第1章

数据挖掘概论

1.1 什么是数据挖掘

1.1.1 对数据挖掘的需求

随着信息技术的迅速发展，人类社会已进入大数据时代。根据互联网数据中心（IDC）的研究，在未来几年，受到物联网设备的普及和生成式人工智能等技术的驱动，数据量将继续大幅增长，到 2028 年全球数据量将增长到 393.8ZB[1]。如何从海量数据中挖掘出有效的知识，成为工业界和学术界共同关注的问题。数据挖掘（data mining）作为一种系统的知识发现过程，能够帮助人们从海量数据中提取有价值的模式与信息，是现代社会利用海量数据所需的关键技术。

数据科学家 Clive Humby 提出“数据是新时代的石油”（Data is the new oil），这一观点揭示了数据在现代社会的重要地位，但也指出数据就像石油一样，只有经过处理和分析，才能释放真正的价值[2]。数据挖掘正是从数据中提炼有用信息、发挥数据价值的关键技术。数据挖掘的应用已遍及各个行业，涵盖商业经济、科学研究、公共管理的方方面面。例如，在商业领域，电商、银行、民航等企业积累了大量用户行为和交易数据，通过分析用户偏好，可以制定针对性的营销策略和产品推广计划。在科学研究中，数据挖掘已成为加速创新、提升研究效率和拓展知识边界的重要工具，通过自动化的数据分析和知识发现，帮助科学家从数据中提取出隐藏的模式和有价值的信息，推动科学进步。在公共管理领域，通过从海量的公共数据中提取有价值的信息，帮助决策者优化政策、提高资源配置效率、促进社会治理。

数据挖掘善于处理海量数据，提升了大数据分析的深度和广度。大数据具有四个主要特征：数据量（volume）、多样性（variety）、速度（velocity）和价值（value）[3]。大数据的“4V”特征对传统的数据处理方法提出了挑战。在数据量方面，现代数据规模庞大，传统的小样本分析方法在存储、计算和处理海量数据方面能力不足。在多样性方面，现代数据来源广泛，类型多样，涉及文本、图像、视频等非结构化数据，传统方法缺乏处理和分析非结构化数据的能力。在速度方面，数据实时生成和更新速度快速提升，传统方法难以满足实时性需求。在价值方面，要发挥大数据的价值，需要在降低成本、提高收益、改善运营模式等多个

方面综合考虑，传统方法难以应对。数据挖掘赋予大数据意义，帮助从庞大的数据集中提取出有价值的知识。

现代社会的各个领域对数据挖掘提出了强烈的需求，而大数据的“4V”特征及传统方法的局限进一步推动了数据挖掘技术的发展。数据挖掘不仅是数据处理和分析的工具，更是推动决策和发现的核心技术。通过数据挖掘，可以从海量数据中自动挖掘出有价值的模式和信息，为经济发展、科学研究和社会治理等提供强有力的支持。

1.1.2 数据挖掘的起源与发展

数据挖掘起源于20世纪80年代后期和90年代初期，随着数据存储和计算能力的快速发展，对数据分析的需求愈加迫切。传统的统计学研究数据的收集、整理、分析和解释，帮助研究者从数据中做出推断。虽然统计学提供了分析数据的基础，但随着数据规模、复杂性和多样性的快速增长，融合统计分析、机器学习、数据库、信息检索等技术的数据挖掘应运而生。数据挖掘的发展历程大致分为以下几个阶段。

20世纪80年代，随着数据库规模的增大，传统的统计方法和数据库查询功能不足以从大量数据中提取有用信息。为了应对这一挑战，计算机领域的学者提出知识发现(knowledge discovery in databases，KDD)的概念。知识发现是一个多步骤的过程，包括数据选择、数据清理、数据集成、数据挖掘和模式评估等多个阶段。Fayyad等系统地描述了KDD过程，将数据挖掘定义为从数据库中自动提取有意义模式的步骤[4]。

20世纪90年代，随着计算能力的提升和机器学习技术的进步，数据挖掘迅速发展，许多数据挖掘算法被开发出来，包括决策树、关联规则发现、聚类算法、支持向量机等。这一时期的数据挖掘算法主要关注模式的发现、分类、聚类等，解决一些具体的业务需求，如零售业的客户购买模式分析、金融行业的欺诈检测等。

进入21世纪，互联网的发展使得数据规模进一步扩大，数据类型也更为丰富，包含文本、图像、视频等非结构化数据。此时大数据的概念逐渐兴起，数据挖掘面临新的挑战。传统的算法和工具无法满足大数据处理的需求，数据挖掘开始引入并行计算和分布式计算技术，Hadoop和MapReduce框架的出现使得大规模数据处理成为可能。基于云计算的分布式数据库和计算模型帮助数据挖掘适应海量数据的需求。

2010年之后，随着深度学习的崛起，数据挖掘迎来新的发展。深度学习通过多层神经网络结构，能够更好地处理图像、文本等非结构化数据。这一时期，数据挖掘的重点逐渐向人工智能方向靠拢，从以发现模式为主，转向预测和智能决策。深度学习在语音识别、图像处理、文本分析等领域的应用使得数据挖掘技术更加智能化，广泛应用于社交媒体分析、个性化推荐、医疗诊断等领域。

近年来，随着自动化机器学习(AutoML)的发展，数据挖掘的自动化程度进一步提高。AutoML通过自动化选择数据预处理步骤、特征工程、选择算法和参数优化，降低了数据挖掘的门槛，使得数据挖掘的应用更加普及。这一阶段的数据挖掘工具更注重用户友好性和实用性，使得非专业人士可以应用数据挖掘技术处理业务问题。

1.1.3 数据挖掘的概念

数据挖掘是一个多学科交叉领域，对数据挖掘的理解也因不同研究视角而有所差异。

一些学者对数据挖掘进行了描述或定义。

Fayyad 等将数据挖掘定义为“从大量数据中自动或半自动地发现有趣的模式和知识的过程”[4]。这一定义强调了数据挖掘的自动化特点，指出其主要目标是发现“有趣”的模式。

Witten 与 Frank 将数据挖掘定义为“数据挖掘是使用计算技术从数据中发现模式的过程”[5]。这个定义突出了数据挖掘依赖计算工具来揭示数据中的模式。

韩家炜（Jiawei Han）在著作《数据挖掘：概念与技术》中，对数据挖掘的定义是“数据挖掘是从大量数据中挖掘有趣模式和知识的过程”[6]。这一定义与 Fayyad 的观点基本一致，突出了“有趣模式”的发现过程。

陈封能在著作《数据挖掘导论》中，将数据挖掘定义为“数据挖掘是从大型数据库中自动发现有用信息的过程”[7]。这一定义关注数据挖掘的“自动发现”特性，强调了“有用信息”的提取。

Aggarwal 在著作《数据挖掘：原理与实践》中，将数据挖掘定义为“数据挖掘是指对数据进行收集、清洗、加工和分析并从中获取有用知识的过程”[8]。这一定义涉及数据处理的多个步骤，强调了知识发现的全面性。

基于对数据挖掘的理解，可以认为数据挖掘是一个从海量数据中自动或半自动地发现有趣模式或知识的过程。它涉及数据的收集、清洗、分析、应用等步骤，利用统计分析、机器学习、数据库等技术，支持对实际问题的分析、优化和决策支持。

1.1.4 数据挖掘的学科特性

数据挖掘是问题导向的，它融合不同领域的技术，为大规模数据的知识发现提供全面的解决方案。数据挖掘是一门交叉学科，主要融合了计算机科学、统计学、机器学习、数据库、信息检索等多个学科的知识和技术。

数据挖掘的核心是处理大规模数据，这需要计算机科学中的算法设计、数据结构和复杂性分析。计算机科学还提供了数据挖掘中非常重要的分布式计算和并行计算技术，这些技术使得处理海量数据变得可行。

数据挖掘处理的数据大多存储在数据库中，因此数据库系统在数据挖掘中起到关键性作用。数据库技术提供了高效的存储和访问大量数据的方法，同时还开发了用于数据预处理的抽取、转换、加载（ETL）工具，使得数据挖掘模型能够在清理后的数据上构建。此外，关系数据库和 NoSQL 数据库的扩展技术，如数据仓库和在线分析处理（OLAP），也为数据挖掘提供了便捷的分析环境。

统计学为数据挖掘提供了许多基础方法，尤其在数据分析和模式识别方面。统计学中的探索性分析、假设检验、回归分析、聚类分析等技术在数据挖掘中被广泛使用，用于发现数据中的潜在规律和关系。统计学强调数据的可解释性和模型验证，为数据挖掘结果的可靠性提供了重要支持。

机器学习是数据挖掘的核心支柱之一，其主要任务是让计算机从数据中自动学习模式和知识。监督学习（如分类、回归）、无监督学习（如聚类）以及强化学习等机器学习方法广泛应用于数据挖掘。机器学习的发展推动了数据挖掘技术的进步，使得数据挖掘方法更加智能和高效。随着深度学习的发展，许多人工智能算法被应用到数据挖掘中，用于实现复

杂的模式识别和知识发现任务。深度神经网络在图像、语音和自然语言处理等领域的突破，极大地丰富了数据挖掘的技术体系，使得从复杂数据（如文本、图像、视频）中提取有用信息成为可能。

信息检索与数据挖掘在文本数据处理上有较多的重叠。特别是在文档分类、文本聚类、主题模型等任务中，数据挖掘与信息检索的方法互相借鉴。信息检索技术支持数据挖掘的文本分析任务，如文本挖掘、社交媒体分析等。

1.1.5 数据挖掘的未来

过去几十年，数据挖掘的理论、技术和方法日渐成熟，在很多领域得到广泛应用，成为社会经济发展的关键技术。近年来，随着大数据和人工智能的快速发展，数据挖掘在技术、应用、方法等方面均展现出更大的潜力，但也面临法律、伦理等方面的挑战。可以预计在未来较长时间，数据挖掘将继续推动技术创新，在智能社会和数字经济的建设中发挥重要作用。

在技术方面，数据挖掘受益于多项前沿技术的发展，包括自动化机器学习、深度学习、复杂数据挖掘、大数据与云计算等。自动化机器学习旨在简化机器学习模型的构建和训练过程，该技术将进一步成熟，特征工程和模型优化逐渐标准化，从而提升模型的普适性和准确性。深度学习扩展了数据挖掘的能力，尤其在图像、语音、文本等非结构化数据中表现出色。深度学习结合强化学习、自监督学习等先进方法，能够实现更深层次的知识发现。云计算提供了强大的分布式计算资源，使得大规模数据处理更为便捷。大数据与云计算的结合将推动实时数据挖掘的发展，通过分布式存储和并行计算，降低成本和复杂性。

在应用方面，数据挖掘在应用场景上不断拓展，将涵盖实时数据分析和知识图谱等前沿应用。实时数据挖掘使得物联网、金融监控等场景中的快速响应成为可能。流数据处理技术将进一步优化，能更迅速地从动态数据中提取信息。图数据挖掘通过分析图结构数据（如社交网络、基因网络等）揭示深层关系。知识图谱在智能搜索、问答系统中得到了广泛应用，实现复杂知识推理和分析。

为满足复杂数据挖掘需求，数据挖掘在方法层面不断创新，特别是在隐私保护和模型解释性方面取得突破。随着数据隐私问题日益受到重视，隐私保护技术在数据挖掘中变得必不可少。模型的可解释性在医疗、金融等领域尤为重要。未来的数据挖掘技术将逐渐增强解释性，帮助用户理解复杂模型的决策过程，提升模型在应用中的透明性和可控性。

数据挖掘技术的广泛应用引发了关于数据伦理的讨论。为了确保数据的合法、合规使用，数据伦理在数据挖掘发展中将扮演重要的角色。数据使用政策的透明性也将进一步提高，增强用户对数据使用的信任。数据挖掘的伦理标准将继续完善，特别是在涉及敏感数据和个人隐私的领域。未来可能会制定更详细的数据挖掘操作规范，确保技术应用符合道德和法律要求，避免不公平或歧视性算法的使用。

1.2 数据挖掘过程

广义的数据挖掘是一个完整的知识发现过程，它不仅包括狭义的数据挖掘建模阶段，还包括数据准备、数据预处理、模型评估、部署等多个阶段。随着数据挖掘的发展，人们提

出多种不同的数据挖掘方法论，用以指导数据挖掘实践。例如，KDD 过程模型强调知识发现的整体流程。CRISP-DM 提出了一个灵活、迭代和分阶段的流程，适用于各种行业。SEMMA 侧重技术流程。ASUM-DM 引入项目管理思想，更适合大规模商业项目。这些方法论帮助在不同层面实现数据挖掘项目的系统化和标准化，使数据挖掘过程更加规范、高效，更具有针对性。

1.2.1 KDD 过程模型

1996 年，Fayyad 等提出 KDD(knowledge discovery in databases)过程模型[9]，将 KDD 定义为一个系统化的过程，而不仅仅是数据挖掘算法的应用，开启了对数据挖掘过程模型的研究。KDD 明确了数据预处理、数据转换、数据挖掘、模式评估等环节，并强调了知识发现过程的迭代性。随着数据量和计算能力的增加，KDD 成为后续数据挖掘方法论的基础。

在 KDD 过程模型中，KDD 是从原始数据中获得有价值知识的完整过程。数据挖掘是 KDD 过程的一个阶段，应用算法从数据中发现模式。目前学者对数据挖掘多采用广义的理解，将数据挖掘看作从海量数据中获得有趣模式、知识的过程。

KDD 从获取数据开始，目的是从数据中提取知识，这一过程包括若干基本步骤[9]，如图 1-1 所示。

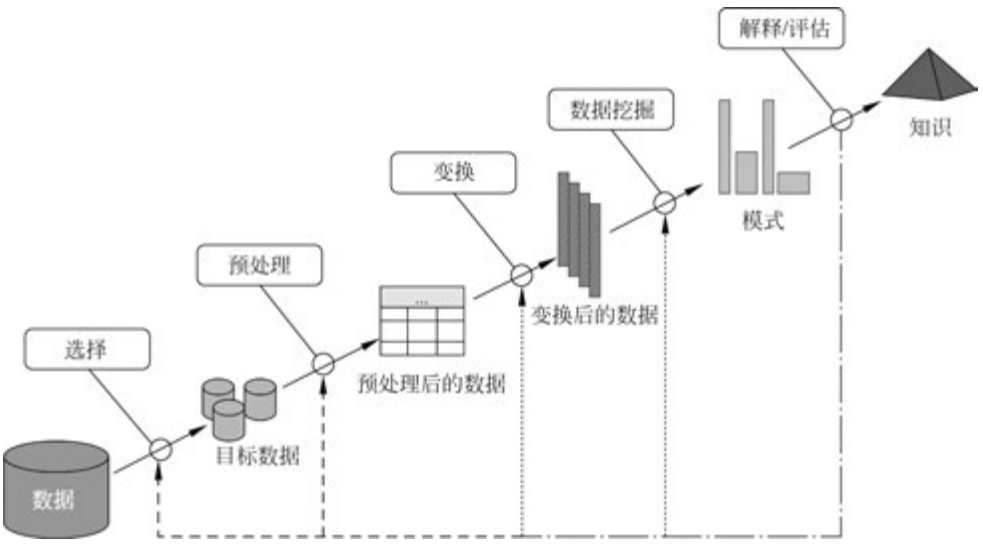

图 1-1 KDD 过程模型

KDD 过程包括 9 个基本步骤，分别是：

(1) 了解问题所在的领域，获得先验知识，从客户的角度确定 KDD 流程的目标。

(2) 创建目标数据集。选择一个数据集，确定需要关注的变量，或进行抽样获得样本子集。目标数据集是进行知识发现的基础。

(3) 数据清理和预处理。基本操作包括去除噪声、收集必要信息了解噪声的情况，或对噪声进行说明，决定处理缺失数据的策略，说明时间序列信息和已知的变化。

(4) 数据缩减和投影。根据任务目标，找到有用的特征来表示数据。通过降维或变换，

减少变量的数量，或找到数据的不变表示。

(5) 将 KDD 流程的目标与特定的数据挖掘方法相匹配，如分类、回归、聚类等。

(6) 探索性分析，选择模型和假设。选择数据挖掘算法，并选择用于搜索数据模式的方法。这一过程包括决定合适的模型和参数，将特定的数据挖掘方法与 KDD 过程的总体目标相匹配。

(7) 数据挖掘。以一种特定的表示形式或一组表示形式搜索感兴趣的模式，包括分类规则或分类树、回归和聚类等。

(8) 解释挖掘出的模式。这一步还可能涉及对提取的模式和模型进行可视化，或根据提取的模型对数据进行可视化。

(9) 根据提取的知识采取行动。行动的方式很多，例如直接使用这些知识，或将这些知识纳入另一个系统以采取进一步行动，或者只是将其记录下来并报告给相关方。这一过程还包括检查和解决与先前知识之间的潜在冲突。

KDD 过程的各个步骤并不是完全线性的，可能涉及大量迭代，在后面阶段发现存在问题，可能需要返回到前面的某一步重新执行。经过多次迭代，直到达成项目目标。

KDD 是早期经典的数据挖掘方法论，为数据挖掘项目的系统性和规范性提供了基础，为数据挖掘实践提供了支撑。它强调知识发现的系统性，流程完整，具有灵活性和迭代性。但缺乏对各个步骤的详细指导，对领域知识的依赖性较强，逐渐被一些更现代、更高效的框架补充和取代。

1.2.2 CRISP-DM 方法论

跨行业数据挖掘标准过程(CRoss-Industry Standard Process for Data Mining，CRISP-DM)是目前广泛使用的数据挖掘方法论之一，它是由 SPSS、NCR、DaimlerChrysler 等公司组成的专家委员会联合提出的，2000 年发布了 CRISP-DM 1.0 用户指南[10]。在过去的二十多年，CRISP-DM 是数据挖掘的事实标准[11]，广泛适用于目标明确和过程驱动的数据科学项目。

CRISP-DM 是一个通用的、结构化的过程模型，它是行业中性的，不依赖具体的领域知识，可以用于指导不同行业的数据挖掘项目。它将数据挖掘项目划分为多个阶段，每个阶段有明确的目标、任务和输出，这种结构化的流程模型有助于项目团队系统性地解决问题。CRISP-DM 流程还具有高度的灵活性和迭代性，在项目进行过程中可以根据需要调整流程，以应对数据复杂性、业务需求变化的挑战。

CRISP-DM 是一个层级化结构模型，从一般到特殊进行了四层抽象，分别是：阶段(phase)、一般任务(generic task)、具体任务(specialized task)、过程实例(process instance)。在第一层将数据挖掘项目划分为若干阶段。在第二层描述每个阶段包含的一般任务。在第三层根据实际的特定场景，将一般任务明确为需要执行的具体任务。第四层是对实际数据挖掘工作中行动、决策、结果的记录。

CRISP-DM 参考模型对数据挖掘项目生命周期进行了概括，将数据挖掘项目划分为 6 个阶段，各个阶段存在交互关系，如图 1-2 所示。

图 1-2 只是象征性地画出阶段之间的重要关系，各阶段的顺序不是一成不变的，有时需

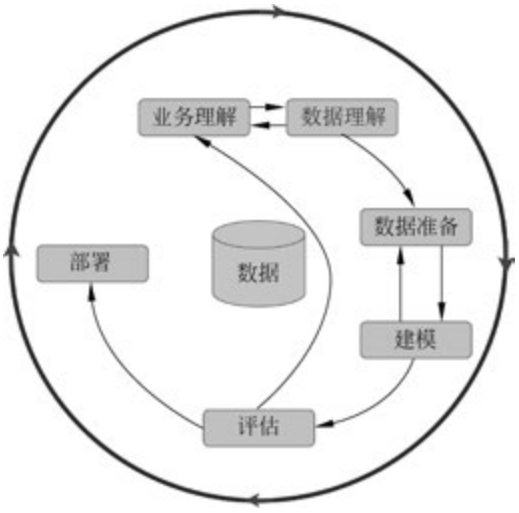

图 1-2 CRISP-DM 生命周期模型

要在不同阶段之间进行往返，项目可能包含多次迭代和循环。外层的圆圈表示数据挖掘的周期性，解决方案部署后并不意味着整个过程结束了，而是会周而复始多次执行。在以前过程中发现的问题，以及从已部署的解决方案中吸取的经验教训，会引发新的、更有针对性的业务问题，又启动新一轮的数据挖掘过程。

CRISP-DM 强调从业务实际出发，通过数据挖掘实现业务目标。项目周期包括六个阶段，每个阶段的描述如下。

(1) 业务理解(business understanding)

这一阶段的重点是从业务角度理解项目的目标和需求。根据业务目标和需求，明确数据挖掘问题，确定数据挖掘的目标，并拟定实现目标的初步计划。

(2) 数据理解(data understanding)

根据数据挖掘目标和相关资源，进行初步的数据收集，对数据进行探索，了解数据，发现数据质量方面存在的问题。获得对数据的初步理解，发现感兴趣的数据子集，形成对隐藏模式、信息的初步假设。

(3) 数据准备(data preparation)

在对数据初步了解的基础上，根据建模工具对数据的要求，对数据进行选择、清洗、集成、转换等操作，为数据挖掘模型提供满足质量要求的数据集。数据准备任务没有预先设定的顺序，可能需要多次执行。

(4) 建模(modeling)

通常情况下，同一类型的数据挖掘问题有多种技术可供选择。在这一阶段选择和应用各种建模技术，并将模型参数调整到最佳值。有些数据挖掘技术对数据形式有特定要求，因此通常需要回到数据准备阶段。

(5) 评估(evaluation)

在项目的这一阶段，已经建立了一个(或多个)模型，仅从数据分析的角度看，模型似乎有很高的质量。然而在对模型进行部署之前，需要从业务角度对模型进行更彻底的评估，并审查构建模型的步骤，检查是否遗漏了业务问题，确保模型能正确实现业务目标。在这

一阶段结束时，决定如何应用数据挖掘结果。

（6）部署（deployment）

项目的最终目标并不是创建模型，而是在实际业务中发挥作用，实现业务目标。在这一阶段将数据挖掘模型以及挖掘得到的知识应用于实际业务，即部署。部署的形式千差万别，可能是生成一份报告，也可能需要将模型嵌入决策过程，或者在企业实施可重复的数据挖掘流程。很多时候是客户执行部署步骤，在这种情况下，数据分析师需要告知客户如何使用模型和知识。

CRISP-DM 用户手册对六个阶段的目标、任务与输出进行了详细论述，主要内容见表 1-1。

表 1-1　CRISP-DM 六个阶段的目标、任务与输出

阶段	目　　标	主 要 任 务	输　　出
业务理解	深入理解业务问题并明确数据挖掘的目标	1. 确定业务目标	背景，业务目标，业务成功准则
		2. 态势评估	可用资源，需求、假设与约束，风险与应对，成本与收益
		3. 确定数据挖掘目标	数据挖掘目标，数据挖掘成功标准
		4. 制订项目计划	项目计划，工具与技术评估
数据理解	初步了解和分析数据的特征和质量	1. 收集初步数据	初步数据收集报告
		2. 描述数据（格式、数量、字段等）	数据描述报告
		3. 探索数据（数据分布、变量关系等）	数据探索报告
		4. 数据质量评估（缺失值、异常值、噪声、错误）	数据质量报告
数据准备	处理并准备建模所需的数据集	1. 选择数据	数据选择/排除原因
		2. 清理数据（处理缺失值、异常值、重复值）	数据清理报告
		3. 数据生成与变换	派生属性与生成记录
		4. 数据集成（合并多数据源）	合并数据
		5. 数据格式转换	规范化的数据
建模	构建和优化满足目标的模型	1. 选择建模技术（如决策树、神经网络等）	建模技术，模型假设
		2. 设计测试方案	测试方案
		3. 建模	模型参数，模型描述
		4. 评估模型性能	模型的性能
评估	确保模型符合业务目标并具有实用性	1. 使用业务场景评估模型	模型评估报告，批准的模型
		2. 评审建模过程	评审结果报告
		3. 确认是否进入部署阶段	下一步行动决策
部署	将模型投入业务应用并监控其表现	1. 制订部署计划	部署计划
		2. 制订监督与维护计划	监督与维护计划
		3. 编制最终报告	项目最终报告
		4. 项目总结	项目总结报告

CRISP-DM 是一种经典的数据挖掘方法论，凭借其标准化、跨行业适应性强和强调业务理解等优点，成为数据挖掘领域的事实标准。它重视业务理解，确保数据挖掘过程始终

围绕业务目标展开，避免了仅从技术角度进行数据分析。它提供了一个清晰、标准化的过程框架，帮助从业者以系统化的方式组织数据挖掘项目，很多数据挖掘工具（如 SPSS Modeler、RapidMiner、KNIME 等）都以 CRISP-DM 为参考框架。它强调数据挖掘是一个迭代、灵活的过程，能够根据项目的不同需求进行调整。它强调“跨行业”，被金融、电信、零售、制造等多个行业广泛采用。然而，随着大数据、实时数据和深度学习等技术的发展，CRISP-DM 的局限性逐渐显现，尤其是在高维数据、大数据和实时数据处理方面的不足。CRISP-DM 需要与新的技术和方法相结合，以适应现代数据挖掘的需求。

1.2.3 SEMMA 方法论

SEMMA 是 SAS 公司提出的一种数据挖掘方法论[12]。SEMMA 是 Sample、Explore、Modify、Model、Assess 五个单词的首字母缩写。SEMMA 是 SAS 公司为它的数据挖掘产品设计的框架，但也具有一定的通用性，能够为数据挖掘过程提供系统指导，帮助分析人员从数据中挖掘出有用的信息和模式。

SEMMA 将数据挖掘过程划分为五个步骤，如图 1-3 所示。

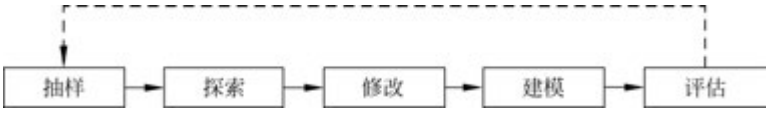

图 1-3 SEMMA 方法论

每一步都有明确的目标和任务，以下是各步骤的介绍。

（1）抽样（sample）

在数据挖掘的初始阶段，从大量数据中选择一个代表性的子集，以提高数据处理效率并降低计算复杂度。这一过程确保选择的数据足够有代表性，同时减少计算成本。

（2）探索（explore）

在探索阶段，分析人员使用统计方法和数据可视化手段理解数据的分布、变化趋势和异常点。此过程帮助发现数据的基本特征和潜在关系，为后续建模奠定基础。

（3）修改（modify）

通过数据清洗、转换、归一化和特征工程等手段，对数据进行处理。目的是优化数据质量和结构，提高模型的效果。此阶段可能包括去除噪声、处理缺失值以及数据编码等操作。

（4）建模（model）

在建模阶段，分析人员选择适当的数据挖掘模型，如决策树、神经网络等，来分析数据和提取模式。模型的选择取决于具体的任务需求，如分类、回归或聚类。

（5）评估（assess）

评估阶段对建好的模型进行性能评估，以确保模型的有效性和可靠性。通常会采用交叉验证、混淆矩阵等方法来评估模型的准确性、稳定性等指标，以确定模型的实际应用价值。

SEMMA 提供了一个有条理的、分步的框架，帮助数据分析人员系统性地完成数据挖掘项目，注重数据理解、预处理和模型选择，以实现对数据模式的深入洞察。与 CRISP-DM 相比，SEMMA 主要关注数据挖掘的技术流程，以模型开发为核心，尤其适用于采用 SAS 软

件的场景。SEMMA 没有包括业务理解和部署阶段，在实现业务目标和项目管理方面略显不足。

1.2.4 ASUM-DM 方法论

为了应对数据分析项目的复杂性和多样化需求，2015 年 IBM 在 CRISP-DM 的基础上提出 ASUM-DM（analytics solutions unified method for data mining）方法论。相比于 CRISP-DM，ASUM-DM 增加了许多面向企业应用的阶段和操作步骤，支持大型数据分析项目的端到端实施。ASUM-DM 是对 CRISP-DM 的扩展，包含一些新的和改进的措施、模板、规范和活动。ASUM-DM 也是应用中立的，可用于各种领域和行业。

ASUM-DM 将项目管理与数据分析活动结合起来，提供了实施和执行数据挖掘及预测分析不同阶段的分步指南，包括活动、职责、角色分配、模板等，目的是提高数据挖掘的效率，缩短项目实施时间。

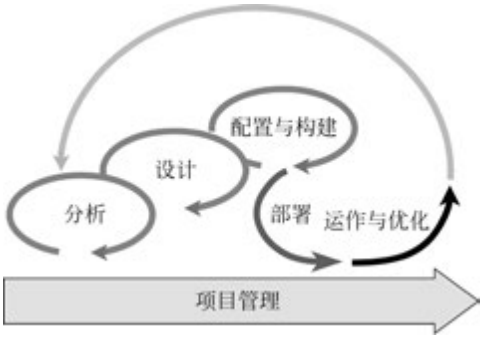

图 1-4 ASUM-DM 数据分析项目流程

ASUM-DM 定义了数据挖掘和预测分析的五个阶段，所有阶段都由项目管理伴随和监控，负责各个阶段之间的一致性、协作和沟通。五个阶段是：分析（analyze），设计（design），配置与构建（configure & build），部署（deploy），运作与优化（operate & optimize），如图 1-4 所示。每个阶段可能存在多次迭代，阶段之间也存在多次往返。

在分析阶段，定义数据挖掘和预测分析解决方案的要求和目标。参与解决方案的每个人必须就要求和目标达成一致。

在设计阶段，定义各个解决方案组件及其相互依赖性，确定所需的资源并提供开发环境。创建该解决方案的第一个原型。

在配置与构建阶段，配置、实施和测试各个组件，并构建完整的解决方案。

在部署阶段，将解决方案交到用户手中，并为持续运行做好准备。

在运作与优化阶段，用户实际使用解决方案，使用过程包含维护任务和检查点，确保解决方案在其生命周期内保持良好状态，并通过不断优化消除缺陷。

ASUM-DM 对每个阶段进行了逐层分解，对每个阶段提供了说明、任务分解、团队分工、交付标准等。例如，部署阶段的工作流如图 1-5 所示。

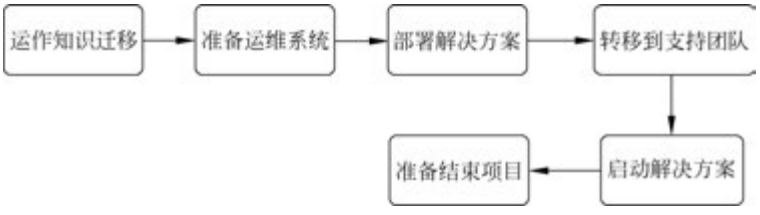

图 1-5 ASUM-DM 部署阶段工作流