复杂系统影响因素研究的数据驱动分析方法

李海林　林春培　编著

清華大學出版社
北　京

内 容 简 介

本书聚焦于复杂系统影响因素研究的数据驱动分析方法(DAC)，为应对大数据和人工智能时代复杂系统问题提供创新思路与实用工具。第 1 章阐述了传统分析方法在处理复杂系统多变量、非线性和动态变化等特征时的不足，而 DAC 凭借先进的数据挖掘和机器学习算法，通过数据获取、数据处理与变量测量、聚类分析、决策树分析和贝叶斯网络分析 5 个关键阶段(步骤)，为决策制定和优化助力。第 2 章强调指标选取的依据、选取原则等，依据数据类型选择合适量化方法，并通过实例演示如何将实际问题转化为可量化数据集，保障后续分析质量。第 3 章详细介绍数据采集、统计分析、变量选取、校准处理(引入云校准概念)等数据预处理内容。第 4 章讲解基于聚类算法的异质性群体的多种分析。第 5 章使用决策树分析了异质性群体对象的影响因素交互效应。第 6 章运用贝叶斯网络和相关算法探究变量间的作用关系和影响路径。第 7 章通过后发企业创新绩效案例分析，展示 DAC 在实际研究中的应用优势。

本书特色鲜明，内容紧密围绕解决复杂管理问题，案例丰富且分析透彻，从多领域实际问题出发，旨在增强读者对方法的理解与应用能力。本书中代码示例详细，可操作性强。本书适用于工商管理、管理科学与工程、经济与金融等专业的本科生和研究生，为他们开展学位论文研究和学术探索提供新颖视角和方法，帮助他们掌握这一跨学科融合的研究范式。

图书在版编目(CIP)数据

复杂系统影响因素研究的数据驱动分析方法 / 李海林, 林春培编著.

北京 : 清华大学出版社, 2025. 4. -- ISBN 978-7-302-68545-6

I. TP274

中国国家版本馆 CIP 数据核字第 2025SV3022 号

责任编辑：王　定

封面设计：周晓亮

版式设计：思创景点

责任校对：马遥遥

责任印制：刘　菲

出版发行：清华大学出版社

网　　址：https://www.tup.com.cn，https://www.wqxuetang.com

地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084

社 总 机：010-83470000　　邮　　购：010-62786544

投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn

质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn

印 装 者：大厂回族自治县彩虹印刷有限公司

经　　销：全国新华书店

开　　本：185mm×260mm　　**印　　张：**15　　**字　　数：**245 千字

版　　次：2025 年 4 月第 1 版　　**印　　次：**2025 年 4 月第 1 次印刷

定　　价：69.80 元

产品编号：110651-01

前言

PREFACE

当今时代，云计算、数字化、5G 和人工智能等信息技术迅猛发展，人类社会经济系统所产生的信息与数据呈指数级爆炸式增长。在此背景下，如何于海量数据中精准挖掘出有价值的信息与知识，探寻其中的变化规律和因果关系，进而推动更为高效的管理决策，成为数字经济时代迫切需要解决的现实难题。数字经济环境下，客观系统具备动态性、随机性，而人类认知存在模糊性、有限性，二者共同构成了管理研究问题的不确定性内涵。传统那种单一线性的因果关系解释方法，在日益复杂多元的管理情境面前显得力不从心。面向复杂系统影响因素研究的数据驱动分析方法（Data-driven Analysis methods for the study of influencing factors in Complex systems，DAC）应运而生。DAC 融合了定量研究和定性研究的优势，借助数据挖掘、机器学习等大数据技术与方法，对复杂系统关键核心因素的间的作用机制展开研究。它整合了定量研究、定性研究以及大数据分析的多维优势，从数据特征、问题特征和管理决策特征的角度出发，构建了将大数据技术应用于管理实践问题的新型研究框架。这一框架能协助决策者依据实际管理问题情境剖析系统因素间复杂的交互机制，从而实现关键知识发现和最优资源配置。

DAC 涵盖以下 5 个核心模块。

其一，数据采集与预处理模块。在研究问题时，此模块依据对当前复杂系统关键前因和结果特征的筛选与检测，运用爬虫、数据解析、数据库技术等手段，将多源异构数据收集起来并处理成结构型数据。同时，此模块利用数据分析手段对数据集进行异常检测和处理，以此提升数据分析质量，为后续研究奠定坚实的数据基础。

其二，指标量化与校准模块。该模块致力于把原始数据转化为可量化的指标，并对其进行校准，以消除数据间的异质性和不一致性。在实际操作中，该模块根据管理决策问题的需求选择合适的指标体系和量化方法，将数据转化为可计量形式的数据。特别地，该模块通过运用云模型进行校准，充分考虑到指标间的模糊性和不确定性，使得量化后的指标能更精准地反映实际情况。

其三，异质性群组划分模块。该模块将研究对象划分为具有相似特征的异质性群组，以便更好地理解和分析不同群组间的差异及影响机制。在此过程中，该模块采用诸如 K-Means、层次聚类、AP 聚类等聚类算法对研究对象开展聚类分析，并依据轮廓系数、簇内平方和等主流聚类评价准则来确定最佳的群组划分数量，将相似性高的个体归入同一群组，进而深入探索不同群组之间的关系与差异。

其四，决策规则提取模块。此模块旨在从数据中挖掘潜在的决策规则，揭示不同因素对决策结果的影响和作用方式；通过运用决策树算法或其他机器学习方法，分析数据中的特征和标签之间的关系，构建决策规则模型，为决策和预测提供有力支持，帮助决策者深入理解决策背后的因果关系和决策规律。

其五，因果关系识别模块。该模块聚焦于识别和分析不同因素之间的因果关系；利用贝叶斯网络等概率图模型，基于统计推断和概率分析方法，探寻不同因素间复杂的因果关系；通过对数据进行建模和推理，发现潜在的因果路径和影响机制，从而更全面地理解和阐释复杂系统中的因果关系。

本书特色

从内容的实际价值来看，本书紧密围绕解决复杂因素影响下的管理问题这一核心。无论是面对工商管理中的市场策略制定，管理科学与工程中的流程优化，还是面对经济与金融领域的风险评估等问题，本书所阐述的 DAC 都能提供切实可行的解决方案。本书阐述了 DAC 每个模块在解决实际问题中的应用方式，使读者能够清晰地理解如何运用 DAC 应对各种复杂管理情境中的挑战，将理论知识转化为解决实际问题的有力工具。

在案例分析的可理解性方面，本书精心挑选了具有代表性的案例，并对每个案例进行了深入细致的剖析。从案例的背景介绍、问题分析到运用 DAC 解决问题的全过程，本书都进行了逐步讲解。通过清晰的逻辑结构和生动的表述方式，使读者在阅读案例时能够轻松理解复杂因素是如何相互作用的，以及 DAC 是如何在实际场景中发挥作用的。这种案例分析方式就像是为读者搭建了一座桥梁，让读者能够顺利地从理论知识学习过渡到实际应用能力掌握，提升对复杂问题的分析和解决能力。

在代码可操作性方面，本书为读者提供了丰富且易于理解的代码示例。针对 DAC 中涉及的各个技术环节，如数据采集与预处理中的云校准代码、聚类算法中的代码实现、决策树算法和贝叶斯网络相关的代码片段等，本书都进行了详细的展示和讲解。这些代码不仅注释清晰，而且经过精心设计和调试，确保读者能够轻松地将其应用于实践中。无论对有编程基础的读者，还是对正在学习编程的新手来说，这些代码示例都能成为他们快速掌握 DAC 并将其应用于实际问题的有力助手，真正实现了理论与实践的无缝对接。

使用对象

DAC 运用基于云模型的数据校准、聚类、决策树、贝叶斯网络等数据挖掘工具，对复杂因素影响机制研究过程中的数据预处理、多元情境识别与分析、影响因素分析和目标提升路径展开研究，以实现复杂系统影响因素机制研究目的。这种方法体现了管理学、统计学与计算机科学等相关学科的交叉融合，创新发展了传统管理研究范式，已在工商管理、管理科学与工程、经济与金融等学科的复杂系统影响机制问题研究领域中得到有效应用，为相关专业本科生和研究生的学术研究提供了一套新颖且可行的方法。希望这本书能成为广大学子和研究者在探索复杂因素影响机制的道路上的一盏指路灯，助力他们在相关领域

取得更大的研究成果。

编写致谢

在编写本书的过程中，我们得到了诸多教师、学生和朋友的关心与支持，在此向他们表达衷心的感谢。

首先，要感谢张丽萍、万校基、蔡林峰、谭观音等老师。在编写本书的各个阶段，他们凭借深厚的专业素养和丰富的教学经验，给我们提供了大量宝贵的建议。无论是在内容架构的搭建上，还是在具体知识点的阐述上，他们的指导都犹如明灯，照亮了我们在编写中的迷茫之路。他们的支持是我们完成本书不可或缺的力量，他们的智慧和热情深深感染着我们，让我们在面对困难和挑战时充满勇气和决心。

其次，要感谢编者的研究生团队：周文浩、田慧敏、李虎峰、廖杨月、龙芳菊、汤弘钦、黄梦婷、陈多、陈美婷、付梦等。他们在资料整理、方法研究、案例分析、代码实现等各个方面都付出了辛勤的努力，作出了重要贡献。在资料整理过程中，他们认真细致地收集、筛选和梳理大量的文献资料，为本书内容的丰富性和权威性奠定了坚实的基础。在方法研究方面，他们积极探索，勇于创新，为复杂因素影响机制的研究提供了新的思路和方法。在案例分析环节，他们深入剖析案例，使得案例更具代表性和启发性。在代码实现过程中，他们精心编写和调试代码，确保了本书技术内容的可操作性和准确性。他们的努力和付出是本书得以顺利完成的重要保障，他们的才华和专注让本书更加精彩。

本书免费提供教学课件、教学大纲和电子教案，读者可扫描下列二维码获取。

教学课件　　教学大纲　　电子教案

编　者

2025 年 1 月

目录

CONTENTS

第 1 章

导 论

在大数据和人工智能时代，各行业数据呈现爆发式增长。传统分析方法难以应对复杂系统中多变量、非线性和动态变化等特征要求。用于复杂系统影响因素研究的数据驱动分析方法采用先进的数据挖掘和机器学习技术和算法，能够更加全面、准确地识别和量化复杂系统中的关键影响因素，为决策制定和优化提供有力支持。该方法的技术流程主要包括数据获取、数据处理与变量测量、聚类分析、决策树分析和贝叶斯网络分析 5 个关键阶段(步骤)。为有效开展研究，使用者需要掌握扎实的统计学和机器学习理论基础，以及熟练的编程和数据处理技能。同时，使用者还须熟悉并能够配置 Python、Graphviz 和 Netica 等专业软件工具。用于复杂系统影响因素研究的数据驱动分析方法的应用广泛，涵盖城市交通系统优化、生态环境保护、金融风险评估和医疗健康管理等多个领域。随着技术的不断发展和应用场景的拓展，该方法正在催生新的研究方向和应用前景，为解决复杂系统问题提供了创新性的思路和工具。

1.1 背景意义

复杂系统通常涉及大量相互作用的组成部分，表现出多变量、非线性和动

态变化等特征，传统分析方法往往难以全面把握这些组成部分(系统)的本质和行为规律。同时，大数据、人工智能等技术的迅猛发展为人们提供了前所未有的机遇，使得从海量数据中提取有价值信息成为可能。在这一背景下，复杂系统影响因素研究的数据驱动分析方法结合了数据科学、机器学习和复杂系统理论，为解决复杂问题提供了新的视角和工具。

1.1.1 实际背景

随着大数据、物联网、人工智能、云计算、区块链等前沿技术的不断发展，数字经济和智能经济已经成为推动人类社会经济发展的新动能，并展现出加速发展的态势。数字经济和智能经济是基于大数据及其相关技术的创新和应用而形成的经济形态。数据日益凸显其作用和重要性，其生产、开发和应用已经成为新经济发展的关键因素，为新经济的发展提供了核心支持(李政和周希禛，2020)。随着第五次信息技术革命的持续演进，人类的各种行为和生活轨迹被以数据形式记录下来。目前，大数据已经成为一种社会公共资源，并越来越多地被应用于各个领域的研究和分析。同时，大数据作为一种创新工具，帮助人们更高效地完成工作任务(林甫，2019)。在大数据领域，我国正从跟随者转变为并行者，并在一些以 5G 为代表的通信技术、集成电路、互联网金融等领域取得了领先地位。在这样的背景下，以数据为核心的新技术、新产业、新业态、新模式等逐渐兴起并快速发展，不仅为传统经济注入了新的发展动力，也使国民经济更加“数字化”(李政和周希禛，2020)。根据中国信息通信研究院发布的《中国数字经济发展研究报告(2023 年)》，2022 年，我国数字经济规模达到 50.2 万亿元，与第二产业在国民经济中的比重基本相当，数字经济的质量也得到了显著提升(中国信息通信研究院，2023)。

随着数字经济的迅猛发展，大数据已成为学术界、产业界及政府关注的焦点。中共中央、国务院于 2020 年 4 月发布的《关于构建更加完善的要素市场化配置体制机制的意见》明确将数据纳入“五大生产要素”，标志着大数据已被提升至国家战略的高度。《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》进一步强调，应将大数据作为关键产业进行发展，并着重推动在数据采集、处理、存储、挖掘、分析及可视化处理等领域的技术创新，以培育一个全面覆盖大数据生命周期的产业体系。2022 年 1 月，国务院

发布的《“十四五”数字经济发展规划》提出了在 2025 年初步建立数据要素市场体系，并在 2035 年“力争形成一个统一公平、竞争有序、成熟完备的数字经济现代市场体系”(中共中央、国务院，2022)。这些举措对于推动我国数字经济的持续、健康、快速发展具有重要意义。党的二十大报告也明确指出，要加快发展数字经济，促进数字经济和实体经济深度融合，打造具有国际竞争力的数字产业集群。大数据在多个方面影响着人类社会行为，并改变着人们的思维方式。它为社会经济活动中的诸多方面，如社交关系和经济发展提供了更为直观的呈现方式，并促使传统管理决策逐渐向基于数据的决策转变，为管理学研究提供了新的工具和视角(洪永淼和汪寿阳，2021)。

当前的管理研究方法主要分为三大类。第一类是定性研究方法，包括综述研究、访谈研究、扎根理论、案例研究等方法。这些方法已被周小豪和朱晓林(2021)归纳。第二类是数理模型研究方法，如运筹学和博弈论，由张钹和张铃(1990)所代表。第三类是定量研究方法，以计量经济学及方法为代表，纪园园等(2021)对此进行了阐述。部分学者已经对不同因素相互作用下的复杂“组态效应”进行了探索，并在一定程度上阐明了这些因素组合对结果变量的复杂作用机制(杜运周和贾良定，2017)。随着数字化技术的进步，社会体系的复杂性前所未有地增加，引发了一系列挑战性问题。信息技术的迅猛发展导致现实与虚拟世界的不断融合，人与机器的连接程度日益提高，形成了一个“信息—物理—社会”耦合度日益增强的复杂系统，促使社会治理方式发生了深刻的变革(王芳和郭雷，2022)。在数字经济的背景下，各主体间的作用关系变得更加复杂，存在相互依赖性。描述研究对象的数据呈现出高维度化、多特征化的特点，数据类型多样，包括数值型数据、文本型数据、图片型数据、媒体数据等，甚至包括具有时间维度的高维数据。这些都给管理学研究带来了挑战。传统的管理研究范式越来越难以解释系统内复杂因素间的交互影响机制。因此，有必要对现有方法进行创新，以便更好地探究复杂系统问题。

大数据不断发展和人工智能崛起，管理决策在中国特色复杂情境中的应用日益增多(陈国青等，2021)。人们对复杂系统的理论和应用有了更深入的理解，这促进了计算法学的兴起(申卫星和刘云，2020)，并使得复杂系统管理成为一种新的研究范式(盛昭瀚和于景元，2021)。例如，一些学者(Varian，2014；洪永淼和汪寿阳，2021)分析了大数据和机器学习为经济学研究范式和方法带来的机遇

与挑战，并提出大数据技术有助于促进不同学科之间的融合。其他学者则运用自然语言处理(Kang et al.，2020)、神经网络(胡海青等，2012)等机器学习技术来研究管理问题。他们通过运用爬虫技术和语义分析方法构建指标(胡楠等，2021)，利用文本分析技术挖掘公司年报中的信息(伊志宏等，2019)，以及运用机器学习模型进行决策支持和股票预测(王茹婷等，2022)。

数据挖掘在管理学领域的应用充分展示了大数据技术在管理实践应用中的关键作用。通过数据挖掘，研究者能够揭示更多变量之间的潜在规律，为管理学研究中的知识发现提供有力支持。然而，目前的研究往往采取孤立的视角，导致研究过程缺乏系统性。例如，研究者倾向于单独使用决策树(程平和晏露，2022)、贝叶斯网络(Florio et al.，2018)等机器学习算法来探讨管理问题，或者采用神经网络方法进行管理预测(胡海青等，2012)。在综合分析群体间的差异性，异质性群体内部的前因与结果变量之间的潜在决策规则，以及前因与结果之间的细粒度因果关系时，传统数据挖掘方法仍显得力不从心。

在大数据背景下，本文提出了一种名为数据驱动分析(DAC)的创新方法，旨在研究复杂因素的影响机制，即复杂系统影响因素研究的数据驱动分析方法(Data-driven Analysis methods for the study of influencing factors in Complex systems，DAC)。DAC 巧妙融合了定量研究与定性研究的优势，通过一系列数据挖掘任务，如数据校准、聚类分析、决策树分析和贝叶斯网络分析等，构建了一个研究复杂因素影响机制的框架。该方法旨在帮助企业管理者在特定情境下识别实现预期结果所需遵循的决策规则，从而有助于他们合理分配资源，有效达成管理目标。

1.1.2 重要意义

区别于实证研究范式，从大数据视角出发，DAC 可以深入挖掘不同特征变量与结果间的复杂关系结构，具有一定学术创新和实践意义。

DAC 顺应国家数字经济发展趋势，为针对由此引发的社会系统复杂性问题提供了新的研究方法和思路。大数据被誉为 21 世纪的“钻石矿”，已经成为我国发展的战略性资源，正在深刻改变人类的生产和生活方式。特别是在科学技术研究领域，基于大数据技术的科研手段和工具为研究者提供了实时监测、跟踪和分析海量数据背后行为规律和管理策略的便利。在数字经济时代背景下，

DAC有助于人们梳理社会系统各主体间的复杂关系，对于加快数字经济的发展，促进数字经济与实体经济的深度融合具有重大意义。

DAC 体现了多学科的融合，创新性地发展了传统的管理研究范式。它依托于系统论、信息论和控制论的原理，将系统的演变视为多种不同因素的综合作用。通过数据挖掘算法，DAC 能够识别系统内样本的异质性特征，并结合所关注的管理决策问题，提取不同样本空间内影响因素的决策规则，在此基础上，进一步分析复杂因素的影响机制。DAC 体现了管理学、统计学与计算机科学等学科的交叉融合，降低了知识获取的复杂性。它为传统管理研究范式带来了新的思考，以决策问题为导向，以数据挖掘算法为技术支撑，为挖掘海量数据背后的重要知识和管理启示提供了一个有效的“工具箱”。

DAC 为解决复杂管理问题提供了新的方法和思路。在大数据时代，它为经济、管理等学科中复杂因素影响机制的研究提供了新的研究路径。大数据时代的到来导致大规模、多变量、非结构化数据的不断涌现。一些传统的统计方法和实证分析方法已不再适用于复杂系统的研究分析。DAC 从数据挖掘的角度深入分析复杂变量之间的非线性、异质性和离散性等关系及其联动作用机制，不仅能够实现样本内数据的拟合，还能对样本外数据进行精准预测。它通过数据采集(也称为数据收集)、数据清洗、聚类分析、决策分析、影响机制和敏感度分析等流程，深入剖析复杂系统多因素间的影响机制，并通过可视化手段增强了理论模型的可解释性和预测性。应用 DAC 可以使研究更加科学化、严谨化和精细化，帮助人们在面对问题或挑战时作出更加科学的决策。

1.2 数据挖掘的典型应用

大数据处理构成了一个综合、复杂且多维度的系统，涵盖了众多处理模块。作为大数据处理体系中的一个独立分支，数据挖掘技术与其他模块相辅相成，共同进步。经过多年的发展，数据挖掘研究已经建立了一套坚实的理论基础，涵盖了分类、聚类、模式挖掘和规则提取等领域。Huber 等(2019)的评估表明，在数据量庞大时，数据驱动的方法相较于传统分析方法具有明显优势。杜鹃(2020)指出，大数据技术的显著优势在于能够从庞大的数据集中揭示事物间的关

联性，并对事物的未来发展进行预测，从而更精确地反映事物的全貌。数据挖掘技术在自然科学和社会科学领域都得到了广泛的应用。谭春辉和熊梦媛(2021)利用隐含狄利克雷分布(Latent Dirichlet Allocation，LDA)模型对国内外数据挖掘研究进行了对比分析，发现国内研究更倾向于社会科学应用，而国外则更偏向自然科学应用；数据挖掘领域的研究重心正逐渐从理论研究转向应用研究，并且与大数据技术结合，催生了许多新兴的发展方向。机器学习方法，如随机森林和神经网络，在工程、商业和科学领域研究中得到了广泛应用(Mjolsness & Decoste，2001)。接下来，研究者将从自然科学和社会科学两个领域探讨数据挖掘的应用。

1.2.1 自然科学领域的应用

数据挖掘在自然科学领域的应用广泛，涵盖了预测、模式发现、规律识别以及辅助决策等方面。在众多学科中，预测是一个备受关注的课题，旨在利用现有数据对未来进行预测。例如，Fernández 等(2020)利用神经网络构建了地铁能量耗散模型，成功预测了能量消耗；Li 等(2020)基于反向传播神经网络(Back Propagation Neural Network，BPNN)对吸能装置的碰撞响应进行了反演预测；赵然杭等(2021)提出了一种基于时序分解的神经网络模型，能够对降水时序数据进行有效挖掘和预测；陈冲等(2021)处理了战场气象环境的历史数据，并提出了一种基于历史数据挖掘的未来战场气象环境数据模糊预测算法。识别和挖掘飞行员操纵数据属于时间序列预测问题，构建模型时需要结合实际应用背景和时间序列预测方法。王志刚等(2021)提出了一种基于长短期记忆网络(Long Short-Time Memory，LSTM)模型的飞行历史数据挖掘模型构建方法，充分利用数据，提取飞行员的飞行智慧，以用于后续型号的研制。

数据挖掘同样在模式发现与规律识别方面得到广泛应用，涉及水利水电、灭火系统、中医药等多个领域。唐凤珍等(2020)为解决梯级电站效益评估中不同电站间经济效益相互关联的复杂性问题，提出了一种基于数据挖掘的电站效益关联分析方法，为多个电站效益评估方法的研究开辟了新途径。在消防领域，数据挖掘的应用主要包括火灾模型建立、分析和识别火灾发展规律等。韩光等(2020)利用数据挖掘技术和深层网络，建立了一个用于自动喷水灭火系统的模型。在中医药研究领域，数据挖掘尤其在用药规律研究和名老中医经验传承方

面发挥着重要作用(刘嘉辉等，2020)。例如，学者基于方剂信息的关联规则算法，减少了对中药复方的分析、整理工作，并在此基础上进一步挖掘出中药数据的客观规律；聚类算法以无监督方式挖掘高维中医数据中属性间的固有联系，探索诊断与用药的模式，提高中医诊疗的创新性(陈志奎等，2020)。

数据挖掘对于辅助决策也具有重要意义。医学数据具有多态性、信息缺失性、时序性、冗余性等特点，传统计算方式无法完全满足卫生医学数据的分析需求。采用分类算法，可以根据患者实际情况，模拟医生的诊断方式，进行客观分析，辅助医生作出合理判断(陈志奎等，2020)。

此外，数据挖掘在自然科学中的应用还包括优化、地理信息挖掘、图像识别、故障诊断等方面。数据挖掘技术能够有效解决交通道路和物流网络间的优化问题。随着我国城市化进程的加快和人民生活水平的提高，汽车数量不断增加，导致严重交通堵塞等问题的出现。Gumus 和 Yiltas-Kaplan 等(2020)基于人工神经网络实现了拥堵问题的优化。信息科学及复杂系统领域的许多学者已经针对地理大数据分析和挖掘开展了大量研究(刘耀林等，2022)。地理信息挖掘在智慧城市(姚晓婧等，2019)、公共安全(Liu et al.，2018)、环境保护(Shan et al.，2014)、气候变化(Liess et al.，2017)、流行病防控(Huang et al.，2021)、矿产资源勘查(Guan et al.，2021)等领域均发挥了重要作用。在计算机领域，将数据挖掘技术用于影像识别已较为普遍，如人脸辨识、指纹辨识等。时庆涛等(2020)将数据挖掘技术应用于多光谱图像特征数据处理，提出了一种基于 Contourlet 变换的图像纹理特征挖掘方法。该方法具有高挖掘精度、短挖掘时间、低成本消耗，并能获得均匀度较好、深浅度适中的数据的优点。数据挖掘还被用来检测异常数据，排除故障。例如，朱圳等(2022)提出了一种对通信网络故障进行分类的数据挖掘方法。

1.2.2 社会科学领域的应用

通过数据挖掘技术，研究者可以将信贷数据转化为分类规则，揭示其中与金融政策相悖的信息，为政府的金融干预政策提供依据。李海林等(2022)运用分析与回归树(Classification And Regression Trees，CART)算法，探索了影响杰出学者达成高绩效目标的关键特征因素，并挖掘了规则路径中的非线性复杂关系结构，为实施精心设计的绩效激励措施奠定了基础。

数据挖掘技术能够揭示变量间的因果关系，常见的方法包括贝叶斯网络等。高晶鑫等(2015)构建了一个基于贝叶斯网络的居民出行目的地选择模型，并对模型中的父节点与子节点进行了概率相关性分析，从而揭示了居民出行目的地选择的规律及其影响因素特征。刘建荣和刘志伟(2022)通过贝叶斯网络分析了各种因素对老年人使用公共交通的总体满意度和意愿的影响程度。贝叶斯网络使得设计者能够考虑用户的行为特征和经验需求，并识别出影响用户行为的真实因素。胡康等(2023)利用贝叶斯网络技术深入分析消费者需求和行为，并对快递包装回收进行了系统化设计。李海林等(2023)建立了贝叶斯网络，通过预测分析、原因诊断和贡献率测度等方法，探究了科技补贴和人才补贴与企业创新活动之间的因果作用机制。

数据挖掘技术为政府、平台和企业提供了科学、高效的辅助决策支持。通过大数据技术，研究人员可以迅速获取有价值的信息，及时发现并纠正问题，为社会和政府提供有效的决策参考(顾肃，2021)。物联网和大数据技术在养老行业中有广泛的应用前景。屈芳等(2017)提出了一种“互联网+大数据”的养老方式。该方式基于多源异质信息的汇集与数据融合挖掘，结合通信技术、数据挖掘技术和人工智能技术，利用传感器和智能计算对老年人信息进行实时分析，并提供智能化辅助决策。邱国栋(2018)提出了一种以数据为中心，以算法为手段，以平台为支撑的“数据—智慧”决策模型，为政府指挥决策和社科研究提供了新的理论视角。随着大数据技术应用的普及，数据驱动的决策优化已成为企业科学管理的发展方向(陈国青等，2020)。如何结合企业领域知识和合理运用数据以动态优化企业决策，进而提升企业竞争力并改善消费者体验，是企业运营管理和数智技术发展领域的重要研究问题，也反映为数字经济健康长期发展所需要的企业基础能力之一(Mochon et al.，2017；Zhang & Wedel，2009)。张诚等(2023)融合运营管理和营销领域知识，提出一个基于深度增强学习的动态促销框架。该框架结合仿真技术和机器学习技术，实现预测与决策分析的协同，为大数据时代的协同分析研究提供了重要参考和思路。王霄(2019)引入数据驱动相关方法研究舱位分配的优化问题，为航空公司作出合理舱位控制决策提供了借鉴。熊浩和鄢慧丽(2022)基于数据驱动与运筹优化结合的视角研究外卖平台派单问题，发现将机器学习、运筹优化和仿真分析相结合的分析方法能帮助外卖平台优化智能派单决策，提高运行效率。

数据挖掘在社会科学领域的应用还包括预测、规律识别、个性化推荐和文本挖掘等。王欣和张冬梅(2018)在收集读者阅读数据，以及挖掘和预测个性化阅读需求的基础上，提出了一种推荐机制，实现了面向读者的个性化、智能化阅读服务。王颖纯等(2018)提出了以知识挖掘为基础的智能推荐服务，包括基于“用户画像”的智能推荐和面向用户需求的智能推荐等。近年来，从大量语言文本中挖掘人类认知模式的研究受到了广泛关注(DeDeo，2022)。Box-Steffensmeier 和 Moses(2021)通过社交媒体信息中的认知偏见和语气来衡量特定群体的认知表达模式，探究其在传染病流行期间对信息传播和公众反应的介导作用。Carrasco-Farré(2022)以在线新闻为研究对象，探究了不同类别错误信息的语法及词汇特征、情感极性和社会认同，进而衡量其中的情绪和道德内容。

数据挖掘在自然科学和社会科学研究中的应用均体现了大数据技术的重要性。数据挖掘有助于研究者明确更多变量间的潜在规律，为研究中的知识发现提供参考。然而，现有研究大多从独立视角展开，研究过程较为零散，如单独运用决策树(程平和晏露，2022)、贝叶斯网络(Florio 等，2018)等机器学习算法探究问题，或运用神经网络方法进行预测(胡海青等，2012)。传统数据挖掘方法在综合剖析群体间的差异性，异质性群体内前因与结果变量间潜在决策规则，前因与结果间细粒度因果关系时仍存在不足。因此，需要提出一种整体研究框架，以系统性地解决复杂性问题。

1.3 基本框架与流程

复杂系统影响因素研究的数据驱动分析方法主要包括 5 个关键阶段(步骤)：数据获取、数据处理与变量测量、聚类分析、决策树分析和贝叶斯网络分析。在数据获取阶段，从各种来源收集相关原始数据；在数据处理与变量测量阶段，对原始数据进行清洗、转化和规范化；在聚类分析阶段，识别数据中的潜在模式和分组；在决策树分析阶段，通过决策树分析探索变量间的层次关系和预测规则；最后在贝叶斯网络分析阶段，基于贝叶斯网络分析揭示变量间的概率依赖关系。通过这一系统化的流程，研究者可以有效地从海量数据中提取有价值的信息，揭示复杂系统的内在机制和关键驱动因素，为后续的决策和优化提供支持。

1.3.1 基本框架

DAC 研究框架如图 1.1 所示，大致分为数据采集与预处理、指标量化与校准、异质性群组划分、决策规则提取、因果关系识别 5 个重要模块。

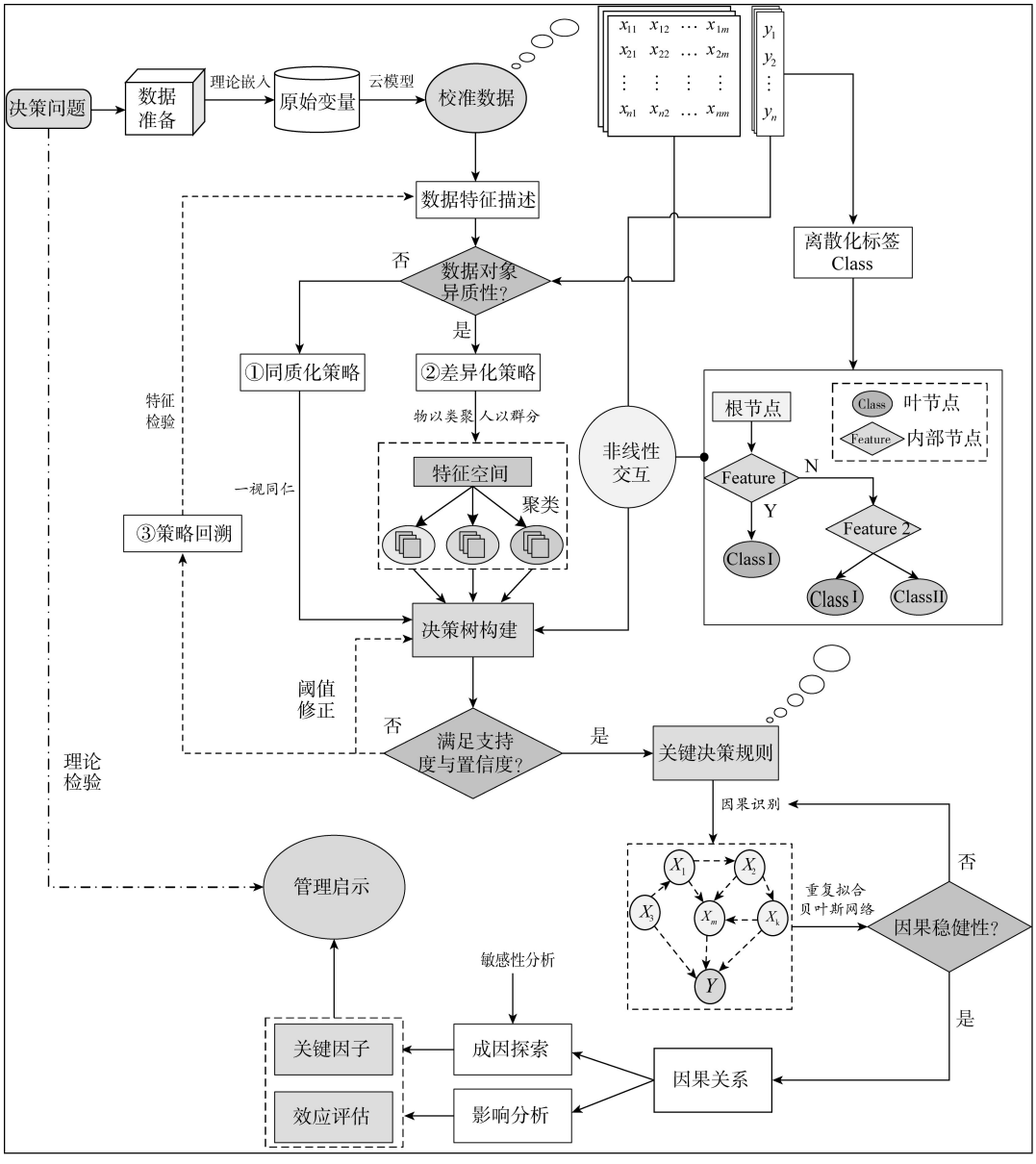

图 1.1 DAC 研究框架

在数据采集与预处理模块，首先根据研究问题，依据指标选取原则，选定相应的前因和结果变量，并采用多种方法从多个来源采集数据，以形成原始数

据集。其次，对数据执行异常检测与处理，确保数据的干净与完整性。

在指标量化与校准模块，对于无法直接观测或采集数据的变量，根据其数据类型，选择适当的指标量化方法进行度量，从而获得原始变量数据。采用云模型方法，消除量纲和极值的影响，对原始变量数据进行校准，得到取值分布在 0～1 的前因和结果变量数据值。

在异质性群组划分模块，对校准后的数据进行异质性分析。如果发现不同样本前因间存在显著的异质性，则利用聚类算法对样本空间进行划分，设置相关参数，将样本划分为不同的群体，并对这些群体进行特征分析和可视化。在决策规则提取模块，将结果变量数据纳入研究，针对不同的群体分别进行决策树分析。如果不同样本前因间不存在显著的异质性，则直接对总体样本进行决策树分析。设置决策树的相关参数和剪枝策略，构建决策树模型。如果生成的决策树分枝的支持度和置信度均不高，则通过调整相关参数和剪枝策略来优化决策树，直至满足要求。最终，通过提取决策树中感兴趣的决策规则，进行目标因变量的非线性复杂影响因素交互效应分析。

在因果关系识别模块，针对上述感兴趣的决策规则所对应的研究对象进行深入分析，运用爬山算法识别变量间的相互依赖关系，并构建贝叶斯网络模型。通过敏感度分析，细致探究变量间的影响关系，揭示复杂前因变量对结果变量的影响路径。同时，结合相关管理理论，分析决策规则和影响路径，得出能够指导管理实践的研究结论。

1.3.2 基本流程

DAC 基本操作流程如图 1.2 所示。从图 1.2 中可以得知，DAC 大致包括数据获取、数据处理与变量测量、聚类分析、决策树分析、贝叶斯网络分析 5 个关键阶段(步骤)。

1. 数据获取

当前数据呈现的形式多种多样，如调查问卷数据、实验仿真数据、文本类型数据、网络类型数据等。为了确定研究所需的数据形式和来源，必须仔细分析问题。对于大规模数据集，可以利用 Python 网络爬虫等技术进行采集，或者直接从专业数据库中导出。

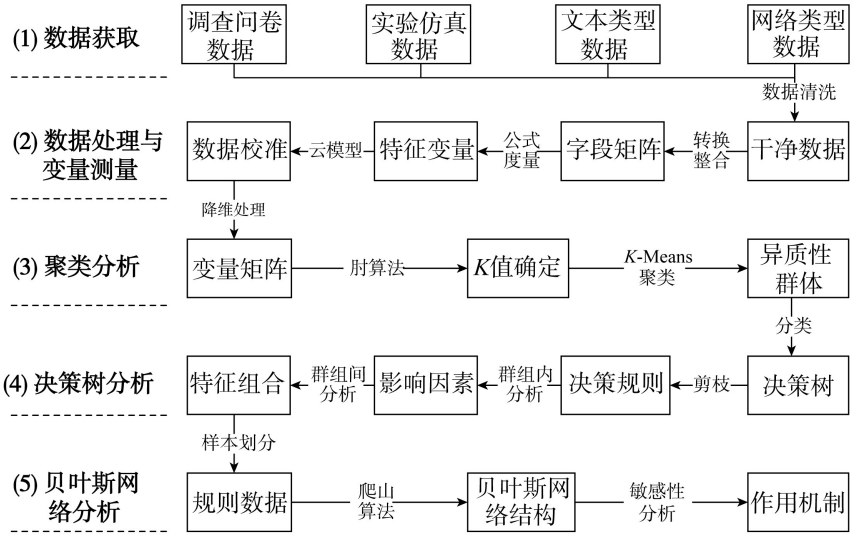

图 1.2　DAC 基本操作流程

2. 数据处理与变量测量

由于主观和客观因素的影响，收集到的原始数据通常包含错误，可能存在如重复记录、格式不统一或数据缺失等问题。为了确保后续研究的精确性，必须对这些原始数据进行清洗和识别，同时实现数据的整合与存储。数据处理流程主要包括统一数据格式，处理无效和缺失值，异常值处理，变量标准化，以及数据的离散化和连续化等步骤。在获取原始数据后，需要进行指标量化，将无法直接观测的指标转化为可操作的指标形式，为深入研究提供坚实的基础。针对不同数据类型，可以采用相应的指标量化方法。例如，对于文本数据，可以使用词频统计、词频-逆文档频率(Term Frequency-Inverse Document Frequency，TF-IDF)、潜在语义分析(Latent Semantic Analysis，LSA)等技术进行度量；而对于网络数据，可以采用社会网络分析法、基于图卷积神经网络的网络嵌入技术等进行评估。在数据访问控制(Data Access Control，DAC)研究中，研究者提出了使用云模型(李德毅，2000)对变量进行数据校准的方法。该方法消除了变量间的量纲和极值影响，将连续数值转换为[0, 1]区间的数值，便于运用数据挖掘技术对复杂因素进行比较和机制研究。

3. 聚类分析

运用单一模型对全部样本数据进行分析往往会导致模型的拟合度下降，并且研究得出的结论可能缺乏普遍适用性。通过在特征变量中识别数据对象的相

似性，并据此将数据划分为不同的群体，然后对每个群体应用不同的模型进行分析，不仅可提升模型的拟合度，还增强了分析结果的针对性。聚类算法基于“物以类聚”的原则，将具有相似特征的样本归入同一簇中，确保簇内数据的相似度(Intracluster Similarity)较高，而不同簇间数据的相似度较低。一些常见的聚类算法包括 K-均值聚类算法(K-Means clustering algorithm，K-Means)(Celebi et al.，2013)、近邻传播聚类算法(Affinity Propagation clustering algorithm，AP)(Frey & Dueck，2007)以及基于密度的聚类(Density-Based Spatial Clustering of Applications with Noise，DBSCAN) (Schubert et al.，2017)等。DAC 能够利用聚类等技术手段，将样本数据细分为多个群体，以便后续对不同群体对象进行不同模型的分析，做到“具体问题具体分析”。

4. 决策树分析

为了深入理解在不同情境中自变量的条件属性如何与因变量产生非线性效应，DAC 运用决策树模型对异质性数据集进行细致的决策规则分析。在决策树模型中，节点的分裂规则是基于划分后子空间数据的不纯度来确定的，不纯度越低，意味着分裂规则越有效。通过易于理解的树状结构，研究者可以记录并总结变量之间的映射关系，从而识别哪些特征或特征组合对因变量的决策分类具有最大的影响。此外，这种分析还能揭示异质性群体内部变量间的影响机制，并协助决策者根据已知的变量特征进行属性预测。通过应用 ID3(Iterative Dichotomiser 3)(Quinlan，1986)、C4.5(Quinlan，1993)以及 CART(Denison et al.，1998)等算法，研究者可以进一步分析特征变量组合与结果变量之间的交互作用，明确前因与结果之间的内在联系，并对未知样本数据进行有效预测。

5. 贝叶斯网络分析

DAC 在处理异质性数据群体时，不仅仅局限于单一的前因变量，而是综合考虑了不同前因组合对因变量的交互影响程度。这种方法的核心在于以决策规则中不同特征组合数据为中心，深入挖掘数据之间的内在联系。为了实现这一目标，DAC 采用了爬山(Hill Climbing)算法(Tsamardinos et al.，2006)。这是一种经典的启发式搜索算法，能够有效地识别变量间的逻辑关系。在识别出变量间的逻辑关系后，DAC 进一步构建了贝叶斯网络(Pearl，1986)。这是一种基于概率图模型的因果推断方法。贝叶斯网络能够以图形化的方式表示变量间的条件依赖关系，从而揭示因果关系的复杂结构。

在构建了贝叶斯网络之后，DAC 并没有停止，而是继续深入分析系统内各影响因素间及其与因变量间的因果关系。通过对这些因果关系的识别和判断，DAC 能够对各前因变量的条件概率进行准确估计。条件概率的估计是理解变量间因果关系的关键，因为它能够反映出在给定某些前因变量的条件下，因变量发生的概率变化。最后，DAC 通过条件概率的变化程度进行敏感度分析，以评估不同前因变量对因变量的影响程度。敏感度分析是一种评估模型对参数变化敏感性的方法。通过这种方法，DAC 能够识别出对因变量影响最大的关键前因变量，从而为决策提供更为科学的依据。

1.4 相关软件及工具准备

进行 DAC 研究，一般采用 Python、PyCharm、Graphviz、Netica 等软件工具。图 1.3 展示了各个流程环节所涉及的软件应用，其中虚线箭头指示了在特定分析步骤中使用到的软件。

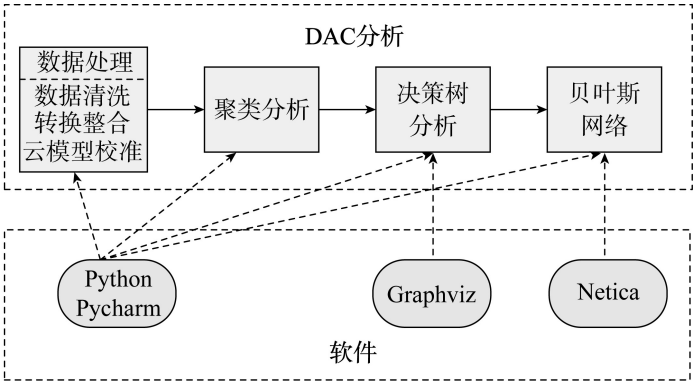

图 1.3　DAC 研究中使用的软件

1.4.1　Python 软件

1. 软件介绍

在 DAC 分析中，校准、聚类、决策树、贝叶斯等关键环节均会运用到 Python 软件。该软件在 DAC 分析中扮演着贯穿始终的角色，而且所附的运行代码也是基于程序语言 Python 编写的。Python 是一种解释型的面向对象编程语言，以简

洁的语法和丰富的库资源著称，在人工智能分析领域尤为便捷。例如，sklearn 库支持机器学习任务，numpy 和 scipy 库适用于数值计算，PyBrain 库便于神经网络研究，而 matplotlib 库则助力于数据可视化。

Python 的功能多样。首先，它在机器学习领域中大有可为。在人工智能、机器人技术、图像识别、自然语言处理等研究领域，Python 都扮演着关键角色。其次，Python 可用于编写“爬虫”程序。所谓“爬虫”，是指根据特定规则编写代码，通过自动化工具有针对性地收集和处理数据，以获取所需信息。再次，Python 适用于网站开发。它提供了丰富的免费数据处理库和网站模板系统，以及与网站服务器交互的库，使得网站开发变得简单高效。最后，Python 还广泛应用于嵌入式软件开发、游戏开发等其他领域。

2. 下载安装与配置环境

Python 可在官网下载页面 https://www.python.org/downloads/下载，用户可选择适合自己计算机系统版本的安装包下载即可，具体步骤如下。

(1) 打开 Python 下载页面https://www.python.org/downloads/(见图 1.4)，可看到 Window、Linux/UNX、macOS 等各个平台安装包的下载地址，选择与计算机系统相匹配的安装包下载。这里以 Windows 操作平台为例。

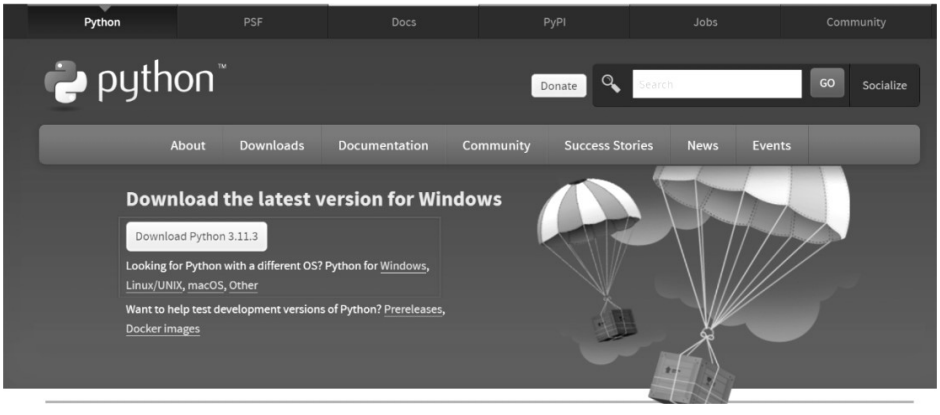

图 1.4 Python 下载页面

(2) 在上个页面往下滑，进入选择下载界面。如图 1.5 所示，选择所需要的 Python 版本号，单击“Download”。这里选择 Python3.8.0 安装包下载，也可选择其他更新的版本。

(3) 因需用到 Windows 下的解释器，所以在运行系统中可选择对应的 Windows 版本，有 64 位与 32 位两种选择。executable 指“可执行”，该版本需

图 1.5　Python 下载界面

在安装后使用，安装过程较容易，一直单击选择默认即可。embeddable 指“嵌入”，该版本解压后可使用。这里选择的是画红线框中这个 64 位的版本。单击“Windows x86-64 executable installer”下载安装包，如图 1.6 所示。

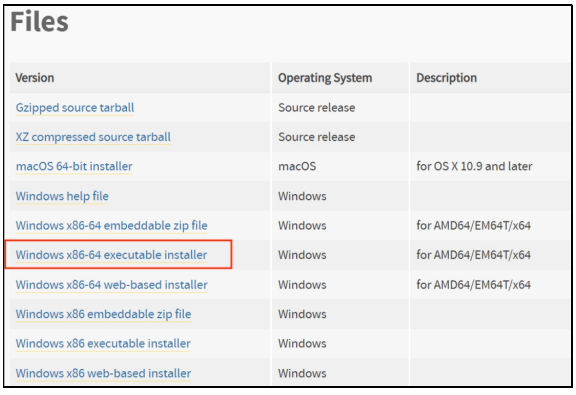

图 1.6　选择下载安装包

(4) 下载后找到安装包，双击打开，单击“运行”按钮(见图 1.7)，进入 Python 安装向导界面。

图 1.7　单击“运行”按钮

(5) 在安装界面中，勾选左下角“Add Python 3.8 to PATH”复选框，然后选择第二个自定义安装(Customize installation)，如图 1.8 所示。

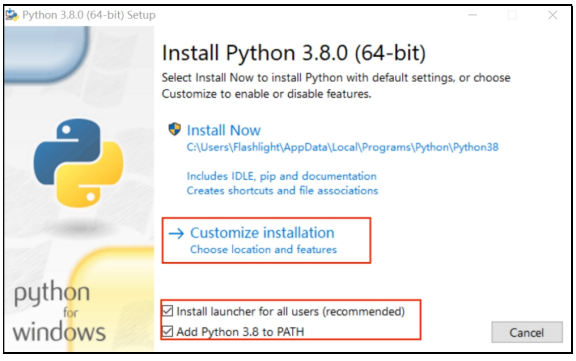

图 1.8　选择自定义安装

(6) 选择默认选项即可，单击“Next”按钮，进入安装步骤的下一步，如图 1.9 所示。

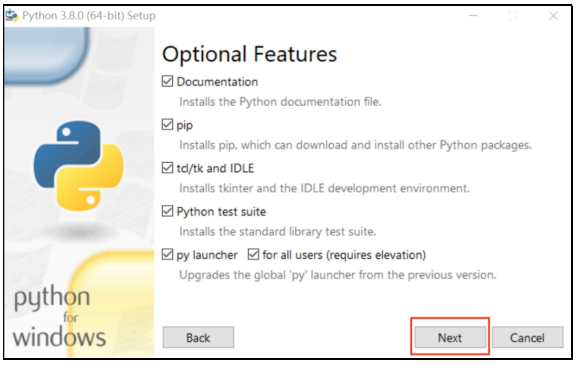

图 1.9　单击“Next”按钮

(7) 选择自定义安装路径，单击“Install”按钮，进行安装，如图 1.10 所示。

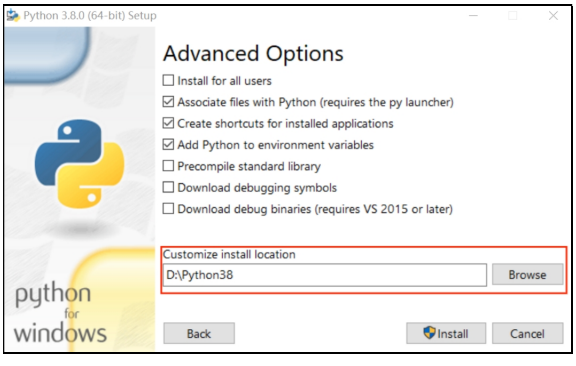

图 1.10　单击“Install”按钮

(8) 等待进度条安装完毕，单击“Close”按钮退出，如图 1.11 所示。

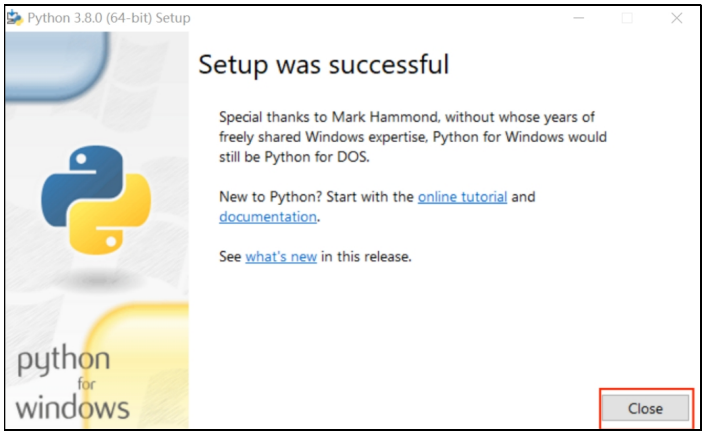

图 1.11　单击“Close”按钮退出

(9) 检查是否成功安装 Python，须按 Win+R 组合键，在弹出的运行框中输入“cmd”，如图 1.12 所示。

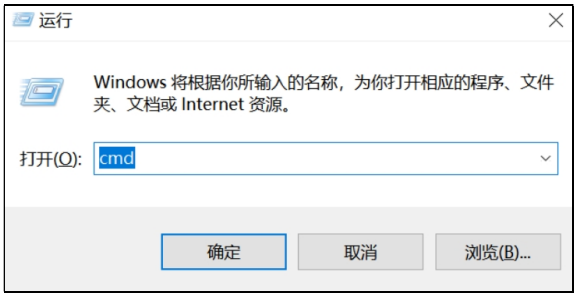

图 1.12　输入“cmd”命令

在弹出来的 cmd 框中输入“python”，若显示出 Python 的版本信息，就为安装成功，如图 1.13 所示。

图 1.13　显示 Python 版本信息

1.4.2 PyCharm 软件

1. 软件介绍

PyCharm 是一款功能强大的 Python 集成开发环境(Integrated Development Environment，IDE)配置软件。它为我们提供了便捷的代码编辑和 Python 调试功能，是 DAC 中不可或缺的工具。除了 PyCharm，Python IDE 的家族还包括 IDLE、Anaconda、Jupyter Notebook 和 Spyder 等成员。本书选择 PyCharm 作为主要的讲解对象。PyCharm 配备了一系列实用工具，包括调试器、语法高亮显示、项目管理以及代码导航等。这些工具极大地提升了开发者的工作效率。此外，在 Django 框架的支持下，PyCharm 还为专业级网站开发人员提供了众多高级功能。

2. 下载安装与配置环境

在官网 https://www.jetbrains.com/pycharm/download/下载 PyCharm 安装包，注意要和 Python 版本相匹配，安装完成后须配置环境。为了后续研究，须在 PyCharm 软件中安装 sklearn、pandas、matplotlib 等工具库。

(1) 打开 PyCharm 下载网址 https://www.jetbrains.com/pycharm/download/，会看到如下页面(图 1.14)。可看到，网站提供了适合 Windows、macOS、Linux 操作系统的 PyCharm 安装包。这里介绍 PyCharm 在 Windows 下的安装。页面中，“Professional”表示专业版，“Community”是社区版，推荐安装社区版，因为是免费使用的。

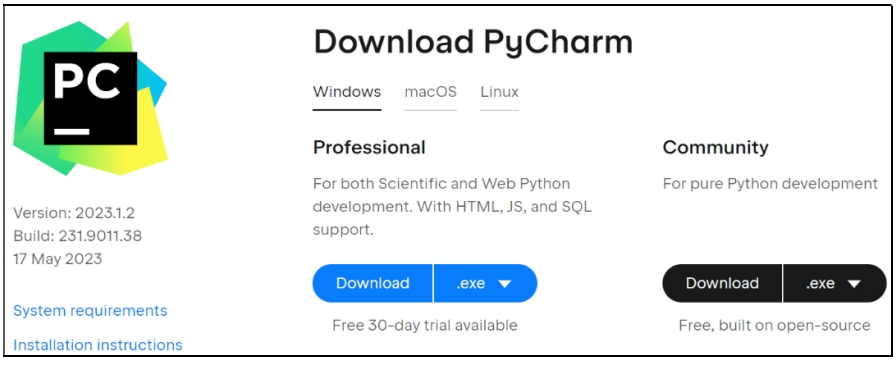

图 1.14 PyCharm 官方下载页面

这里以“pycharm-community-2022.3.3”安装包为例介绍下载步骤，单击“2022.3.3-Windows(exe)”(图 1.15)，也可选择适合自己计算机配置的安装包下载。

图 1.15　单击“2022.3.3-Windows(exe)”

(2) 下载后找到安装包，双击打开，单击“运行”，进入 PyCharm 安装向导(图 1.16)。

图 1.16　运行安装包

(3) 修改安装路径，这里放的是 D 盘，修改好以后，单击“Next”按钮(图 1.17)。

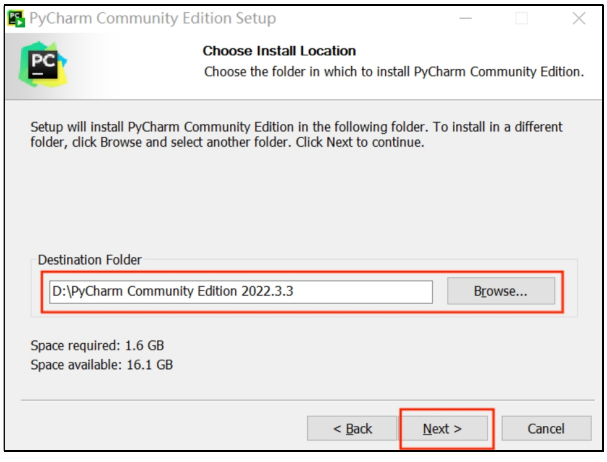

图 1.17　修改安装路径为 D 盘

(4) 勾选相关安装选项复选框，单击“Next”按钮，如图 1.18 所示。

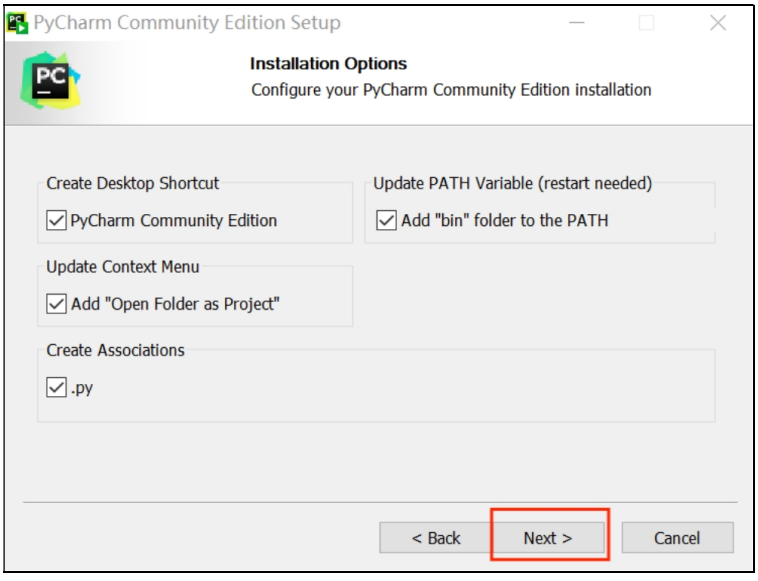

图 1.18　勾选相关安装选项复选框

(5) 选择“开始菜单”目录，以在该目录中建立程序快捷方式，也可建一个新文件夹。单击“Install”按钮，等待安装，如图 1.19 所示。

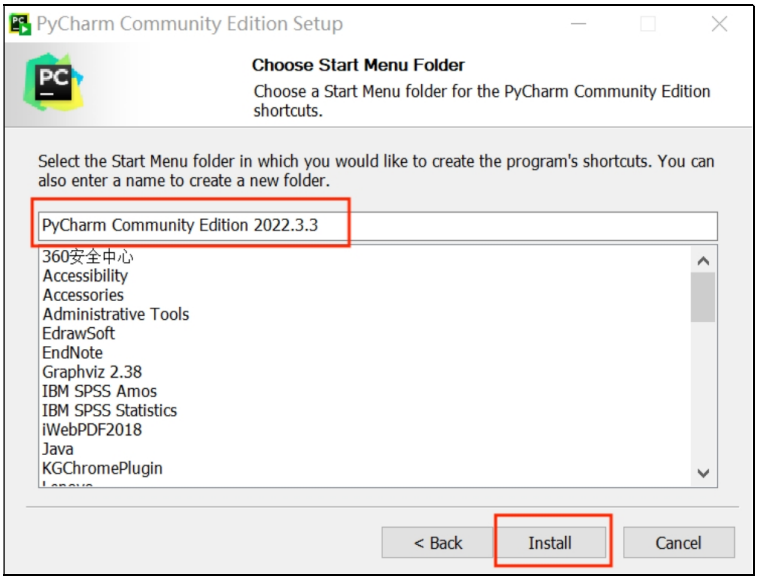

图 1.19　选择目录等待安装

(6) 必须重新启动计算机，才能完成 PyCharm 社区版安装。可选择立即重新启动或稍后手动重启，如图 1.20 所示。

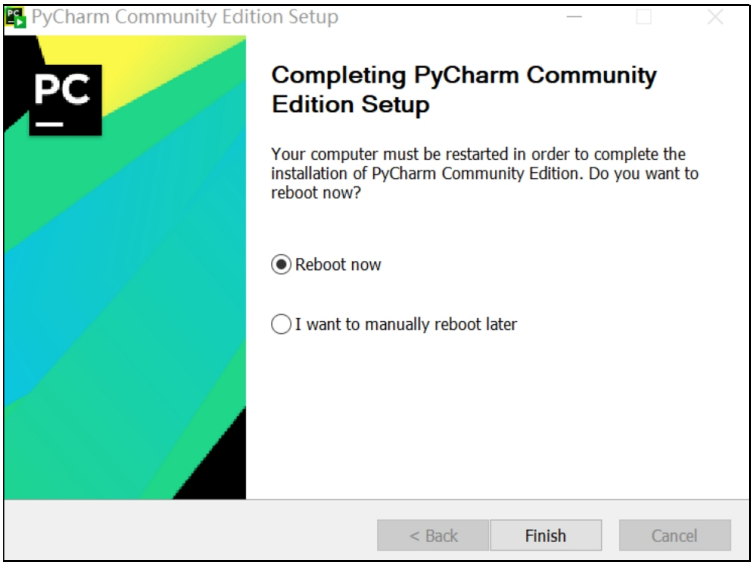

图 1.20　选择重启方式

(7) PyCharm 装好后，双击图标进入该软件，单击“New Project”(图 1.21)。

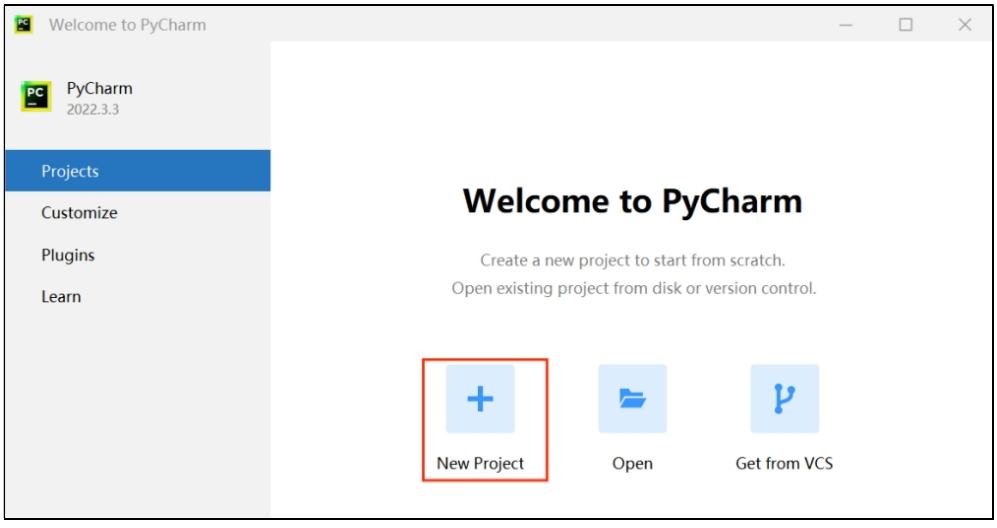

图 1.21　单击 New Project

(8) 接下来配置环境，包括设置路径和选择 PyCharm 对应的 Python 解释器。“Location”是存放工程的路径，在对话框中单击第一个，可以选择“Location”的路径，所选择的路径需要为空，不然无法创建。第二个“Location”是系统默认的，不用修改。在 Base interpreter 中，选择 PyCharm 对应的 Python 解释器。这里选择已下载好的 Python3.8.0 版本的解释器。单击“Create”按钮即可设置，如图 1.22 所示。

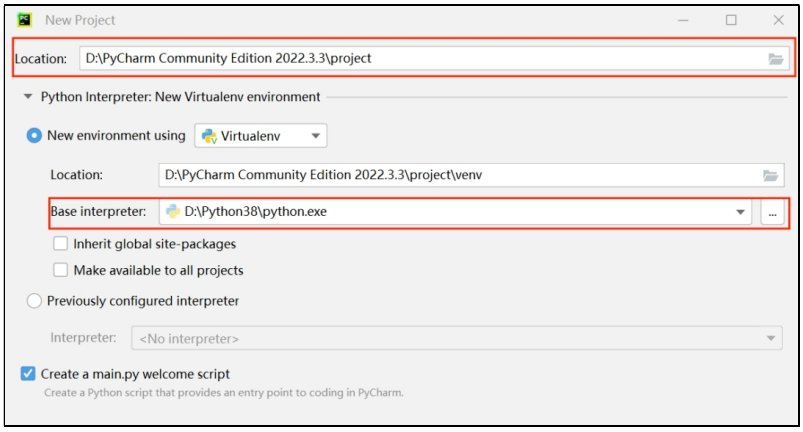

图 1.22 配置环境

(9) 创建一个编译环境。单击“File”，然后单击“New”(图 1.23)，在弹出的列表中选择“Python File”(图 1.24)，将文件命名为“hello”(图 1.25)，即可创建新文件。可在该文件中编写代码，运行程序。

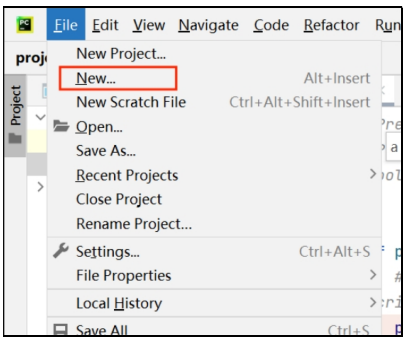

图 1.23 单击“File”

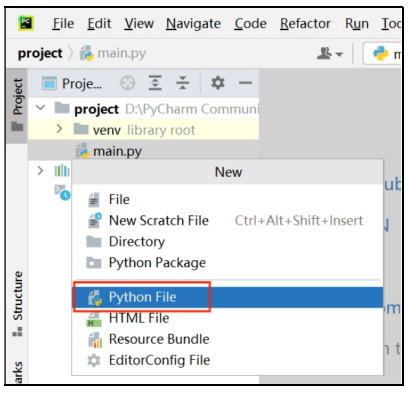

图 1.24 选择“Python File”

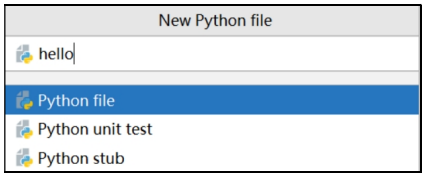

图 1.25 将文件命名为“hello”

(10) 下载工具库。在 DAC 研究中需用到 sklearn、pandas、matplotlib 等工具库。以 sklearn 工具库为例。单击“File”→“ Settings”→“ Python Interpreter”，然后单击右侧的“+”，在“搜索”栏里键入“sklearn”，然后单击“Install Package”，即可安装 sklearn 工具库。其他工具库的安装方式与之相同，如图 1.26～图 1.28 所示。

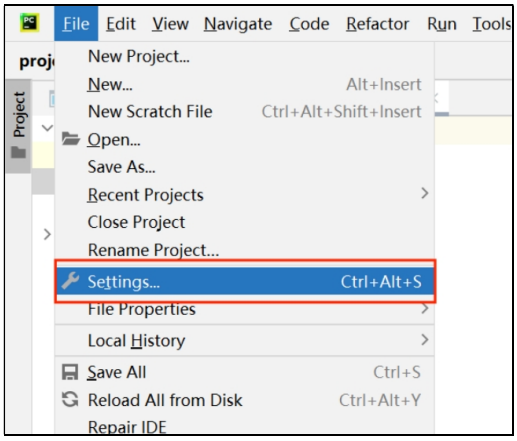

图 1.26　下载工具库 1

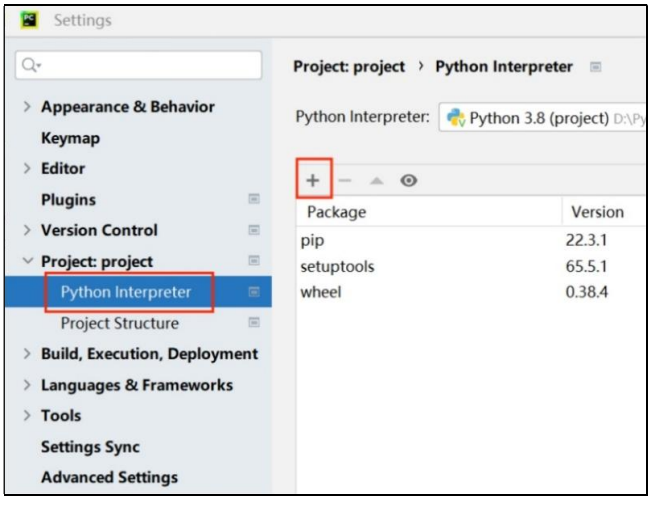

图 1.27　下载工具库 2

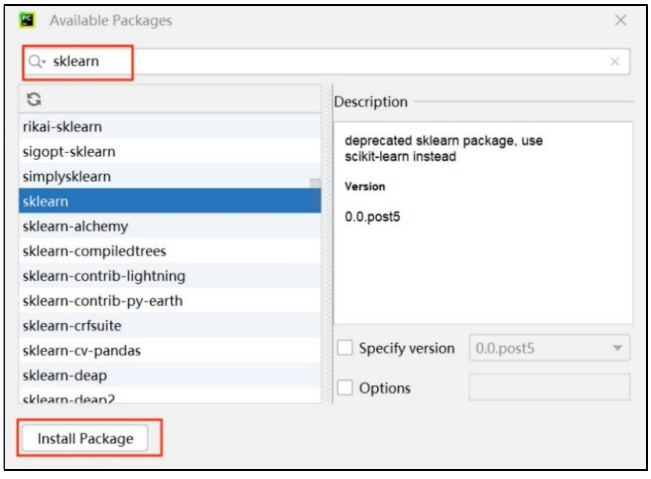

图 1.28　下载工具库 3

以上就是 PyCharm 软件的下载和安装、环境配置以及工具库下载的过程。

1.4.3 Graphviz 软件

1. 软件介绍

Graphviz 是一个图形可视化软件，被用于本书中绘制决策树图。作为一款开源的图形可视化软件，Graphviz 采用 DOT 语言作为脚本，通过其内置的布局引擎对脚本进行解析，从而实现图形的自动化布局。在使用 Graphviz 进行图形绘制时，用户需要编写 dot 脚本来描述各个节点之间的关系，而无需手动处理布局问题。此外，Graphviz 还允许用户自定义图形元素的详细属性，包括节点的字体、颜色以及线条样式等。它支持多种输出格式，如常见的图片格式、SVG 以及 PDF 等。Graphviz 还提供了多种布局选项，例如，dot(用于有向图)、neato(用于无向图)、circo(用于圆环布局)等。

2. 下载安装与配置环境

Graphviz 安装时，要先在官网 http://www.graphviz.org/download/下载安装包，按前述方法，安装完成后配置环境变量，打开 cmd，输入命令“dot-version”，若显示版本信息，说明安装成功。用 Graphviz 画决策树图，还需要在 Python 环境(PyCharm)中安装 Graphviz 工具库，画图时需调用该工具库。

(1) 进入下载地址：http://www.graphviz.org/download/，里面有适合 Windows、Mac、Linux 等系统的软件安装包，这里以 Windows 版本的为例。选择适合自己计算机操作系统(64 位或 32 位)的版本，这里以“graphviz-2.49.0(64-bit) EXE installer[sha256]”安装包为例，进行下载(图 1.29)。

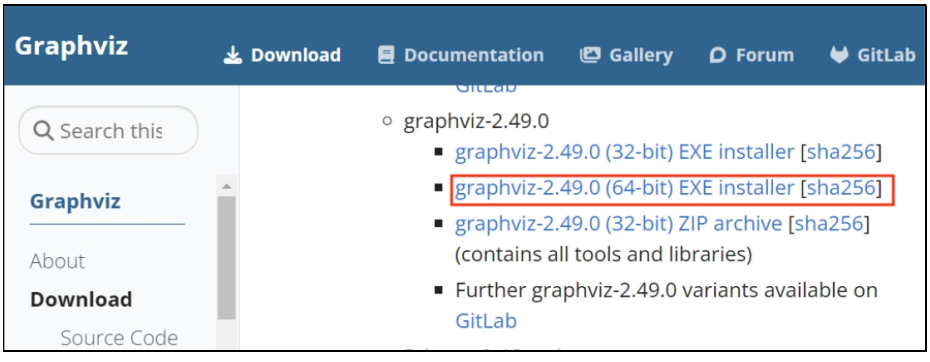

图 1.29 下载安装包

(2) 下载后，双击安装包，单击“下一步”。在“Graphviz 安装”对话框中，

选择“Add Graphviz to the system PATH for all users”(将 Graphviz 添加到所有用户的系统路径中)，单击“下一步”，如图 1.30 所示。

图 1.30　选择软件安装包

(3) 选择安装路径，单击“下一步”(图 1.31)。

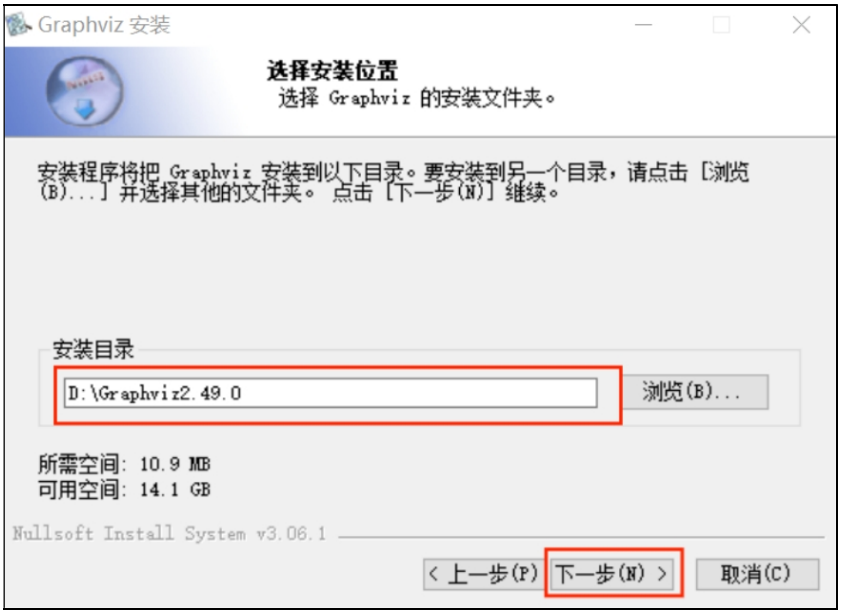

图 1.31　选择安装路径

(4) 选择开始菜单文件夹用于创建快捷方式，单击“安装”(图 1.32)。

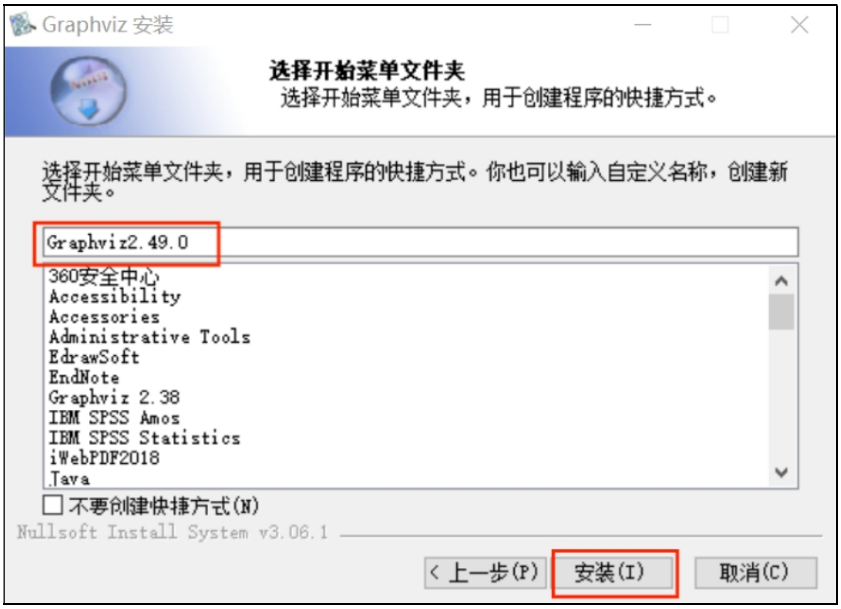

图 1.32　选择开始菜单文件夹

(5) 确认环境变量是正确配置：鼠标右键单击“此电脑”→“属性”→“高级系统设置”→“环境变量”(图 1.33—图 1.35)选择“系统变量”里的“Path”，单击“编辑”(图 1.36)。可以看到，已经在系统变量“Path”中添加了“Graphviz2.49.0”的 bin 文件夹路径(图 1.37)，若没有则需要手动添加路径。

图 1.33　单击“属性”

图 1.34　单击“高级系统设置”

图 1.35　单击“环境变量”

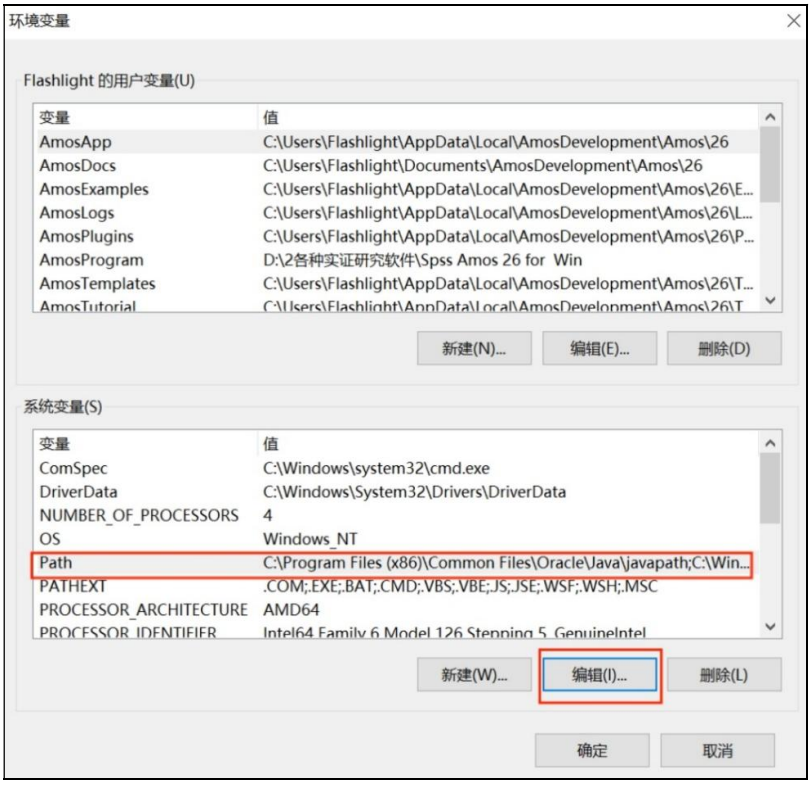

图 1.36　选择“Path”

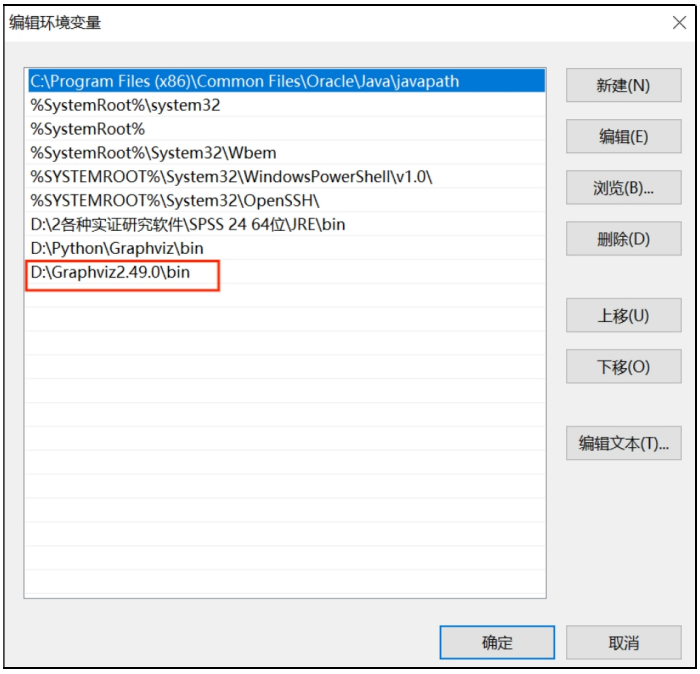

图 1.37　添加了 bin 文件夹路径

(6) 查看安装软件的正确性：在 windows 的命令行界面中，键入“dot-version”，单击“确定”，当 Graphviz 的有关版本信息出现时，说明系统已安装并配置了 Graphviz 软件。

(7) 安装 Python 的 Graphviz 工具。

在 cmd 中输入命令“pip install graphviz”，安装完后，再输入命令“pip list”，以确认是否安装成功(图 1.38)。

图 1.38　安装 Graphviz 工具并确认

(8) 完成以下 2 个设置，Python 才能调用 Graphviz 工具。首先，打开 cmd 语句输入以下命令：

```
echo process1 = subprocess.Popen(command1,stdout=subprocess.PIPE,shell=True)
```

其次，在 Python 安装路径下，找到文件 subprocess.py 并双击打开(图 1.39)。

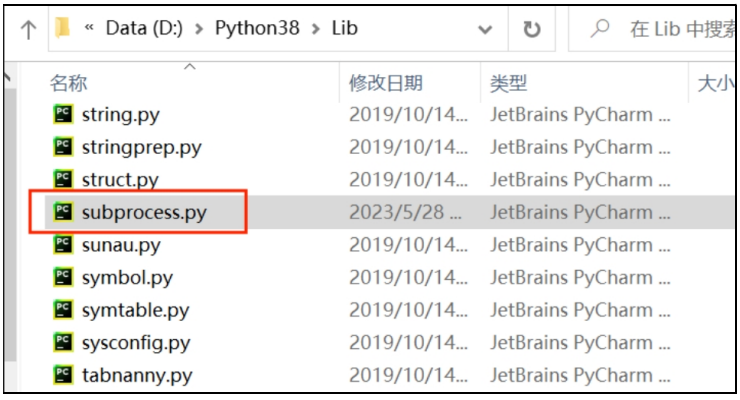

图 1.39　找到 subprocess.py 文件

找到 Popen(object)类，找到这个类的 init 方法，把“shell=False”修改为“shell=True”，如图 1.40 所示。

```
class Popen(object):
    """Execute a child program in a new process...."""
    _child_created = False  # Set here since __del__ checks it

    def __init__(self, args, bufsize=-1, executable=None,
                 stdin=None, stdout=None, stderr=None,
                 preexec_fn=None, close_fds=True,
                 shell=True, cwd=None, env=None, universal_newlines=None,
                 startupinfo=None, creationflags=0,
                 restore_signals=True, start_new_session=False,
                 pass_fds=(), *, encoding=None, errors=None, text=None):
        """Create new Popen instance."""
        _cleanup()
```

图 1.40　修改为 shell=True

修改好之后，创建一个文件，输入以下代码测试 Graphviz 是否运行正常。

```
from graphviz import Digraph
dot = Digraph('测试')
dot.node("1","Life's too short")
dot.node("2","I learn Python")
dot.edge('1','2')
dot.view()
```

如果看到图 1.41 所示内容，说明 Python 能调用 Graphviz 工具了。

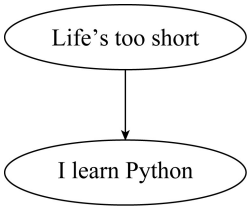

图 1.41　可调用 Graphviz 示意图

1.4.4　Netica 软件

1. 软件介绍

构建贝叶斯网络时，Netica 软件是不可或缺的工具。作为目前应用广泛的贝叶斯网络分析软件，Netica 因其简便性、稳定性和高效性而受到全球众多知名企业的青睐，成为政府机构在决策过程中不可或缺的工具。

首先，构建网络结构。在 Netica 中，可以创建三种类型的节点：状态节点(Nature Node)、决策节点(Decision Node)和效用节点(Utility Node)。状态节点是最为常见的，代表了各个变量可能存在的状态，并展示了这些状态对应的概率值。一旦父节点的概率被确定，子节点就会基于上层节点计算出条件概率，并能够动态地调整状态概率的变化。

其次，创建状态节点。双击黄色椭圆，可以创建多个状态节点；单击黄色椭圆，则创建一个状态节点。构建完成后，退出时需要再次单击黄色椭圆以确认。

再次，创建并指示关系方向。单击父节点名称，再单击子节点名称，并使用箭头连接不同的状态节点，即可完成关系方向的创建。单击状态节点，鼠标拖拽，即可调整节点的位置，箭头会自动随之调整。

最后，对节点进行重命名。双击状态节点，弹出“属性”对话框，在其中输入节点的名称(Name)。然后输入状态(State)，操作方法为：右击“Modify”→“Set States”，每个状态占据一行，之后单击“OK”。使用相同的方法为其他状态节点输入所有状态(States)。

2. 下载安装

Netica 软件可在官网下载页面 https://www.norsys.com/download.html 下载，用户可选择适合自己计算机操作系统版本的安装包进行下载。具体步骤如下。

(1) 打开 Netica 软件网址https://www.norsys.com/download.html，里面有适合 Windows、Mac、Linux 等系统的软件安装包。这里以 Windows 版本的安装包为

例进行下载。单击“Download”下载安装包(图 1.42)。

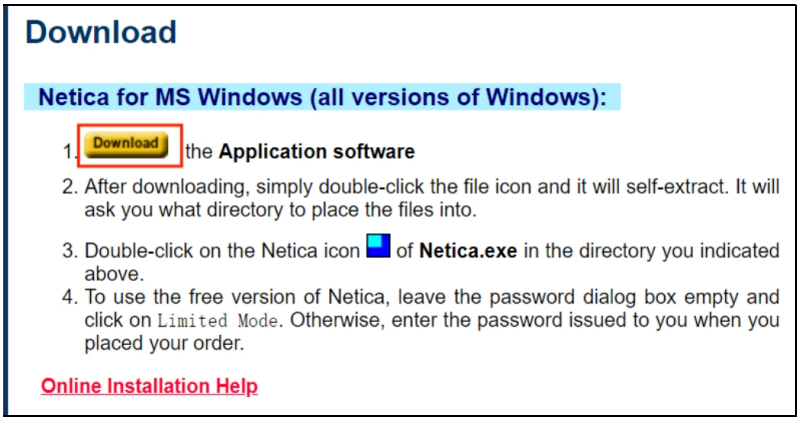

图 1.42　单击“Download”下载安装包

(2) 下载后双击安装包，单击“运行”按钮(图 1.43)。

图 1.43　单击“运行”按钮

(3) 输入 Netica_Win.exe 的解压路径，单击“Unzip”按钮，将 Netica_Win.exe 中的所有文件解压到指定的文件夹中，如图 1.44 所示。解压完成后会弹出如图 1.45 所示对话框，单击“确定”。

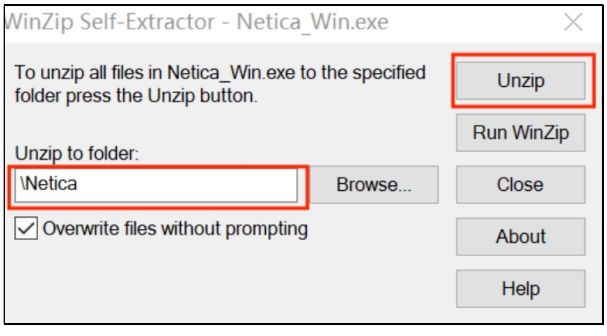

图 1.44　单击“Unzip”按钮

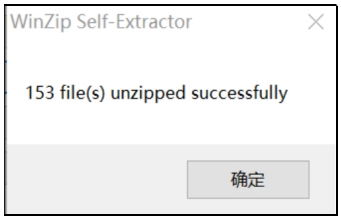

图 1.45　解压完成对话框

(4) 双击上述目录中 Netica.exe 的 Netica 图标(图 1.46)。

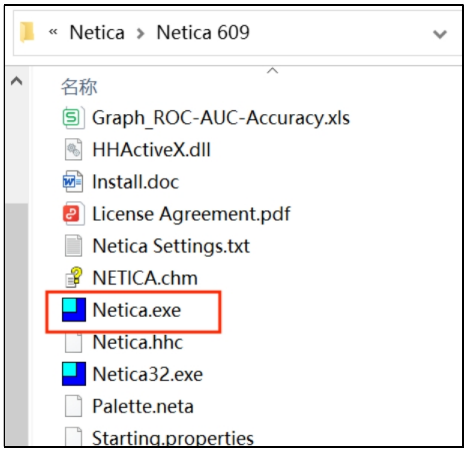

图 1.46　双击 Netica 图标

(5) 在弹出的界面(图 1.47)中，要使用免费版的 Netica，将密码对话框保留为空，然后单击“Limited Mode”。若要获得全部功能，需要购买该软件，并获取密码。最终软件初始界面如图 1.48 所示。

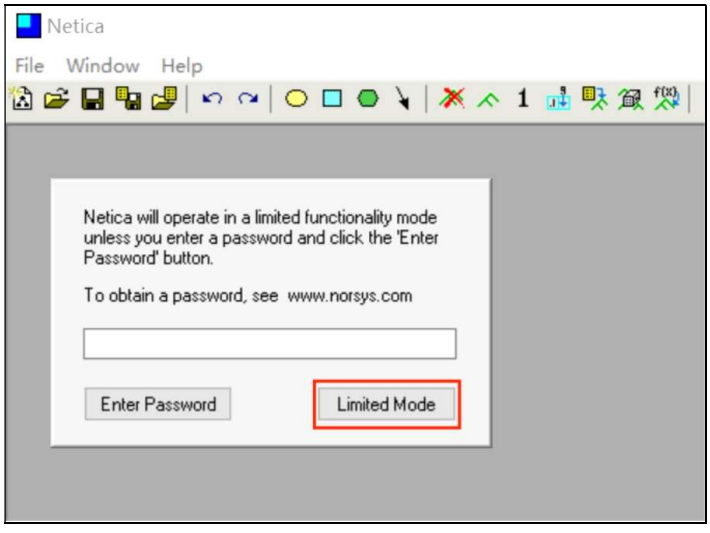

图 1.47　Netica 界面

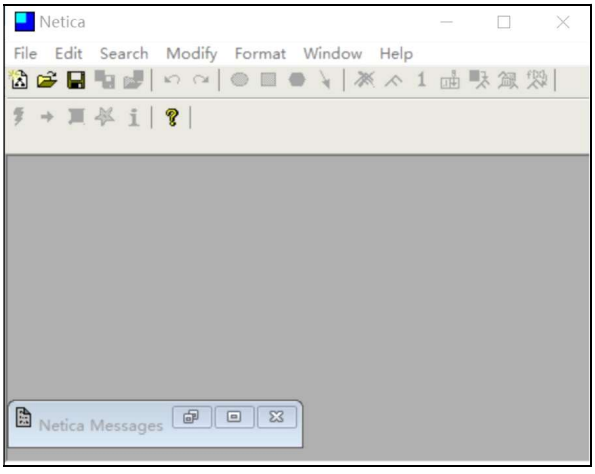

图 1.48　最终软件初始界面

1.5 机器学习方法

数据决策分析致力于运用机器学习算法来解读系统内部复杂因素的运作机制并发掘知识。其核心在于以管理学研究问题为指导，深入探讨管理研究中出现的情境差异、因素间的相关性以及复杂因果关系的识别等关键问题。本节重点阐述 DAC 框架所涉及的机器学习算法。这些算法包括云模型、聚类算法、决策树算法、随机森林、贝叶斯网络以及爬山算法等。

1.5.1　云模型

李德毅院士，一位在人工智能领域享有盛誉的学者，提出了一个名为“云模型”的创新概念。这个模型不仅巧妙地桥接了定性概念与量化表达之间的鸿沟，而且深刻地模拟了人类在认知过程中所固有的模糊性和随机性。这一理论的提出，为处理不确定性问题的人工智能研究领域注入了新的活力，并开辟了新的研究方向(李德毅，2000)。云模型的跨学科特性让它与集对分析、粗糙集理论、机器学习等不同领域的技术优势得以互补。这种互补不仅促进了云模型在评估领域的显著进步，也使得它在多个相关领域中得到了广泛应用(张园等，2023)。

在数据挖掘这个充满挑战的领域，云模型展现出了独特的魅力。它能够深入挖掘过程中的不确定表达，并且能够以定性的方式呈现挖掘结果，为数据的