第 5 章

深 度 学 习

深度学习是机器学习的一个重要分支，它通过构建和训练多层神经网络模型，使计算机能够从大量数据中学习复杂的模式和规律。深度学习在图像识别、语音识别和自然语言处理等领域取得了显著成果，助力了人脸识别、图像识别和自然语言处理等技术的迅猛发展。

强化学习也是一种机器学习方法，它通过让智能体与环境互动，学习如何在不同情境下采取最佳行动，以最大化累积奖励。强化学习在自动驾驶、游戏 AI 和机器人控制等领域展现了强大的潜力。例如，强化学习算法在围棋和其他复杂游戏中已经超越了人类顶尖选手，并且在机器人控制领域实现了自主学习复杂任务的能力。

深度学习和强化学习是机器学习技术的左膀右臂，在近二十年人工智能的高速发展中起到了关键作用。这两种技术的进步推动了人工智能的发展，使得计算机系统在处理复杂任务和适应新环境方面表现出更高的智能和灵活性。

5.1 从机器学习到深度学习

第 1 章讲到，在人工智能研究的波浪式前进中，机器学习起到了至关重要的作用。机器学习是一种让机器通过数据学习和改进的方法，使得机器能够根据过去的经验不断优化自己的性能。深度学习作为机器学习的一个分支，通过构建多层次的神经网络模型，使得机器可以从大量数据中提取抽象的特征和模式，从而实现诸如图像识别、自然语言处理等复杂任务。特别是深度学习也是人工智能研究进入第三个高潮的最重要推手。人工智能、机器学习和深度学习的包含关系如图 5-1 所示。

图 5-1 人工智能、机器学习和深度学习的包含关系

5.1.1 机器学习的发展历程

机器学习的发展分为三个阶段，涵盖传统机器学习、深度学习和大规模预训练模型。

1. 第一阶段是传统机器学习的发展阶段

这个阶段主要是基于统计学和数学方法的传统机器学习算法，例如线性回归、支持向量机、决策树、朴素贝叶斯等。早在 20 世纪五六十年代，就已经出现了感知机、逻辑回归和 K-近邻算法等，直到 1980 年，在美国卡内基-梅隆大学举行了第一届机器学习研讨会，才标志着机器学习研究作为一个独立方向在全世界范围内的兴起，之后传统机器学习快速发展了

二十多年。

线性回归是一种用于预测连续数值输出的回归算法，它基于对输入特征和输出之间的线性关系进行建模，通过最小化预测值和真实值之间的误差得到最佳的模型参数。支持向量机是一种二分类的有监督学习算法，在数据集中找到一个最优决策边界，以尽可能最大化两个不同类别之间的间隔，可以高效地处理高维数据，并且具有较强的泛化能力。决策树是一种基于树状结构的分类和回归算法，通过对特征值的逐步划分构建一个树状的决策流程，易于理解和解释，并且可以处理数值型和离散型数据。朴素贝叶斯是一种基于贝叶斯定理的分类算法。它假设所有特征之间相互独立，通过计算后验概率得到最佳分类结果，计算简单，效果良好。这些传统机器学习算法在数据处理、模式识别、分类、预测等方面都得到了广泛应用，常用于房价预测、文本分类、股票趋势分析、金融风险评估等任务。

然而，当面对复杂数据和高维特征空间时，传统机器学习方法也展现出一定的局限性。首先面临维度灾难：随着特征维度的增加，传统机器学习方法的性能可能迅速下降。在高维特征空间中，数据变得稀疏，将高维空间的分类结果投影在低维空间中，容易使分类器学习过多的样本数据的异常特征（即噪声），出现过拟合问题，在新数据上的泛化能力不佳。其次是特征工程困难：在高维特征空间中，如何选择合适的特征并进行特征工程变得更加困难。传统方法需要大量的人工专业知识来进行特征选择和提取，耗时更多，且难以保证最佳性能。此外，高维特征空间中可能存在大量的特征交互关系，利用传统方法难以捕捉，从而限制了模型的性能。

2. 第二阶段是深度学习的发展阶段

深度学习的概念由辛顿于 2006 年在《科学》（*Science*）上发表的论文《深度学习》（*Deep Learning*）中提出，是一种基于深度神经网络的机器学习方法，通过多层次的神经网络模型学习数据的特征和模式。作为新一代机器学习方法，深度学习依托多层神经网络结构和大规模数据训练的能力，能够自动地从原始数据中学习高层次的抽象特征，从而在计算机视觉、自然语言处理和语音识别等领域取得了突破，可以说深度学习使得机器学习大放异彩。

卷积神经网络（convolutional neural network，CNN）是深度学习最具代表性的方法之一，在图像分类、目标检测和图像生成等计算机视觉任务上取得了巨大成功。与此同时，循环神经网络（recurrent neural network，RNN）作为另一种代表性深度学习方法，因其出色的序列建模能力，在自然语言处理、语言识别、时间序列预测等任务上大放异彩。

深度学习在自然语言处理、图像识别、语音识别等领域取得了重大突破和成功，并得到了广泛应用，例如智能语音助手 Siri、谷歌翻译等。2016 年，谷歌的围棋人工智能 AlphaGo 在比赛中战胜围棋世界冠军李世石，最终总比分为 4∶1。这一事件对于人工智能领域的发展具有重要意义，因为围棋的复杂性和不确定性远超出国际象棋等传统游戏，AlphaGo 的胜利显示了深度学习和强化学习等技术在处理复杂决策问题方面的潜力。这些重要事件，从特定领域的竞技，到自然语言处理，再到复杂决策，凸显了人工智能的不断进步，展示了人工智能在模拟智能能力方面的成功尝试。

3. 第三阶段是大规模预训练模型的发展阶段

以 CNN、RNN 为代表的深度学习模型需要在大量的标注数据上进行训练，对数据品质的依赖性高，且容易出现过拟合问题。由此诞生了预训练模型这一范式。早在 2003 年，本吉奥（Bengio）等提出的神经网络语言模型（neural network language models，NNLM）就已

经通过在大量语料上的训练,用词的分布式表示实现对自然语言序列的建模。

预训练模型是指在大规模数据上预先训练好的神经网络模型,然后在大规模未标记数据上进行自监督学习或多任务学习得到(也有部分计算机视觉领域的预训练模型为有监督学习)。对于特定任务,只需要在少量标注样本上进行微调或迁移学习,就可以达到与从头开始训练相当甚至更好的效果,大大减少了训练时间和样本需求。

随着数据和模型规模的大幅增加,大规模预训练模型(以下简称"大模型")应运而生。大模型的参数通常在十亿以上,模型大小可以达到数百 GB 甚至更大,具有强大的表达能力和学习能力。大模型在自然语言处理领域展现了先进的能力,如问答、机器翻译、摘要提取、代码生成、写作润色等。大模型在计算机视觉领域同样得到了广泛应用。未来大模型将为产品交互、企业生态、商业模式、个人创作等带来深刻的变革。

5.1.2 深度学习的发展脉络

深度学习被学界熟悉并广泛应用之前,绝大多数机器学习和信号处理技术都利用浅层神经网络结构,这些结构一般包含最多一到两层的非线性特征变换。浅层结构包括高斯混合模型(Gaussian mixture model,GMM)、线性或非线性动力系统、条件随机场(conditional random field,CRF)、最大熵模型(maximum entropy models,MaxEnt)、支持向量机(support vector machine,SVM)、逻辑回归(logistic regression,LR)、核回归等方法。浅层结构在解决很多简单的或者限制较多的问题上可以取得良好的效果,但是由于其建模和表示能力有限,在遇到自然语言、视觉图像等数据更复杂的情境时就会遇到各种困难。

作为解决上述困难的有效方法,深度学习起源于对 ANN 的研究。前馈神经网络或多层感知机[3]被认为是最早的深度神经网络(deep neural network,DNN)。具有里程碑意义的反向传播(back-propagation,BP)算法流行于 20 世纪 80 年代,是广为人知的一种有效学习网络参数的算法。更高的算力和更好的学习算法也促使了 DNN 的成功。在训练过程中,深而宽网络的使用不仅显著提高了 DNN 的建模能力,而且创造出了许多接近的最优配置。2003 年,杨立昆(Yann LeCun)提出的随机梯度下降算法[4]在大多数训练集较大且冗余的情况下是最有效的算法。但是,在优化目标为非凸函数的 DNN 中,来自局部最优化或其他最优化问题的挑战普遍存在,当使用批量梯度下降或随机梯度下降的 BP 算法时,目标函数经常会陷入局部最优的情况。随着网络层数的加深,局部最优的情况也就会变得越来越严重,这些挑战通常是学习中面临的主要困难。

2006 年,辛顿提出了一种高效的无监督学习算法——深度置信网络[5](deep belief network,DBN),其由一组受限玻尔兹曼机堆叠而成,经验性地缓解了与深度模型相关的最优化难题。多层感知机或 DNN 通过无监督的 DBN 来进行预训练,然后通过 BP 微调来优化。实验证明,使用配置好的 DBN 来初始化多层感知机的权重取得了比随机初始化的方法更好的结果。除了具有好的初始点,DBN 还有一些颇具吸引力的优点,首先它的学习算法可以有效使用未标注的数据,其次它可以看作一个概率生成模型,最后过拟合和欠拟合问题都可以通过预训练方式解决。

对于 DNN 学习的高度非凸优化问题,由于优化是从初始模型开始的,所以更好的参数初始化技术将会打造出更好的模型。一种与 DBN 性能相当且有效的初始化方法是对 DNN 进行逐层预训练,通常将每两层视为一个除噪自编码器,该除噪自编码器通过将输入节点的

随机子集设置为零而进行正则化。另一种方法是使用压缩自编码器，它通过使输入变量具有更好的鲁棒性来达到同样的目的。除了无监督预训练外，有监督的预训练也被证明是很有效的，并且在标记的训练数据充足的情况下比无监督的预训练技术表现得更好。

有研究人员还分析了深度学习在语音和图像中捕获了哪些信息，他们发现，DNN 的隐藏激活向量保留了与多个尺度上的特征向量相似的结构。我们有理由认为，深层网络的强大之处在于它们拥有在提取合适特征的同时做判别的能力。

2009 年，斯坦福大学的李飞飞教授创建了包含超过 1400 万张标注图像的大型数据集 ImageNet[6]，并在 2010 年启动了 ImageNet 大规模视觉识别挑战赛（ILSVRC）。创立 ImageNet 数据集和挑战赛的目的是评估大型数据集上的图像分类架构，它带来了许多新颖的、强大的、有趣的视觉架构。

2012 年，辛顿团队在该挑战赛上使用了一个名为 AlexNet 的 CNN[7]，一举夺得冠军，将错误率从 26%降低到 15%。AlexNet 网络的成功证明了深度卷积神经网络可以很好地处理视觉识别任务，也引发了 CNN 在图像识别领域的革命。

此后几年，VGGNet[8]、GoogLeNet[9]等卷积架构的网络相继出现，它们的架构不断变大，并且也取得了更好的效果，但网络深度的增加也给训练带来了困难。为此，2015 年，何恺明提出了 ResNet[10]，通过引入残差连接解决了深层网络的训练问题，并且进一步将错误率降低到 3.6%。随后，深层次的网络架构井喷式出现，并在 ImageNet 图像挑战赛上不断刷新纪录，同时也在图像分割、目标检测、人脸识别等图像相关任务上取得了显著效果。

因在深度神经网络概念和工程上的突破，本吉奥、辛顿和杨立昆三位学者获得了 2018 年的图灵奖。

5.2 机器学习的问题与方法

5.2.1 机器学习问题

研究机器学习，首先要对机器学习问题进行定义。如图 5-2 所示，我们希望对不同的图像内容识别出其中物体的类别，例如“轮船”“汽车”“花朵”。该问题可以转换为一个机器学习 $y=f(x)$。其中 x 表示输入的图像特征，f 为预测函数，代表机器学习方法，y 为预测函数的输出。采用训练集中的数据对机器学习模型进行优化，期望预测函数能够应用到新的样本中，并获得准确的结果。

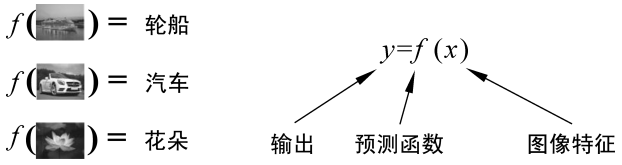

图 5-2 机器学习问题

机器学习的目的是研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构，使之不断改善自身的性能。机器学习方法是人工智能的核心，可以应用到人工智能的各个领域。

常见的机器学习方法包括监督学习、半监督学习、无监督学习和强化学习等,如图 5-3 所示。

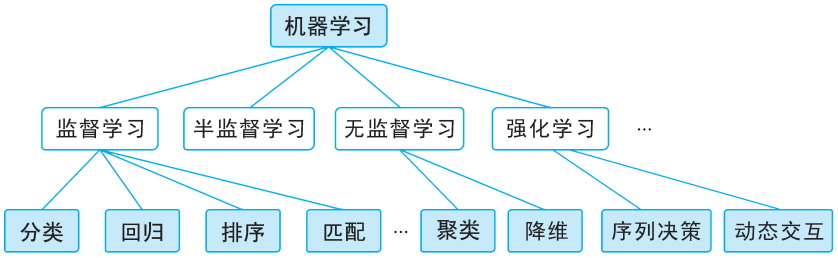

图 5-3 机器学习的分类

监督学习(supervised learning)[13]:是从标记的训练数据来推断一个功能的机器学习任务。训练数据包括一套训练实例。在监督学习中,每个实例都是由一个输入对象(通常为矢量)和一个期望的输出值(也称为监督信号)组成的。监督学习算法是分析该训练数据,并产生一个推断的功能,其可以用于映射出新的实例。目前,监督学习被广泛应用到分类、回归、排序、匹配等多种任务中。

无监督学习(unsupervised learning)[14]:无监督学习的问题是,在未加标签的数据中试图找到隐藏的结构。因为提供给模型学习的实例是未标记的,因此没有错误或奖励信号来评估潜在的解决方案。无监督学习可以减少人工标注的时间,降低人工成本。无监督学习算法常用于聚类和降维任务。

半监督学习(semi-supervised learning):这是一种结合了监督学习和无监督学习的技术。同时使用"少量标记数据"和"大量未标记数据",在降低人工标注成本的同时,少量标记数据也可以指导学习。半监督学习往往采用一些假设,如平滑假设、聚类假设、流形假设等。核心优势在于降低标注成本,提升模型性能,适应数据稀缺场景。半监督学习在自监督学习、对比学习中的应用也在不断扩展。

强化学习(reinforced learning)[15]:令模型以"试错"的方式学习,当模型学习正确的时候,给模型一个奖励。强化学习是智能系统从环境到行为映射的学习,学习目标是使奖励信号(强化信号)函数值最大。由于外部环境提供的信息很少,必须靠自身的经历学习。通过这种方式,模型在行动-评价的环境中获得知识,改进行动方案,以适应环境。它适用于许多需要序列决策或动态交互的任务。

机器学习问题可以根据数据的"有监督/无监督"以及输出数据是"连续/离散"进行分类,分为统计分类、回归分析、聚类分析、降维四类典型问题,如图 5-4 所示。有监督的学习方式要求训练集包括输入和输出,也可以说是特征和目标。其中训练集中的目标是人工标注的。区别于有监督的数据,当训练集没有人工标注结果时,则是无监督的。

下面分别介绍统计分类、回归分析、聚类分析和降维四个典型的机器学习问题的方法。

1. 统计分类

统计分类通过部分已知的离散的观测数据来分析事物的规律,从而对其他数据做出分类。以经典的统计分类问题"泰坦尼克号之灾"为例,使用机器学习创建一个模型,预测哪些乘客在泰坦尼克号沉船事故中幸存下来。在这个问题中,可以使用乘客数据(如姓

	有监督	无监督
离散	classification 统计分类	clustering 聚类分析
连续	regression 回归分析	dimensionality reduction 降维

图 5-4 机器学习问题

名、性别、年龄、舱位等）建立一个预测模型来分析“什么样的人更有可能生存”，如图 5-5 所示。

姓名	性别	年龄	同船亲属	船票	金额/英镑	舱位	登船港口
Braund	男	22	1	A/5 21171	7.25	NaN	S
Cumings	女	38	1	PC 17599	71.28	C85	C
Heikkinen	女	26	0	STON/O2.	7.92	NaN	S
Futrelle	女	35	1	113803	53.10	C123	S
Allen	男	35	0	373450	8.05	NaN	S

输入：乘客信息（姓名、性别、年龄、舱位……）

→ 是否存活？

0 or 1

输出

图 5-5 经典案例：泰坦尼克号之灾

采用统计分类方法可以对此类问题进行建模。在通常情况下，数据集有 N 个训练对象 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$。对于每个样本 $\boldsymbol{x}_i$，提供标签 t_i 描述其所属类别。其中 t_i 是离散标量，$\boldsymbol{x}_i$ 是 D 维的特征向量。每个对象都是一个 D 维的特征向量。对于测试集中给定的对象 $\boldsymbol{x}_{\text{new}}$，分类任务需要预测它的类别 t_{new}。常用的分类方法有贝叶斯分类器、逻辑回归、K-近邻算法、支持向量机等。

2. 回归分析

回归分析是一种统计方法，用于研究一个或多个自变量（independent variables）和因变量（dependent variable）之间的关系。回归分析同样是一种监督学习方法，但与分类问题不同的是，它的输出数据是连续的。一个经典的回归分析案例是房价预测问题。通过多项关于房屋信息的特征（如卧室数量、占地面积等）来预测每套房屋的最终价格，如图 5-6 所示。

编号	卧室数量	占地面积/m²	街道	占地形状	泳池面积/m²	销售月	销售年	…
1	60	8450	石板路	规则	0	2	2008	…
2	20	9600	石板路	规则	0	5	2007	…
3	60	11250	石板路	略不规则	0	9	2008	…
4	70	9550	石板路	略不规则	0	2	2006	…
5	60	14260	石板路	略不规则	0	12	2008	…

输入：住宅特征

→ 房价

（连续数值）

输出

图 5-6 经典案例：房价预测

回归分析是一种预测性的建模技术,通常用于预测分析、时间序列模型以及发现变量之间的因果关系,在许多领域都有着广泛应用,包括经济学、营销学、社会学、医学、工程学等。常见的回归模型有线性回归、岭回归、多项式回归等。

1) 线性回归

线性回归(linear regression)假设因变量与自变量之间呈线性相关,换句话说,因变量可以通过自变量的线性组合来解释。如果回归分析中只包含一个自变量和一个因变量,且二者的关系能够拟合为一条直线(图 5-7),称为一元线性回归。如果包含两个以上的自变量,且因变量和自变量之间的关系是线性的,称为多元线性回归,拟合的是多维空间中的一个超平面。

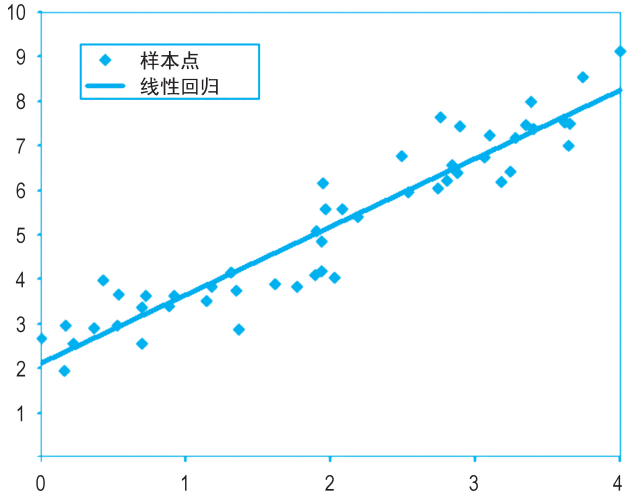

图 5-7　一元线性回归示意图

2) 岭回归

岭回归(ridge regression)是线性回归模型的一种正则化改进方法。当自变量特征维度较高时,回归系数的估计可能方差过大,对数据扰动敏感,导致模型过拟合。为解决这一问题,岭回归在原始损失函数中引入 L2 正则化项,并且乘以一个非负参数正则化强度 α。

图 5-8 展示了在岭回归方法中不同 α 与回归系数值之间的关系。随着 α 增加,回归系数的绝对值趋向于减小,从而减少模型的方差,但可能增加偏差。岭回归的一个关键优点是它能够在保证模型具有一定程度准确性的同时,减少模型的复杂性和过度拟合的风险,特别适用于那些特征维度高于样本量的情况。

3) 多项式回归

许多情况下,数据的分布难以拟合为线性模型,常用的做法是多项式回归(polynomial regression)。多项式回归是一种在线性回归的基础上进行扩展的方法,用于建模因变量和自变量之间的非线性关系。与简单的线性关系不同,多项式回归允许自变量的多项式函数来拟合数据,从而更灵活地捕捉数据之间的复杂关系。

如图 5-9 所示,图中的直线代表用线性回归拟合数据的结果,曲线表示用二次多项式拟合数据的结果,针对图中数据的分布,多项式回归对数据的拟合能力更强。

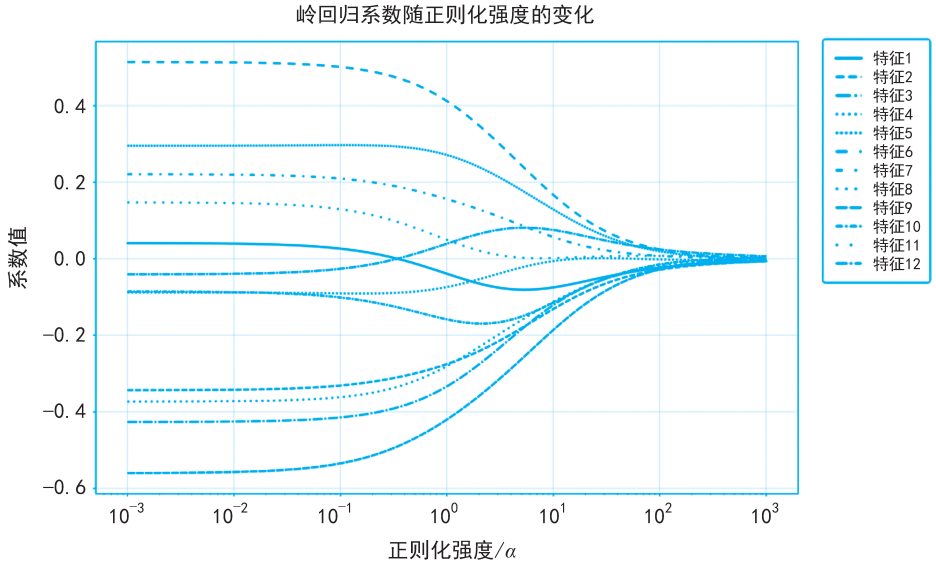

图 5-8　岭回归示意图

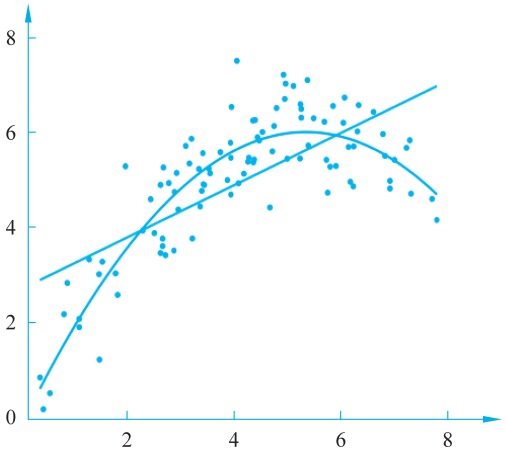

图 5-9　多项式回归和线性回归对比

选择了合适的回归分析方法之后，需要选择一组合适的参数，使模型最好地拟合样本数据，称为参数估计。常用的参数估计方法包括最小二乘法、最大似然法等。

最小二乘法通过最小化误差的平方和寻找数据的最佳函数匹配。利用最小二乘法可以简便地求得未知的数据，并使得这些求得的数据与实际数据之间误差的平方和最小。

最大似然法采用最大似然估计预测结果。当从模型总体随机抽取 n 组样本观测值后，最合理的参数估计量应该使得从模型中抽取该 n 组样本观测值的概率最大。

3. 聚类分析

“物以类聚，人以群分”。个体的属性往往存在某些倾向和共性，因此理论上可以通过定义某种规则，将个体按其属性划分到不同的组（簇）内，从而将整体划分成多个部分，以获得对整体分布的某种认识。如图 5-10 所示，左侧是由苹果、香蕉、葡萄等多种水果个体混杂在

一起组成的一个水果“整体”。假设我们并不具有对这些水果的先验认识，我们仍可以按形状、颜色、尺寸、气味等个体属性，将同一类型的水果聚合到一簇中，这有助于对这些水果的进一步考查和认识。

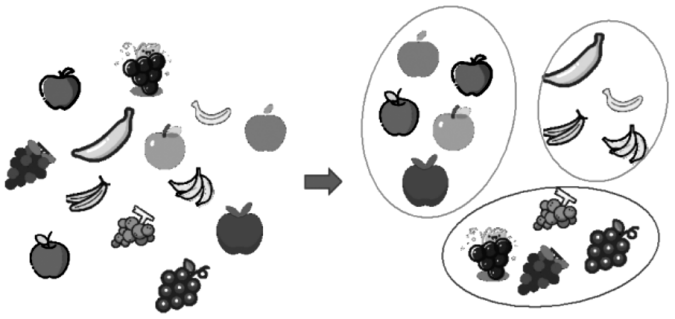

图 5-10　聚类分析

聚类算法正是研究这种分类问题的机器学习算法。它试图发现数据中自然存在的分组结构。它的核心目标是将一组未标记的数据对象（样本个体）划分成多个组（称为“簇”），使得簇内相似性尽可能高，而簇间相似性尽可能低，即同一个簇内的对象彼此之间应该尽可能相似，而不同簇的对象彼此之间应该尽可能不同，只有这样，划分才是有意义的。

聚类算法是一种无监督学习方法。聚类算法完全基于数据本身的属性规律进行探索性分类，不需要知道类别的意义是什么或存在哪些类别，而需要算法的执行者，即人类根据对数据代表的实际对象和相关的先验知识理解和应用。作为对比，分类算法则是基于已知的分类规则，其意义在于将少量已知数据的标注反映的分类规律扩充到大量未知数据。聚类算法是数据挖掘和知识发现中不可或缺的关键技术。

1）划分方法（Partition Methods）

划分方法的核心思想是，预先指定要划分的簇数 k，将数据对象划分到 k 个互斥的簇中，并通过迭代优化（如最小化簇内距离平方和）来改进划分。这 k 个簇需要满足：①每一个分组至少包含一个数据记录；②每一个数据记录属于且仅属于一个分组。算法从一个初始划分开始，通过迭代重定位（将对象从一个簇移动到另一个簇）来优化划分质量。优化的核心准则是：最大化簇内相似性（或最小化簇内距离），最小化簇间相似性（或最大化簇间距离）。划分方法一般是基于距离度量的。

K-Means 算法是最经典和广泛使用的算法，假设簇是凸形且大小相近的，选择 k 个初始质心，将每个点分配到最近的质心，重新计算质心，迭代直到质心稳定或达到最大迭代次数。其他代表性算法包括 K-Medoids（PAM、CLARANS）等。

划分方法为达到全局最优，可能需要穷举所有划分，计算量巨大，故实践中广泛采用启发式方法寻找局部最优解，渐进地提高聚类质量，逼近局部最优解。划分方法是一类基础算法，擅长发现中小规模数据集中的球状或凸形簇，对初始质心敏感，对噪声和离群点敏感。为了发现具有复杂形状的簇和对超大型数据集进行聚类，需要进行进一步的扩展。

2）层次方法（hierarchical methods）

与划分方法相比，层次方法不预设簇数 k，而是构建一个簇的层次结构。该结构可以通过两种策略生成：①凝聚法（自底向上）：初始时每个对象自成一簇。在每一步迭代中，将

最相似(距离最近或密度最高)的两个簇合并,直到所有对象聚合成一个大簇或满足终止条件。②分裂法(自顶向下):初始时所有对象属于一个簇。在每一步迭代中,将最不相似(距离最远或密度最低)的簇分裂,直到每个对象自成一簇或满足终止条件。代表算法包括BIRCH、CURE、CHAMELEON 等。层次方法可以是基于距离、密度或连通性的。

层次方法通过生成不同粒度的簇划分,可以提供更全面的数据结构视图,且树状图本身具有直观的可解释性。

3) 基于密度的方法(density-based methods)

基于密度的方法假设簇是数据空间中的高密度区域,被低密度区域分隔。这类算法不依赖全局距离,而是基于局部密度进行聚类:只要某个区域内的数据点密度超过给定阈值,就将这些点及其邻近的高密度区域连接起来,形成一个簇。它能够有效识别任意形状的簇,代表算法包括 DBSCAN、OPTICS、DENCLUE 等。

基于密度的方法能发现任意形状的簇,对噪声和离群点有较好的鲁棒性,且通常不需要预先指定簇数 k,特别适合处理空间数据。但它对密度阈值参数敏感,在高维数据或密度差异大的数据上效果可能下降,且全局密度参数可能难以适用于所有区域。

4) 基于模型的方法(model-based methods)

基于模型的方法假设数据是由潜在的概率分布过程生成的,这类算法为每个簇假定一个特定的数据生成模型(如概率分布模型),然后寻找最能拟合数据的最佳模型及其参数(即簇),代表性算法分为统计模型(如 GMM)和神经网络模型(如 SOM)两类。

基于模型的方法能提供簇的统计框架描述,具有坚实的理论基础,能给出对象属于各簇的概率(软聚类)。对某些特定类型的数据(如符合假设分布的)效果很好。其局限性在于模型结构的选择,当模型假设与真实数据结构差异较大时,可能导致聚类效果不佳。此外,复杂的模型可能计算量过大,需考虑高效的优化方法。

5) 基于网格的方法(grid-based methods)

基于网格的方法将数据空间划分为有限数量的网格单元(cell),然后将聚类操作转化为对网格单元的统计和处理。所有对象根据其属性值被映射到相应的网格单元中,代表性算法包括 STING、CLIQUE、WAVE-CLUSTER 等。

基于网格的方法处理速度极快,因为处理时间主要取决于空间划分的网格单元数,而与原始数据对象的数量 N 基本无关,因此非常适合处理海量数据集,也易于与其他聚类方法(如基于密度的方法)集成。其局限性在于聚类的质量依赖网格的粒度和划分方式,边界处理可能影响结果精度,且可能丢失数据细节。

6) 现实挑战

上述传统聚类方法在低维数据上取得了显著成功。但面对现实世界中日益普遍的高维数据(如基因序列数据,可能有成百上千甚至更高维度),上述传统聚类方法面临着严峻挑战。首先,高维空间中的数据分布极其稀疏,对象间的距离趋于均等化,使得基于距离的相似性度量失效。传统方法(尤其是基于距离和密度的)的性能在高维空间急剧下降,难以发现有效的簇结构;其次,噪声在高维空间中被显著放大,由于大量不相关或冗余属性的存在,在所有维度上同时存在有意义的簇变得极为困难。高维聚类分析已成为聚类领域最活跃的研究方向之一,是推动数据挖掘深度应用、实际应用的关键。