市

大数据还被应用于改善城市。例如基于城市实时交通信息、利用社交网络和天气数据优化最新的交通情况。目前很多城市都在进行大数据的分析和试点。

9. 金融交易

大数据在金融行业主要应用于金融交易。高频交易是大数据应用得比较多的领域。其中,大数据算法应用于交易决定,现在很多股权的交易都是利用大数据算法进行的,这些算法现在越来越多地考虑了社交媒体和网站新闻的数据,从而决定在未来几秒内是买进还是卖出。

以上就是大数据分析应用的 9 大领域,随着大数据的应用越来越普及,还会产生很多新的大数据应用领域以及新的大数据应用。

5.4.2 大数据分析应用应注意的问题

开展数据分析工作的目的在于将数据分析的成果转化为业务发展的成效,使数据分析成果进一步优化企业的管理和决策,进而真正实现资源的优化配置,并促进各项业务的快速高效发展。对于企业业务部门而言,在数据分析成果的应用实施中,应该充分注意以下几点。

(1) 加强数据安全管理是数据分析成果应用的前提。

数据分析的成果和结论中普遍含有大量的、精华的、最具有参考意义的业务发展数据信息,这些信息同样是企业经营发展的机密信息,不仅对于企业的高级管理人员有重要的参考价值,对于企业业务部门的市场竞争对手同样有重要的战略参考价值,数据分析成果的泄露对企业造成的损失也是难以估量的,因此在使用数据分析成果的同时,一定要加强数据安全管理。

加强安全管理意识,同时制定严格的数据保密制度,并强化泄密责任追究制度;企业在应用和下发分析成果时必须按需下发、分级下发,不允许将完整的数据分析报告随意下发至任何人,必须在保障数据分析成果充分应用的同时尽可能减少数据分析成果的掌握人群,尽可能降低数据泄露的风险;一旦发现数据泄露的情况,要及时向上级管理部门汇报,以便及时采取措施,降低数据泄露对企业造成的损失。

(2) 树立“以量化分析指标为依据进行决策管理”的意识是数据分析成果应用的基础。

数据分析的意义不仅是让管理人员加强对业务知识的理解,还是为管理人员提供决策的核心数据依据,使企业的经营决策能够从“定性分析”和“经验指导”转变为“定量分析与定性分析相结合”和“数据指导与科学论证相结合”,避免出现“决策拍脑袋、事后拍大腿”的情况。如果企业各级管理人员不能从根本上树立“以量化分析指标为依据进行决策管理”的意识,那么数据分析成果就不可能得到真正深入的应用和实施,数据分析的应用价值也就无从发挥。因此,树立“以量化分析指标为依据进行决策管理”的意识是数据分析成果应用的

基础。

(3) 市场调研工作是数据分析成果应用的重要组成部分。

开展数据分析工作的基础在于数据,而数据的来源绝不仅仅是企业信息化系统提供的电子数据,一部分关键信息是信息化系统所不具有的,例如竞争对手的发展情况、兄弟企业的先进经验及发展趋势、市场份额占比的变化趋势等,只有包含这些信息,数据来源才是完整的,在此基础上进行数据分析而得出的结论才是最科学、最有参考价值的,因此在数据分析工作开展和成果应用的同时,业务部门必须积极配合技术部门的市场调研工作,完善数据采集机制,使数据分析成果更有针对性和实用性。

在数据分析成果应用的过程中,同样需要加强市场调研工作,原因有两点:一是需要市场调研验证数据分析成果和结论在本地市场的可行性;二是需要将数据分析成果与市场调研的结论相结合,在此基础上形成完整的营销策划书或者管理措施等,因为数据分析的结论和成果是不会直接转化为经营效益的,只有当营销策划方案和管理措施的实施和落实后,其才能转化为企业发展的真正成效。

(4) 完善的沟通协调机制是用好数据分析成果的关键。

数据分析成果在形成后不是一成不变的,在这些成果的使用过程中,需要各级管理部门及时反馈使用情况,并且随着时间的推移,在各项业务发展的内外部环境变化的情况下,还需要进一步调整和优化分析方案,以形成新的分析成果。也可以说,数据分析工作是一个螺旋式推进的过程,只有在各级业务部门和技术部门之间建立起完善、长效的沟通机制,才能促进数据分析工作的长效、科学发展,才能使数据分析成果更容易发挥实效。

(5) 完善的市场营销体系是用好数据分析成果的保障。

近年来,大企业的市场部一直在强调完善市场营销体系的建设,就数据分析成果的运用而言,要想真正发挥其作用,同样依赖于完善的市场营销体系,所有分析成果都需要依靠完整的营销调研、方案策划、落地实施、质量控制、成效反馈等完善的市场营销闭环体系以发挥其实用价值。因此,企业业务管理部门必须在应用数据分析成果的同时,将其融入市场营销体系建设和市场营销的日常管理工作中去。

(6) 提升管理人员的数据敏感度及需求挖掘能力是数据分析成果能够发挥长效作用的重要手段。

数据分析成果就是结论和数据,同样的成果在不同的管理人员眼中的价值点不同,能发挥的作用也不同,这就取决于管理人员的数据敏感度及其对分析成果向实用化方案的转化能力。因此,提升管理人员的数据敏感度并使其形成根据分析结论进行决策的习惯是数据分析成果能够长久发挥作用的重点环节之一。

数据分析工作目前已经形成长效机制,经营管理人员必须提升自身的需求挖掘能力,及时提出数据分析需求的着眼点,并及时与数据分析团队进行充分沟通,使有限的数据分析能力能够用在企业生产经营和发展最急需的地方,这也是数据分析成果能够发挥最大作用的关键环节之一。