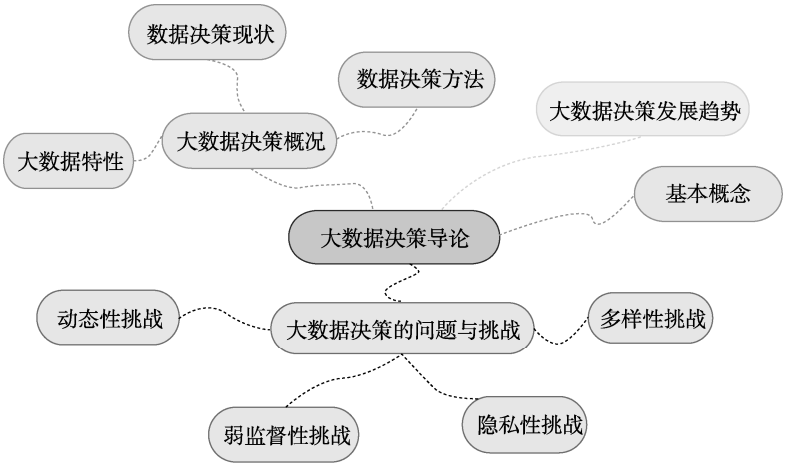
第 1 章

大数据决策导论

学习目标

1. 了解大数据决策的基本概念和方法。
2. 了解大数据决策的发展趋势。
3. 了解大数据决策所面临的问题与挑战。

知识图谱



应用场景导入

信息技术的迅猛发展，使得我们处于一个数据爆炸的时代。各行各业的决策者面临着前所未有的机遇与挑战：如何从海量、多源、高速变化的数据中提取有价值的信息，进而做出科学、及时、高效的决策？传统的基于经验和有限数据样本的决策方式已难以应对当今复杂多变的环境。

以现代智能制造为例，制造企业通过部署物联网传感设备，实时采集生产线上的各类数据，这些数据涵盖了设备运行状态、产品质量参数、环境条件等多个维度。借助大数据决策平台，企业能够对这些多源异构数据进行融合分析，不仅能够发现设备故障的先兆信号，预测产品质量波动趋势，还能优化生产调度和资源配置。与传统决策模式相比，基于大数据的决策具有更强的全局性和前瞻性，能够实现从“事后处理”到“预测

预防”的转变。

在商业领域，企业利用大数据技术深入挖掘消费者行为模式和市场趋势，从而制定更加精准的营销策略和产品开发计划。与传统的市场调研相比，大数据分析能够捕捉到更细微的消费者需求变化，发现传统方法难以察觉的市场机会，使企业能够更敏捷地响应市场变化，提供个性化的产品和服务。

这些应用场景展示了大数据决策如何改变传统决策模式，从基于经验和直觉的定性分析，转向基于全面数据和科学模型的定量分析；从静态的分析框架转向动态的决策体系；从满足一般需求转向满足个性化需求。大数据决策不仅提高了决策的准确性和效率，还扩展了决策的边界，使得过去因信息不足而无法解决的复杂问题变得可分析、可预测、可控制。

1.1 基 本 概 念

当今社会正处于一个信息技术高速发展的时期，数据信息的交互、共享与开放程度持续加快，使得各行业领域的数据信息呈爆炸式增长。“大数据时代”如期而至，并成为当今社会的代名词。大数据以其蕴藏的巨大经济、社会和科研价值受到社会各界的广泛关注。2012 年 1 月，达沃斯世界经济论坛发布的大数据报告“Big data, big impact: new possibilities for international development”将大数据列为和货币与黄金同等重要的新经济资产。2012 年 5 月，联合国发布的“Big Data for Development: Challenges & Opportunities”白皮书指出，大数据是联合国和各国政府的一个历史性机遇：利用大数据进行决策，是提升国家治理能力，实现治理能力现代化的必然要求，可以帮助政府更好地参与经济社会的运行与发展。在科研领域，大数据正引领数据密集型科学（Data-intensive Science）的到来，形成继实验科学、理论科学以及计算科学之后的第四科学范式，有望推动传统科学的假设驱动模式向基于大数据探索的数据密集型方法转变。在全球信息化快速发展的背景下，大数据已逐渐成为世界各国的基础性战略资源，运用大数据推动社会经济发展正成为趋势。

现阶段，加快发展智能经济、智能服务和智能制造是我国经济增长的内在需求和必然选择。目前，我国处于工业化和信息化的深度融合时期，我国制造业正处于从价值链的低端向中高端、从中国制造向中国创造转变的关键历史时期，发展基于大数据的人工智能新技术是实现从制造大国向制造强国迈进的战略举措。在此背景之下，国家相继出台了《“互联网+”行动计划》和《中国制造 2025》战略规划，特别是国务院颁布的《促进大数据发展行动纲要》和《新一代人工智能发展规划》都将大数据智能作为重点发展方向，大数据的战略资源地位进一步凸显。近年来，以大数据与人工智能技术为基础的“智能制造”成为推动大数据从概念到落地的重要模式和手段。从大数据的供给需求来看，智能制造的核心要义便是在两化融合的基础上构建智能分析优化系统“工业大脑”，对大数据进行智能化分析进而实现智能决策。

决策存在于人类一切实践活动当中。小到一台机器的操作，大到一个国家的治理，都离不开决策。例如，工业领域的操作优化与资源分配、商业领域的个性化推荐与供应

商选择、交通领域的车流控制与路径导航、医疗领域的疾病诊断与治疗策略等都属于决策范畴。随着社会节奏的持续加快，来自各领域行业的决策活动在频度、广度及复杂性上较以往都有着本质的提高。决策问题的不确定性程度随着决策环境的开放程度及决策资源的变化程度而越来越大。传统的基于人工经验、直觉及少量数据分析的决策方式已经远不能满足日益个性化、多样化、复杂化的决策需求。在当前信息开放与交互的经营环境下，机遇与挑战并存。想要把握机遇，就需要企业或组织具备出色的决策能力。在这个过程中，大数据正扮演着越来越重要的角色。

大数据作为一种重要的信息资产，可望为人们提供全面的、精准的、实时的商业洞察和决策指导。美国应用信息经济学家 Hubbard 认为“一切皆可量化”，并积极倡导数据化决策。纽约大学 Provost 教授等认为数据科学的终极目标就是改善决策。杨善林院士等则指出，大数据的价值在于其“决策有用性”，通过分析、挖掘来发现其中蕴藏的知识，可以为各种实际应用提供其他资源难以提供的决策支持。从数据到知识，从知识到决策，是当前大数据智能的计算范式。研究大数据的意义就是不断提高“从数据到决策的能力”。随着大数据技术的发展，人们传统的决策模式与思维方式正在发生变革，基于大数据的决策方式正逐渐成为决策应用与研究领域的主旋律，大数据决策时代已经到来。大数据能够突破事物之间隐性因素无法被量化的瓶颈，充分阐述生产的主客体和生产全过程、全时段的客观状态，通过智能化分析和预测判断来提高企业的决策能力。

国内外学者对于大数据决策的定义不尽相同。许多国外学者将“大数据决策”定义为利用海量、多元化数据资源，通过先进的分析技术和算法，实现数据到洞察、洞察到行动的转化过程。国内学术界普遍认为大数据智能决策是从数据到知识，再从知识到决策的计算范式。在行业应用中，大数据智能决策被定义为利用工业大脑、商业智能等系统，对多源数据进行智能化分析，从而实现精准的商业洞察、资源优化配置及预测性判断，帮助提升决策水平。例如，沃尔玛通过销售交易大数据的知识获取支持价格策略决策，百度工业大数据监测平台应用于制造行业决策支持。

基于大数据的科学决策，是公共管理、工业制造、医疗健康、金融服务等众多行业领域未来发展的方向和目标。本书将大数据智能决策定义为：以大数据为基础资源，结合人工智能技术，通过数据收集、处理、分析和知识挖掘，实现从数据到决策的智能化转化过程，旨在解决传统决策方式难以应对的高复杂性、高不确定性问题，为各领域提供全面、精准、实时的决策支持。而如何进行大数据的智能分析与科学决策，实现由数据优势向决策优势的转化，仍然是当前大数据应用研究中的关键问题。对大数据的分析和处理在不同行业和领域均存在着巨大的挑战，大数据的大体量、高通量、多源异构性和不确定性等对传统的数据处理硬件设备和软件处理方法均构成前所未有的挑战。本书将从大数据决策方法出发，覆盖数据采集、数据处理、智能决策等多方面的应用，为充分发挥大数据在决策中的价值提供良好的参考和借鉴。

1.2　大数据决策概况

大数据时代的到来，不仅对人们的生活产生了广泛而深刻的影响，也为企业和组织

提供了全新的决策支持资源。大数据的概念和特性一直是学界和业界关注的热点话题。不同领域的专家学者基于各自的背景和视角，对大数据的定义也给出了不同的阐述。总的来说，大数据是指无法在合理时间内利用现有数据处理手段进行存储、管理和分析的庞大数据集合。

相比传统决策，大数据决策具有诸多独特的特点。首先，大数据决策具有动态性，因为大数据是对事物状态的实时反映，决策分析需要跟随动态数据的变化而及时调整。其次，大数据决策具有全局性，不再局限于单一领域，而是注重多源信息的融合和全面系统的分析。再次，大数据决策存在不确定性，这既体现在大数据本身的不完备性和噪声，也表现在大数据分析处理方法的局限性。此外，大数据决策正在从传统的因果分析逐步向相关分析转变，以更好地适应大数据环境下信息不确定性的特点。最后，大数据决策正在从满足整体需求向满足个性化需求转变，这是大数据应用带来的显著趋势。

伴随着大数据技术的不断进步，支撑大数据决策的理论方法也在不断创新。一方面，智能决策支持系统正在从传统的静态模型逐步发展为自适应和分布式的智能系统，以更好地适应决策环境的动态变化。另一方面，针对大数据的不确定性特点，基于模糊集、粗糙集、贝叶斯理论等不确定性分析方法的智能决策方法正在得到广泛关注和应用。此外，大数据环境下的多源信息融合技术为决策分析提供了全新的思路，能够有效克服单一数据源的局限性，增强决策的可靠性。同时，面向大数据的增量式学习方法也成为当前研究的热点，尤其在应对数据结构、特征和类别的动态变化方面表现出了良好的适应性。

本节将围绕大数据决策的理论框架、核心技术方法及典型应用案例等方面，全面阐述大数据时代下智能决策支持的前沿研究进展，为读者提供一个系统而深入的学习和参考。

1.2.1 大数据的概念及特性

由于不同领域的大数据在特性上存在差异，并且人们分析大数据的背景和应用大数据的目的不同，因此不同领域的专家对大数据的定义也各不相同。高德纳咨询公司、维基百科、美国国家科学基金会分别从不同的角度给出了大数据的定义。我国的《工业大数据白皮书（2019 版）》还对工业大数据进行了定义。简而言之，大数据就是在合理时间内无法利用现有的数据处理手段进行诸如存储、管理、抓取等分析和处理的数据集合。

有关大数据的特性，业界普遍将其归纳为 4V 特性，如图 1-1 所示：一是数据体量（Volume）大，如一些电商企业日常处理 PB 级别的数据已经常态化；二是数据类型多样（Variety），如在工业大数据中数据类型包含了数值、文本、图片、音频、视频以及传

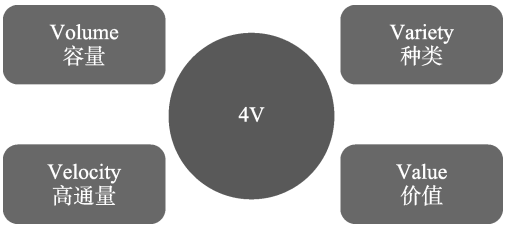


图 1-1　大数据 4V 特性

感器信号等；三是大数据的价值（Value）巨大，但价值密度稀疏，需要通过分析和挖掘来获取数据中有价值的信息；四是大数据的高通量（Velocity），它除了指数据高速产生以外，还意味着数据的采集与分析过程必须迅速及时，以满足用户“及时、实时”的决策需求。

1.2.2　数据决策理论与方法的发展现状

在信息技术快速发展的今天，决策理论和方法正经历着深刻的变革。从最初的简单数据分析工具，到如今融合人工智能、大数据、云计算等技术的智能化决策，相关理论在帮助人们应对复杂决策场景方面发挥着越来越重要的作用。本节将系统介绍数据决策理论与方法的发展现状，包括智能决策支持系统的演进历程、不确定性数据处理方法、多源信息融合技术以及增量分析方法等关键内容，为本书的相关内容提供全面的理论支撑。

1. 智能决策支持系统

决策支持是在管理科学和运筹学的基础上发展而来的一门学科。20 世纪 70 年代，决策支持系统（Decision Support System，DSS）的概念正式提出。DSS 是一种以提高决策有效性为目的，综合利用大量数据，有机地结合各种模型，通过人机交互的方式，辅助各级决策者实现科学决策的计算机系统。最初的 DSS 设计采用两库框架，由用户接口、数据库管理系统、模型库管理系统三部件集成而成。随着人们对 DSS 研究和应用的深入，DSS 逐步引入方法库管理系统、知识库管理系统和推理机，形成了包含数据库、模型库、方法库和知识库的四库框架。经过几十年的发展，DSS 不断与新技术、新学科相互交叉融合，并在体系结构、问题处理模式、功能模块集成等方面发生了巨大变化。

智能决策支持系统（Intelligent Decision Support System，IDSS）是在 DSS 基础上不断升级和发展而来的。20 世纪 80 年代，随着专家系统（Expert System，ES）的广泛应用，研究者将 DSS 与 ES 相结合，充分发挥 DSS 的数值分析能力和 ES 的符号知识处理能力，用于解决定量与定性问题以及半结构化、非结构化问题，有效扩大了 DSS 处理问题的范围。这种 DSS 与 ES 的结合构成了 IDSS 的初期模型。智能决策支持系统利用人工智能和专家系统技术在定性分析和不确定推理上的优势，以及人类在问题求解中的经验和知识，为决策问题的求解提供了更加广阔的思路。近年来，几乎所有有关 DSS 的研究都围绕着人工智能技术的应用而展开。人工智能方法已经逐渐渗透到 IDSS 的体系结构、问题求解方法等各个方面。智能决策系统的研究已经从过去的决策部件功能扩展发展到部件的综合集成，从过去的定量模型发展到基于知识的智能决策方法。

随着社会的发展，信息量激增，这使得管理和决策日趋复杂。单纯依靠某一个决策者做出的决策往往不够完善，因此群决策理论被引入 DSS，形成了群决策支持系统（Group Decision Support System，GDSS）。GDSS 为企业的组织决策提供一种开放与协同的决策环境，通过吸收群体的经验和智慧，实现群体对决策问题的共同求解，从而提高决策质量。目前，分布式环境下的 GDSS 和基于人工智能的群决策方法仍然是该领域的重要研究方向。

传统的 DSS 多采用静态模型，决策过程需要用户自主选择方法和模型，系统缺乏主

动决策机制。为解决这一问题，主动决策支持系统（Active DSS，ADSS）应运而生。ADSS 通过建立人类认知模型，在决策问题求解的不同阶段，给决策者提供不同的方法选择，从而形成不同的问题求解路径。虽然 ADSS 是基于人类先验知识的，但其前提假设是系统运行在静态的决策环境下，因此在实际应用中仍然存在适应性较差的局限性。

为了适应决策环境的变化，自适应决策支持系统（Adaptive Decision Support System，AdDSS）框架被提出。该系统尝试用机器学习和案例推理等方法从大量历史数据和过往经验中发现与决策问题相关的知识，使系统具有随时间和决策过程变化调整自身行为的能力。目前，关于 AdDSS 的研究涵盖了系统结构自适应、领域知识自适应、用户接口自适应等多个方面，自适应性和自学习能力已经成为智能决策支持系统的一个主要特征。

互联网技术的应用使得决策环境出现了新特点：决策分析中的数据不再集中于一个物理位置，而是分散在不同部门或地区。分布式决策支持系统（Distributed Decision Support System，DDSS）正是为适应这类决策问题而建立的信息系统。DDSS 将传统集中式 DSS 发展为网络环境下的分布式并行处理方式，通过网络连接工作平台和分布式数据库、模型库等，支持分布在各地的 DSS 彼此交互，从而为决策问题求解提供高效及时的决策支持。

随着智能体（Agent）技术在人工智能领域的深入研究，IDSS 进一步发展。智能体在决策支持系统中应具备独立能力、学习能力、协作能力、推理能力、智能性等特征。目前，多 Agent 智能决策支持系统已经成为发展趋势，通过加入人机交互 Agent、模型选择 Agent、模型求解 Agent 等，减少决策系统对专家的依赖，实现系统由“模型驱动”转为“问题驱动”，提高决策系统的整体智能性。云计算技术的兴起也为 IDSS 带来了新的发展方向。云计算通过互联网将虚拟化的数据中心和智能用户终端有机地联系起来，为用户提供了便捷的信息服务环境。在大数据环境下，云计算平台可以为大数据的决策分析提供庞大的存储空间和强大的分布式并行计算能力。决策环境的开放性、决策资源的虚拟化、问题求解的分布式协作性使得基于云计算的智能决策具有全新的特征。随着移动智能设备和移动互联网的普及，分布式移动云计算环境下的智能决策方法已成为当前研究热点。

在大数据环境下，企业或组织所面临的内外部环境更加复杂，业务问题呈现非线性、不确定性、多维化和实时性等特点。通过综合运用互联网、云平台和人工智能技术，将大数据的采集、存储、管理、分析、共享、可视化等一系列知识发现技术与现有的智能决策支持技术深度融合，构建基于大数据的 IDSS 已成为该领域的重要发展方向。未来基于大数据的 DSS 有望具备海量数据汇聚融合能力、快速感知和认知能力、强大的分析与推理能力、自适应与自优化能力，实现复杂业务的自动识别、判断，并做出前沿性和实时性的决策支持。

2. 基于不确定系统的智能决策方法

不确定性是指客观事物联系与发展过程中无序的、随机的、偶然的、模糊的、粗糙的、近似的属性。现实世界的多样性、随机性、运动性，以及人类对事物描述和信息表达的不精确性、模糊性决定了人们所能获取的数据本身存在着较多的不确定性。在大数据环境下，数据的多源、多样、增量及不完备等特点，加上人们对数据分析处理需求的

多样性（如数据融合等），使得大数据从宏观上具有相较于传统数据更多的不确定性。大数据的不确定性不仅存在于大数据本身，还体现在大数据的处理过程中。因此，关于大数据不确定性信息的表示与处理成为大数据智能决策理论方法研究中不可缺少的一部分。在不确定性理论方法中，模糊集、粗糙集、贝叶斯理论、灰色系统等在智能决策方法中都起到了关键作用。下面将从大数据不确定性处理的角度对相关方法进行介绍。

（1）粗糙集理论。粗糙集由波兰数学家 Pawlak 于 1982 年提出。粗糙集使用具有精确概念的上近似集和下近似集对一个不精确概念或知识进行近似表示与度量，其独特之处在于不需要主观先验知识，可以直接对数据进行分析与推理，并揭示潜在规律。目前，粗糙集及其扩展理论已经成为处理不精确、不一致、不完备信息的有力工具，并广泛用于数据挖掘、知识获取，以及各类决策问题的求解。为满足粗糙集方法的大数据决策分析需求，研究者从多个方向开展了探索，如基于 MapReduce 的粗糙集并行化层次属性约简方法和并行化优势粗糙集方法近似计算方法。针对大数据常见的不完备特性，已有研究将中性集和粗糙集方法相结合，来处理智慧城市大数据的不完备性问题，通过结合遗传算法研究面向决策粗糙集的大规模数据集的并行化属性约简方法，并成功应用于网络入侵检测等领域。

（2）三支决策方法。三支决策是在决策粗糙集基础上发展出来的理论，是一种综合考虑多种因素并权衡各种可能性以做出最优决策的方法。在这一方法中，决策者需要明确保守型决策、激进型决策和中庸型决策三种主要的决策路径。保守型决策倾向于选择风险最小、回报相对稳定的方案，以确保最小化潜在的损失；激进型决策则更注重高风险高回报的机会，往往在不确定性较高的环境中寻求最大化利益；中庸型决策介于两者之间，既考虑风险也权衡收益，力求在风险和回报之间找到最优平衡点。通过分析和评估这三种不同的决策路径，决策者能够更全面地理解各种可能的结果和影响，从而做出更为明智的决策。这种方法不仅适用于商业决策，还广泛应用于政策制定、项目管理及个人生活中的重要选择。

（3）贝叶斯理论。基于贝叶斯理论的方法已经在人工智能领域中的不确定性推理、机器学习等方面取得了丰富成果。对于不同规模大小的贝叶斯网络，可以分别采用精确推理和近似推理算法予以分析，并提供决策支持。在小规模贝叶斯网络中，精确推理算法能够高效地计算出节点之间的联合概率分布和条件概率分布，从而得出最优决策。而在大规模贝叶斯网络中，由于计算复杂度的显著增加，近似推理算法，如马尔可夫链蒙特卡洛方法和变分推理方法，能够在较短时间内提供较为准确的推理结果。结合这些推理方法，可以更好地处理复杂系统中的不确定性，提升决策的科学性和可靠性。在大数据环境下，贝叶斯方法因其严密的理论基础和灵活的应用特点，在智能决策支持系统中发挥着越来越重要的作用。

（4）模糊集理论。模糊集理论由扎德（Zadeh）于 1965 年首次提出，它通过隶属度函数来描述事物的不确定性和模糊性，为处理复杂系统中的不精确信息提供了有效的数学工具。在模糊集理论中，元素对集合的归属关系不再是传统的非此即彼的二值关系，而是用[0,1]区间内的数值来表示元素属于某个集合的程度。这种表达方式更符合人类的思维习惯和现实世界的复杂性。在大数据环境下，模糊集理论的应用得到了进一步扩展

和深化。研究者提出了多种基于模糊集的大数据分析方法，如模糊聚类分析、模糊综合评判和模糊模式识别等。这些方法能够有效处理大数据中的不精确性和语义模糊性，在数据分类、决策支持和知识发现等方面发挥重要作用。特别地，将模糊集理论与深度学习等人工智能技术相结合，可以更好地处理大规模数据中的不确定性问题，提高决策的准确性和可靠性。同时，模糊集理论在多准则决策分析中的应用也日益广泛，为解决复杂的决策问题提供了新的思路和方法。

（5）灰色系统理论。灰色系统理论由邓聚龙教授于 1982 年创立，是一种专门研究不确定性系统的理论方法。灰色系统理论认为，任何系统都包含已知信息（白色部分）和未知信息（黑色部分），而介于两者之间的不完全确知信息则称为灰色信息。这种理论特别适用于样本量少、信息不完全的“小数据”情况，且在大数据时代仍具有重要价值。在处理大数据决策问题时，灰色系统理论主要通过灰色关联分析、灰色预测和灰色聚类等方法发挥作用。灰色关联分析方法可以揭示系统中各因素之间的相互关系，为多因素决策提供依据；灰色预测方法则能够基于有限的历史数据预测系统的发展趋势，特别适用于具有时序特征的数据分析；灰色聚类方法则可以对具有不完全信息特征的数据进行分类和归类。已有研究将灰色系统理论与其他智能计算方法（如粒子群算法、遗传算法等）结合，这样能够更好地处理大数据环境下的不确定性决策问题，提高决策的科学性和可靠性。近年来，灰色系统理论在环境监测、经济预测、风险评估等领域的应用不断深化，具有广泛的运用价值。

3. 基于信息融合的智能决策方法

多源信息融合是人类固有的一种基本功能。人类本能地将各种感知器官获得的信息与先验知识综合起来，对周围环境和正在发生的事件做出准确的判断。古代印度“盲人摸象”的故事生动地说明了这一点：故事中的盲人各自触摸大象的不同部位，得出了截然不同的认知。这个寓言告诉人们：单凭一种感官获得的信息难以全面认知客观事物，而通过融合不同维度的信息，才能形成对事物的完整认识。多源信息融合技术正是模拟了人脑综合处理多源信息的功能，自动或半自动地将不同来源和时间点的信息转化为统一表示形式，为人们提供有效的决策支持。

大数据环境的出现使得多源信息融合变得更加重要。数据的分布式存储与交互式共享已成为常态，而自治数据源的分布式和分散控制是大数据应用的主要特征之一。从决策应用的角度看，企业或组织在决策时需要收集和整合大量数据，汇集不同观点，才能制定出符合客观规律的决策。随着数据获取的便利性增加，信息的全面性和多源信息的协同作用将更加受到关注。例如，在城市规划决策中，政府部门需要结合路网结构、交通流量、城市人口分布及关注点（point of interest，POI）数据进行综合分析；在医疗诊断中，专家往往需要融合多家医疗机构的诊断结果；在工业生产过程中，可以利用红外图像、超声波音频及其他监控数据来共同判断设备的运行状态。

信息融合作为一种概念框架，其发展历程颇具启发意义。这一技术最早以多传感器数据融合的形式出现在军事领域。20 世纪 70 年代，美国国防部提出了 JDL 模型，旨在对不同源的数据信息进行多层次融合处理，以提高目标识别、身份评估、战况评估和威胁评估的准确性。此后，信息融合技术不断发展，逐渐形成了一个涵盖信号处理、信息

理论、统计学、人工智能和机器学习的多学科研究领域。从信源关系的角度来看，信息融合可以分为三种基本类型：互补型、竞争型和合作型。互补型融合中，各信源互不依赖，分别感知目标或场景的不同方面，通过融合获取目标的全局信息；竞争型融合中，各信源描述相同目标或场景的同一方面，融合过程主要用于冗余校准和增强信任；合作型融合中，各信源相互依赖，从不同角度感知目标，通过融合可以获得全新的信息。从抽象层次来看，信息融合可以划分为数据层、特征层和决策层三个层次。数据层融合直接处理原始数据，保留了最完整的信息，但需要较大的计算和通信开销。决策层融合则是在规则和知识层面进行整合，具有更强的灵活性和更低的通信负载，但可能存在信息损失。特征层融合介于两者之间，通过融合从数据中提取的特征属性来实现信息整合，在效率和信息保留程度上取得了较好的平衡。

多源信息融合对大数据决策的重要性体现在两个方面：首先，它能够帮助深入挖掘数据价值，从众多分散、异构的数据源中获取隐含的价值信息，丰富决策的内涵；其次，通过多源数据的交叉验证，它可以有效降低大数据中潜在的噪声、数据缺失、信息不一致和语义模糊等不确定性因素，从而提高决策的可靠性。随着大数据技术的不断发展，多源信息融合将在智能决策支持系统中发挥越来越重要的作用。

4. 基于增量分析的智能决策

增量性是大数据的固有特性之一。现实生活中广泛分布的传感与监控设备、实时互联的社会媒体等都构成了大数据动态增长的在线场景。基于大数据决策的数据分析，不仅要从历史大数据中获取知识，更要对新增数据进行动态知识发现。传统机器学习方法对历史大数据的挖掘与分析往往建立在数据隐含规律对未来预测有效性的假设之上，或假定决策状态始终处于决策模型的闭环之内。显然，现实世界的复杂多变性决定了从历史数据中获取的知识多数只具备历史有效性，在实用性较强的决策应用领域，特别是决策时效性要求较高的工业控制领域和智能交通领域等，实时动态的增量式知识获取是保证决策质量的必要条件。增量性主要体现在三个方面：一是数据样本的增量；二是样本特征描述信息的增量；三是类别的增量与数据分布的变化。在数据样本增量方面，现有研究表明通过依赖采样方法可以显著提高增量式支持向量机算法的学习效率。在样本特征描述信息的增量方面，有研究者提出了特征增量随机森林学习方法，解决了因传感器增加而形成的数据特征增量问题。另有多粒度增量式属性约简方法，有效避免了数据增加过程中对等价类的重复计算。在类别的增量与数据分布变化方面，开集学习问题旨在寻求对已知类识别的同时，能有效识别未知新类。研究者提出了基于无标签数据增广类学习框架、最近类平均森林算法和支持向量机森林算法等方法。在流式数据处理中，基于概念漂移的增量式学习方法被认为是有效途径之一，如将概念漂移方法用于流式数据的非监督学习中，有效提高了在线异常检测的精度。

通过对以上几种理论与方法的分析，我们可以看到，大数据环境下的智能决策理论体系正在朝着多元化、自适应和实时化的方向发展。从最初的决策支持系统到如今基于云计算和人工智能的智能决策平台，从不确定性数据处理到多源信息融合，从静态数据分析到动态增量学习，这些理论与方法的发展都反映了人们对复杂决策场景的不断探索与应对。在这些理论方法的基础上，我们有必要进一步明确大数据决策的本质内涵，以

便更好地指导实践。决策是人们为实现某一特定的目标，在占有一定的信息和经验（知识）的基础上，根据主客观条件的可能性，提出各种可行方案，采用一定的科学方法和手段，对解决问题的方案进行比较、分析和评价，并最终进行方案选择的全过程。从本质上来讲，决策通常是目标驱动的行为，是目标导向下的问题求解过程，该过程也被广泛地认为是人类的认知过程。大数据决策便是以大数据为主要驱动的决策方式。随着大数据技术的发展，大数据逐渐成为人们获取对事物和问题更深层次认知的决策资源，特别是人工智能技术与大数据的深度融合，为复杂决策的建模和分析提供了强有力的工具。

1.2.3 大数据决策的特征

大数据决策具有两个维度的特征。从固有特性来看，它体现为动态性（数据实时更新）、全局性（多源数据整合）和不确定性。从发展趋势看，相关分析将逐渐取代因果分析，同时决策将更加注重满足用户的个性化需求。

1. 大数据决策的动态性

大数据是对事物客观表象和发展规律的抽象表达，其动态性和增量性是对事物状态的持续反映。不可否认的是，人们在决策过程中的每一步行动都将影响事物的发展进程，并全程由大数据所反映。此时决策问题的描述以及决策求解的策略都需要跟随动态数据给予及时调整，通过面向大数据的增量式学习方法实现知识的动态发展与有效积累，进而反馈到决策执行当中。大数据决策的动态性决定了问题的求解过程应该是一个集描述、预测、引导于一体的迭代过程，该过程须形成一个完整的、闭环的、动态的体系结构。简要来说，大数据环境下的决策模型将是一种具备实时反馈的闭环模型，决策模式将更多地由相对静态的模式或多步骤模式转变为对决策问题动态描述的渐进式求解模式。

2. 大数据决策的全局特性

截至目前，人们已经开发出多种多样的决策支持系统，但多数面向具体领域中的单一生产环节或特定目标下的局部决策问题，往往无法较好地实现全局决策优化与多目标任务协同。在信息开放与交互的大数据时代，大数据的跨视角、跨媒介、跨行业等多源特性创造了信息的交叉、互补与综合运用的条件，这促使了人们进一步提升问题求解的关联意识和全局意识。在大数据环境下，决策分析会更加注重数据的全方位性，生产流程的系统性、业务各环节的交互性、多目标问题的协同性。通过多源异构信息的融合分析，大数据决策可以实现不同信源信息对全局决策问题求解的有效协同。基于大数据的决策系统，对每个单一问题的决策，都将以优先考虑整体决策的优化为前提，进而为决策者提供企业级、全局性的决策支持。

3. 大数据决策的不确定性

一般而言，决策的不确定性来源于三个方面：一是决策信息不完整、不确定而导致的决策不确定性；二是决策信息分析能力不足而导致的决策不确定性；三是决策问题过于复杂而难以建模导致的不确定性。大数据决策的不确定性也不外乎包括以上三个方面。在信息不完整和不确定方面，首先，大数据具有来源和分布广泛、关联关系复杂等特性，对于多数企业而言，即便借助各种先进的数据收集手段尽可能地将各种信源数据进行整

合，但仍难以保证信息的全面性和完整性；其次，大数据固有的动态性决定了大数据的分布存在随时间变化的不确定性；另外，大数据中普遍存在的噪声与数据缺失现象决定了大数据的不完备性和不精确性。在大数据分析能力方面，显然现有的大数据分析处理技术还存在着不足，诸如多源异构数据融合分析、不确定性知识发现及大数据关联分析等方面仍是当前颇具挑战的研究方向。在决策问题建模方面，在一些非稳态、强耦合的系统环境下，建立精确的动态决策模型往往异常困难，比如流程工业中的操作优化决策。现阶段面向大数据的决策问题求解，人们通常使用满意近似解代替精确解，以此保证问题求解的经济性和高效性。这种近似求解方式实际上也反映了大数据决策的不确定性特征。

4. 从因果分析向相关分析转变

在过往的数据分析中，人们往往假设数据的精确性，并通过反复试验的手段探索事物之间的因果关系。但在大数据环境下，数据的精确性难以保证，数据总体对价值获取的完备性异常重要，此时用于发现因果关系的反复尝试方法变得异常困难。从统计学角度看，变量之间的关系大体可以分为两种类型：函数关系和相关关系。一般情况下，数据很难严格地满足函数关系，而相关关系的要求较为宽松，在大数据环境下更加容易被接受，并能满足人类的众多决策需求。在面向大数据智能化分析的决策应用中，相关性分析技术可为正确数据的选择提供必要的判定与依据，同时将其与其他智能分析方法相结合，可有效避免对数据独立同分布的假设，提高数据分析的合理性和认可度。

5. 决策向满足个性化需求转变

在商业和制造业领域，对用户进行精准营销、满足用户的个性化需求是提升客户价值和实现企业竞争力的经营准则。在大数据背景下，产品和服务的提供，以及价值的创造有望更加贴近社会大众的个性化需求。以互联网大数据为基础，企业通过舆情分析、情感挖掘等以用户为中心的数据驱动方法，可以精准挖掘消费者的兴趣与偏好，做出有针对性的个性化需求预测，进而为消费者提供专属的个性化产品与服务。宏观上讲，大数据可以打通企业和消费者之间的信息主动反馈机制。社会大众意见的表达，可以迅速转化为商业经营的决策依据，反向指导产品的设计和制造环节，实现生产与市场需求的有效对接。随着社会化媒体应用的深入，多元主体参与决策有了更高的便捷性和可能性，决策过程中价值多元的作用更加明显，由此传统自上而下的精英决策模型将会改变，并逐渐形成面向公众与满足用户个性化需求的决策模式。

通过以上对大数据决策特点的总结，可以发现大数据决策与传统基于小数据的分析决策有诸多不同之处。首先，大数据决策的特点反映了当前大数据智能决策的研究重点与需求。其次，大数据决策的动态性、全局性、不确定性以及向相关性分析的转变，决定了面向大数据的关联分析、不确定性分析、对增量与多源数据的有效利用等都将是大数据智能决策研究中的关键内容。

1.3　大数据智能决策发展趋势和挑战

大数据可以为人们带来更加科学全面的决策支持，但大数据智能决策的应用研究还

处于初期阶段，并仍面临诸多挑战。当前，尽管大数据技术在数据收集、存储、处理和分析方面取得了显著进展，但其在实际应用中的智能决策能力仍不成熟，存在许多亟待解决的问题。首先，数据质量和数据来源的多样性问题使得数据的可靠性和准确性难以保证。其次，数据隐私和安全问题成为大数据智能决策过程中的重大障碍，如何在保证数据隐私的前提下进行有效的智能决策仍需深入研究。此外，算法的复杂性和计算资源的限制也是当前大数据智能决策面临的重要挑战，现有的算法在处理大规模数据时常常效率不高，难以满足实时决策的需求。本节将讨论大数据智能决策面临的一些问题挑战，并指出潜在的应对方法或未来的发展趋势，包括提升数据质量管理，完善数据安全和隐私保护措施，优化算法效率，以及加强跨学科合作等，以推动大数据智能决策技术的发展和应用。

1.3.1 大数据多样性带来的挑战

大数据的多样性是其复杂性的重要来源之一，也是智能决策面临的主要难题。当综合决策需要整合来自不同来源的数据时，这些数据在类型、分布、频率和密度上可能存在显著差异，这对多源数据的融合分析和信息协同决策构成了巨大的挑战。虽然现阶段已有一些研究成果，但它们大多集中在特定场景和特定类型的大数据上。如何解决多源异构大数据的协同分析问题，消除信息孤岛，实现更具通用性和鲁棒性的大数据智能决策，仍然是一个关键性的研究课题。

多源大数据之间的关系通常是互补型或合作型，单纯通过数据层面的融合决策并不一定有效。目前在特征层实现异构数据融合的方法中，基于深度神经网络（DNN）的成果较为突出。然而，这些方法仅在一定程度上克服了数据类型的多样性问题，对分布、频率等多样性问题仍然难以应对。值得注意的是，任何决策都伴随着风险，数据分析过程的可解释性对决策者至关重要，而这恰恰是 DNN 的短板。基于粒计算的 DNN 可解释性研究有望成为大数据智能分析的一个潜在研究方向。

通过语义层或决策层实现多源数据的综合利用，是解决数据异质性较为有效的方法，能够避免多种异质性问题。在大数据环境下，分布式自治数据源是一个显著特点，去中心化将成为未来的趋势。通过分布式知识获取与协同的方法，可以有效实现多源异构数据的协同感知与交互。所谓协同，指的是对不一致信息的冲突分析。基于粗糙集、模糊集和群体智能决策的冲突分析方法，如何应用到大数据决策中，是未来一个重要的研究方向。

1.3.2 大数据动态性带来的挑战

人、机、物之间的交互活动日益加快，导致数据快速增长成为大数据的一个显著特性。对于决策需求的及时性和准确性而言，大数据的动态性对现有的增量式机器学习方法提出了巨大的挑战。例如，在流式数据处理中，如何在发生概念漂移时及时调整数据分析策略并实现知识库的自适应更新，仍然是一个具有挑战性的课题。

针对大数据的动态增量问题，可以构建一个包括训练学习、执行预测、漂移检测、

漂移理解和漂移自适应的多步骤自适应学习模型。这类模型的重点和难点在于漂移理解与漂移自适应。在漂移理解方面，可以融入高层次的、符合认知的方法，如采用粗糙集、模糊集和商空间等粒计算方法，建立不同粒度层次下的漂移认知模型，实现符合人类认知的层次化概念漂移理解。针对漂移自适应问题，可以通过构建有效的知识距离度量方法，来度量概念漂移的距离与方向，同时综合运用进化计算和神经网络等方法，构建与问题相符的参数自适应模型，实现对学习模型的演化更新。

综合来看，大数据的多样性和动态性为智能决策带来了多方面的挑战。多源异构数据的融合分析、信息协同决策、流式数据处理中的概念漂移应对，都是当前大数据研究的前沿课题。通过深入研究这些问题，不仅能够提升大数据分析和决策的效率和准确性，还能推动大数据技术的进一步发展，为各行业的智能应用提供更强有力的支持。

1.3.3　大数据极弱监督性带来的挑战

大数据的快速增长不仅使数据量急剧增加，还导致了数据的极弱监督性甚至非监督性的问题。在分类学习中，极弱监督性的问题主要表现在两个方面：首先，由于标记数据的稀缺，无法正确和详尽地反映整体数据集的特征，导致学习模型的泛化能力较弱；其次，标记数据的匮乏使得在构建多分类器时，无法充分体现分类器的多样性，进而影响集成学习的效果。极弱监督性的问题强调了无监督学习方法在大数据环境下的重要性，特别是聚类算法的应用价值。无监督学习方法不依赖预先标记的数据，因此在标记数据稀缺的情况下显得尤为适用。此外，大数据的增量性不仅体现在数据样本数量的增加上，还包括属性的不断增加，这进一步增加了处理的复杂性。为了应对这些挑战，可以利用多视角信息、相似领域信息和先验知识，通过大数据耦合与关联分析、大数据与经验知识融合等技术来增加额外的监督信息。这些方法可以在一定程度上缓解标记数据稀缺带来的问题。三支决策模型体现了一种渐进决策的思想，通过逐步利用少量标签信息或领域专家知识，提供了一种有效的解决方案。此外，基于粒计算的方法构建多粒度聚类分析算法模型，也为大数据属性增量式聚类提供了新的解决思路。通过这些方法，可以有效提升大数据分析和决策的效率和准确性，推动大数据技术的进一步发展，为各行业的智能应用提供更强有力的支持。

1.3.4　大数据中隐私问题带来的挑战

在大数据时代，数据成为一种新的资源，推动了科技、商业和社会的快速发展。然而，伴随着大数据的广泛应用，隐私问题也浮出水面。隐私问题不仅关系到个人信息保护，更关系到社会的信任基础和法律框架的完善。

1. 大数据的规模和多样性使得隐私保护变得异常复杂

传统的隐私保护方法主要依赖于对数据的匿名化处理，然而在大数据环境下，匿名化处理的有效性受到了严重挑战。大数据的特征之一是数据来源广泛、类型多样，来自不同渠道的数据可以通过数据融合和数据挖掘技术相互关联，从而重建个体的完整画像，即使原始数据已经被匿名化。研究表明，通过少量的匿名数据和公开的辅助信息，攻击

者可以成功地重新识别个体。因此，简单的匿名化处理已经不足以应对大数据环境下的隐私威胁。

2. 大数据的广泛应用带来了数据滥用的风险

在商业领域，企业通过收集和分析用户数据，以提供个性化的服务和产品。然而，这种数据收集和利用往往缺乏透明度，用户难以了解其数据被如何使用和分享。更糟糕的是，一些企业甚至在未获得用户同意的情况下，非法收集和出售用户数据，导致用户隐私权受到严重侵害。例如，Facebook 的剑桥分析丑闻引发了全球范围内对数据隐私问题的广泛关注。此外，政府和公共机构也在利用大数据进行决策和管理，然而这种数据的收集和使用同样面临滥用的风险，可能侵犯公民的隐私权。

3. 大数据的安全性问题也对隐私保护提出了巨大挑战

数据泄露事件频繁发生，导致大量个人信息被非法获取和利用。黑客攻击、内部人员泄密、系统漏洞等都可能导致数据泄露。大数据的集中存储和处理使得其成为攻击者的目标，一旦发生数据泄露，影响范围和后果都非常严重。例如，2017 年美国信用报告机构 Equifax 发生的数据泄露事件，导致 1.43 亿人的个人信息被泄露，给受害者带来了巨大的经济和心理损失。此外，随着物联网技术的发展，越来越多的设备连接到互联网，产生了海量的数据，这些设备和数据也面临着安全威胁，一旦遭到攻击，将对个人隐私造成严重影响。

4. 大数据中的隐私问题还带来了伦理和法律的挑战

首先是伦理问题，大数据技术的发展和应用需要在隐私保护和数据利用之间找到平衡。一方面，数据的广泛利用可以带来巨大的社会效益，如提高医疗水平、优化城市管理等；另一方面，过度的数据收集和利用可能侵犯个人隐私权，导致伦理问题。如何在技术进步与伦理规范之间找到平衡，是一个亟待解决的问题。其次是法律挑战，现有的法律框架往往无法有效应对大数据时代的隐私问题。例如，许多国家的隐私保护法律仍然停留在传统的数据处理模式上，无法应对大数据环境下的数据流动和处理复杂性。如何建立和完善适应大数据时代的隐私保护法律，是各国政府面临的重要任务。

此外，大数据中的隐私问题还涉及国际间的合作与协调。不同国家和地区在隐私保护方面的法律和政策存在差异，这给跨国数据流动和合作带来了挑战。例如，欧洲的《通用数据保护条例》（GDPR）以其严格的隐私保护要求著称，而美国的隐私保护法律相对宽松，这种法律和政策的差异可能导致跨国企业在数据处理方面面临复杂的法律环境。如何在国际层面上协调和统一隐私保护标准，是一个亟待解决的问题。

扩展阅读 1.1

大数据决策在商业领域的应用案例

大数据中的隐私问题带来的诸多挑战，需要多方面的努力和创新来应对。首先，技术创新是解决隐私问题的重要手段。例如，差分隐私技术通过在数据中引入噪声，确保在数据分析过程中无法识别具体个体，从而有效保护隐私。其次，企业和机构需要加强数据管理和安全措施，确保数据在收集、存储、传输和处理过程中的安全性。

本章习题

1. 请简述大数据的定义及其 4V 特性。
2. 列举两个利用大数据进行决策的典型案例，并简要说明它们的应用效果。
3. 大数据决策有哪些特点？请简要解释。
4. 简述大数据在工业领域的应用及其带来的决策优化效果。
5. 什么是智能决策支持系统（IDSS）？它是如何提升决策有效性的？

即测即练

自学自测　扫描此码