第 5 章　大模型数据分析

随着大数据时代的到来，数据分析已成为推动现代企业和行业发展的核心动力。它与人工智能一起，正逐步改变各行各业的运作模式。本章阐述数据分析的核心概念与方法，运用大模型技术为数据可视化、回归分析及聚类分析等应用提供切实可行的解决方案，并通过实例演示，助力读者掌握相关的技能。

5.1　数据分析概述

大模型不仅在自然语言处理领域具有广泛应用场景，还在数据分析、推荐系统等领域展现出了强大的应用潜力。作为智能化、个性化服务的提供者，大模型正逐步成为企业和个人在数据驱动决策过程中不可或缺的重要工具。其强大的生成能力使得非专业人士也能迅速掌握并应用大模型技术，有效降低了数据分析和人工智能的入门门槛。

5.1.1　数据分析的定义

数据分析是对数据进行系统的收集、清洗、分析、解释和展示的过程，旨在从数据中发现有用的信息和知识。这一过程的核心在于通过技术手段揭示数据背后的规律、趋势和关联，从而为决策提供帮助。大模型在数据分析中的优势主要体现在自动化、智能化和准确性等方面。它能够自动处理海量数据，发现数据中的隐藏模式和关系，提供深入的洞察和决策支持。

数据分析在大模型技术中的重要性不言而喻，它不仅是大模型训练和优化的基础，也是大模型能够在实际应用中发挥效能的关键。以下是对数据分析重要性的解释。

1. 数据分析是大模型训练的基础

大模型依赖大规模数据集进行训练。数据分析能够从这些数据中提取有用的特征和信息，为模型提供学习的基础。通过数据分析，可以识别并去除数据中的噪声和异常值，提高数据的纯净度和质量。这有助于大模型准确地学习数据的分布和模式，从而提升模型的性能。

2. 数据分析支持大模型的优化

数据分析可以帮助识别对模型性能影响最大的特征，并进行特征选择和降维处理。这可以降低模型的复杂度，提高模型的泛化能力，并降低过拟合的风险。在大模型的训练过程

中，数据分析可以用于评估不同参数设置对模型性能的影响。通过数据分析，可以找到最优的参数组合，从而提高模型的性能。

3. 数据分析助力大模型的应用

大模型在实际应用中，如金融风控、医疗诊断和智能推荐等，需要依赖数据分析来进行预测和决策。数据分析能够帮助模型识别出潜在的风险和机会，为实际应用提供有力的支持。通过数据分析，大模型可以了解用户的偏好和需求，从而提供个性化的服务。例如，在电商领域，大模型可以根据用户的浏览和购买历史，推荐符合用户喜好的商品。

4. 数据分析推动大模型技术创新

数据分析能够从数据中挖掘出潜在的知识和模式，为大模型的创新提供灵感和方向。例如，通过数据分析，可以发现某些特征之间的关联关系，从而设计出更有效的算法和模型。数据分析在大模型技术中的应用促进了计算机科学、统计学和数学等多个学科的融合与发展，这种跨学科融合为大模型技术创新与发展提供了新的思路和方法。

尽管大模型在数据分析中取得了显著成果，但仍面临数据偏见、模型可解释性不足等挑战。未来的研究需要更加注重算法的公平性和可解释性，并加强跨学科合作，推动技术创新和应用拓展。

5.1.2　数据分析的特点

1. 数据规模庞大

大模型通常需要处理海量的数据，这些数据规模可能达到 TB 级别甚至 PB 级别。庞大的数据规模对数据的存储、处理和分析能力提出了很高的要求。

2. 数据类型多样

大模型需要处理的数据类型多种多样，涵盖了文本、图像、视频、音频等多模态数据。不同类型的数据蕴含着不同的信息和特征，需要采用不同的处理和分析方法。这对数据分析的灵活性和适应性提出了更高要求。

3. 数据质量要求高

大模型的性能在很大程度上依赖数据的质量。高质量的数据能够提高模型的准确性和泛化能力。因此，在数据分析过程中，需要对数据进行严格的清洗、去噪和预处理，以确保数据的准确性和可靠性。

4. 分析技术复杂

为了从海量、多样的数据中提取出有用的信息和知识，大模型采用的数据分析技术通常较为复杂。这些技术可能涉及深度学习、自然语言处理和计算机视觉等，需要综合运用多种算法和方法。

5. 分析结果具有预测性

通过数据分析，大模型能够揭示数据之间的内在联系和规律，从而对未来的趋势和结果进行预测。这种预测能力对于制定决策、优化资源配置等方面具有重要意义。

5.1.3 数据分析的流程

通常，数据分析包括以下几个阶段。

1. 数据收集

数据收集即收集需要分析的数据，这些数据既包括结构化数据，如数据库中的表格数据、电子表格中的数值数据等，又包括非结构化数据，如文本、图像、视频、音频等多媒体数据。在收集数据的过程中，还需要注意数据的时效性和合法性，确保数据的新鲜度和合规性。

2. 数据清洗

在收集到的原始数据中，往往存在着缺失值、异常值、重复值和格式不一致等问题。这些问题如果不及时处理，会对数据分析结果产生影响。因此，需要对数据进行清理、转换和整理，以确保数据的准确性、完整性和一致性。数据清洗的具体操作可能包括删除或填补缺失值、替换或删除异常值、合并或删除重复数据，以及统一数据格式等。

3. 数据分析

在数据分析阶段，需要使用统计学、机器学习等方法对数据进行分析和探索。通过数据分析，人们可以发现数据中的模式、趋势和关联，揭示出数据背后隐藏的规律和信息。数据分析的方法多种多样，包括描述性统计、推断性统计、聚类分析、回归分析和关联规则挖掘等。

4. 数据解释

数据分析的结果往往是以数学模型、算法或图表等形式呈现的，这些结果对于非专业人士来说可能难以理解。因此，需要将数据分析的结果转换为可理解的形式，并解释其意义和影响。数据解释的过程需要结合业务背景和实际需求，提出有针对性的见解和建议。数据解释可以帮助决策者更好地理解数据分析结果，从而做出更明智的决策。

5. 数据展示

将数据分析的结果以图表、报告等形式进行展示。数据展示是数据分析成果的最终呈现方式，它直接影响数据分析结果的传播和应用效果。因此，需要注重数据展示的可视化效果和易读性，确保数据展示能够清晰地传达数据的意义和价值。同时，还可以根据实际需求，将数据展示结果以不同的形式（如 PPT、在线报告、交互式仪表板等）进行分享和传播，以便更好地沟通和传达数据分析的成果。

5.1.4　数据分析的应用

在大模型技术中，数据分析的应用非常广泛，涵盖了多个行业和领域。

1. 金融领域

（1）风险评估：大模型通过分析大量的金融交易数据、客户信用记录等信息，可以精准地评估贷款、投资等金融活动的风险。

（2）市场趋势预测：通过分析市场数据、经济指标等，大模型可以帮助金融机构预测市场趋势，制订投资策略。

2. 智能制造

（1）质量检测：大模型可以对生产过程中的产品进行实时质量检测，及时发现并纠正质量问题。

（2）预测性维护：通过分析设备的运行数据、维修记录等，大模型可以预测设备的故障风险，提前进行维护，避免生产中断。

3. 零售与电商

（1）消费者行为分析：大模型可以分析消费者的购买历史、浏览记录等数据，了解消费者的偏好和需求，为个性化推荐和精准营销提供依据。

（2）库存管理：通过分析销售数据、库存数据等，大模型可以帮助企业优化库存管理策略，降低库存成本，提高资金周转率。

4. 医疗领域

（1）医学影像分析：大模型能够对医学影像进行自动分析和诊断，辅助医生发现病灶、判断病情。

（2）疾病预测与预防：通过分析患者的遗传信息、生活习惯和体检数据等，大模型可以预测患者患病的风险，并提出预防措施。

5. 教育领域

（1）智能辅导：大模型可以根据学生的学习进度和能力，提供个性化的学习资源和建议，帮助学生提高学习效果。

（2）学情分析：通过分析学生的学习数据、作业完成情况等，大模型可以帮助教师了解学生的学习情况，制订针对性的教学计划。

6. 智慧城市

（1）交通流量预测：通过分析历史交通数据、天气信息和节假日等因素，大模型可以预测未来的交通流量，为交通管理部门提供决策支持。

（2）智能红绿灯控制：大模型可以根据实时交通情况，智能调整红绿灯的配时，缓解交

通拥堵。

综上所述，数据分析在大模型技术中的应用非常广泛，涵盖了金融、制造、零售、医疗、教育、智慧城市和媒体娱乐等多个领域。通过深度挖掘和分析数据中的价值信息，大模型能够为企业和机构提供精准的决策支持和创新服务。

【例题 5-1】 列举 3 个数据分析在大模型技术中应用的例子。

【解】 (1) 金融风控。

① 信贷审批：大模型通过对借款人的历史还款记录、收入状况、信用评分等多维度数据进行分析，可以快速、准确地评估借款人的信用风险，辅助金融机构做出信贷审批决策。

② 反洗钱监测：利用大模型对交易数据进行实时监测和分析，能够识别出异常交易模式，及时发现并报告可疑的洗钱行为，保障金融系统的安全稳定。

(2) 智能制造的预测性维护。

① 设备健康监测：大模型通过对设备的运行数据、维护记录、故障历史等多维度数据进行分析，可以预测设备的健康状况，提前发现潜在的故障风险。

② 维护计划优化：基于设备健康监测的结果，大模型可以自动生成维护计划，包括维护时间、维护内容、所需备件等，帮助企业合理安排维护资源，降低维护成本。

(3) 教育领域的智能评估。

① 学生能力评估：大模型可以通过分析学生的作业完成情况、考试成绩、课堂表现等多维度数据，对学生的学习能力进行客观、全面的评估。

② 教学效果反馈：通过对学生的学习数据进行分析，大模型还可以为教师提供教学效果反馈，帮助教师了解教学情况，及时调整教学策略。

这些例子展示了数据分析在大模型技术中的广泛应用。通过深度挖掘和分析数据中的价值信息，大模型能够为企业和机构提供精准、高效的决策支持和创新服务。

5.2 数据处理方法

数据处理是数据分析流程中的核心环节，它涉及对原始数据进行一系列的操作和转换，以确保数据的准确性、一致性和完整性，从而支持后续的分析和决策过程。数据处理主要包括数据预处理、数据选择、数值操作、数值运算、数据分组和时间序列分析等。

5.2.1 数据预处理

数据预处理是数据分析前的必要准备阶段，其目的是清洗和整理数据，使其适合进一步的分析。数据预处理主要包括缺失值填充、重复值删除、异常值删除或替换等。

缺失值是数据中常见的现象，可能由于记录错误、设备故障或人为疏忽等原因造成。缺失值的处理方法包括删除含有缺失值的记录(在数据量大且缺失值比例较小时可行)、使用均值/中位数/众数等统计值进行填充，或采用插值法、机器学习算法等更复杂的方法进行预测填充。

重复值是指数据集中完全相同的记录，它们可能在数据录入错误时或数据合并过程中产生。重复值的存在会扭曲数据的分布，影响分析结果的准确性。因此，在数据分析前，需要仔细检查数据集是否存在重复值，并予以删除。

异常值是指数据中明显偏离其他数据点的值，可能由于测量错误、数据录入错误或极端事件等原因造成。异常值的存在会对统计分析和机器学习模型的性能产生负面影响，如拉高或拉低平均值、影响回归模型的拟合效果等。处理异常值的方法包括删除异常值、用统计值（如中位数）替换异常值、根据业务逻辑进行特殊处理等。在处理异常值时，需要谨慎判断，以免误删或误替换包含重要信息的数据点。

【例题 5-2】 基于提供的图书管理数据集，分别执行以下操作，并以表格形式返回大模型工具的处理结果。

（1）删除数据集中有重复的记录行；

（2）找出数据集中缺失值，并使用这一列的均值（整数）进行填充；

（3）将库存量大于 30 的值视作异常值并将其替换为 30。

图书管理数据集如下：

图书 ID	书名	作者	出版社	出版日期	定价	库存量
20120001	计算机导论	蔡平	电子工业出版社	2012.8	35	8
20120002	计算思维导论	陈国良	高等教育出版社	2012.10	29	15
20120003	计算思维导论	唐培和	广西师范大学出版社	2012.10	45	12
20220001	大学计算机	王移芝	高等教育出版社	2022.9	43	33
20120001	计算机导论	蔡平	电子工业出版社	2012.8	35	8
20240001	计算机科学导论	战德臣	高等教育出版社	2024.8	82	
20200001	人工智能导论	廉师友	清华大学出版社	2020.10	69.8	
20220002	人工智能通识教程	王万良	清华大学出版社	2022.6	59.8	22
20240002	人工智能引论	吴飞	高等教育出版社	2024.4	56	38
20240003	人工智能导论	莫宏伟	人民邮电出版社	2024.1	69.8	10

【解】 模型回复分别如图 5-1、图 5-2 和图 5-3 所示。

（1）删除重复行后的数据集如图 5-1 所示。

(1) 删除重复行后的数据集

(删除图书ID为20120001的重复行)

图书ID	书名	作者	出版社	出版日期	定价	库存量
20120001	计算机导论	蔡平	电子工业出版社	2012.8	35	8
20120002	计算思维导论	陈国良	高等教育出版社	2012.10	29	15
20120003	计算思维导论	唐培和	广西师范大学出版社	2012.10	45	12
20220001	大学计算机	王移芝	高等教育出版社	2022.9	43	33
20240001	计算机科学导论	战德臣	高等教育出版社	2024.8	82	(缺失)
20200001	人工智能导论	廉师友	清华大学出版社	2020.10	69.8	(缺失)
20220002	人工智能通识教程	王万良	清华大学出版社	2022.6	59.8	22
20240002	人工智能引论	吴飞	高等教育出版社	2024.4	56	38
20240003	人工智能导论	莫宏伟	人民邮电出版社	2024.1	69.8	10

图 5-1　删除重复行后的数据集

（2）填充缺失值后的数据集（均值整数填充）如图 5-2 所示。

(2) 填充缺失值后的数据集（均值整数填充）

- 定价列均值计算（不含缺失值）：(35 + 29 + 45 + 43 + 82 + 69.8 + 59.8) / 7 ≈ 54.86（取整数55）
- 库存量列均值计算（不含缺失值）：(8 + 15 + 12 + 33 + 22 + 38 + 10) / 7 ≈ 19.71（取整数20）
- 填充缺失值：

图书ID	书名	作者	出版社	出版日期	定价	库存量
20120001	计算机导论	蔡平	电子工业出版社	2012.8	35	8
20120002	计算思维导论	陈国良	高等教育出版社	2012.10	29	15
20120003	计算思维导论	唐培和	广西师范大学出版社	2012.10	45	12
20220001	大学计算机	王移芝	高等教育出版社	2022.9	43	33
20240001	计算机科学导论	战德臣	高等教育出版社	2024.8	55	20
20200001	人工智能导论	廉师友	清华大学出版社	2020.10	69.8	20
20220002	人工智能通识教程	王万良	清华大学出版社	2022.6	59.8	22
20240002	人工智能引论	吴飞	高等教育出版社	2024.4	56	38
20240003	人工智能导论	莫宏伟	人民邮电出版社	2024.1	69.8	10

图 5-2　填充缺失值后的数据集

（3）替换异常值后的数据集（库存量＞30 的值视作异常值并替换为 30）如图 5-3 所示。

(3) 替换异常值后的数据集（库存量>30的值视作异常值并替换为30）

图书ID	书名	作者	出版社	出版日期	定价	库存量
20120001	计算机导论	蔡平	电子工业出版社	2012.8	35	8
20120002	计算思维导论	陈国良	高等教育出版社	2012.10	29	15
20120003	计算思维导论	唐培和	广西师范大学出版社	2012.10	45	12
20220001	大学计算机	王移芝	高等教育出版社	2022.9	43	30
20240001	计算机科学导论	战德臣	高等教育出版社	2024.8	55	20
20200001	人工智能导论	廉师友	清华大学出版社	2020.10	69.8	20
20220002	人工智能通识教程	王万良	清华大学出版社	2022.6	59.8	22
20240002	人工智能引论	吴飞	高等教育出版社	2024.4	56	30
20240003	人工智能导论	莫宏伟	人民邮电出版社	2024.1	69.8	10

图 5-3　替换异常值后的数据集

5.2.2　数据选择

数据选择主要是根据实际需求和分析目标，从数据集中选取特定的记录行或列。通过数据选择，可以减少数据的维度，聚焦关键信息，提高数据分析的效率和准确性。本节以上述例题的结果数据为例，进行数据选择。

【例题 5-3】　基于 5.2.1 节处理后的图书管理数据集（如图 5-3 所示），将出版社为“清华大学出版社”的全部图书筛选出来，并以表格形式返回大模型工具的处理结果。

【解】　模型回复如图 5-4 所示。

根据您提供的图书管理数据集，以下是筛选出出版社为"清华大学出版社"的全部图书，并以表格形式返回的处理结果：

图书ID	书名	作者	出版社	出版日期	定价	库存量
20200001	人工智能导论	廉师友	清华大学出版社	2020.10	69.8	20
20220002	人工智能通识教程	王万良	清华大学出版社	2022.6	59.8	22

这就是出版社为"清华大学出版社"的全部图书信息。

图 5-4　大模型工具按出版社筛选后的结果

【例题 5-4】　基于 5.2.1 节处理后的图书管理数据集（如图 5-3 所示），选择库存量大于 10 且定价低于 50 元的图书，并筛选出这些书的“书名”“作者”“库存量”信息，并以表格形式返回大模型工具的处理结果。

【解】　模型回复如图 5-5 所示。

根据您提供的图书管理数据集，以下是筛选出库存量大于10且定价低于50的图书的"书名" "作者"和"库存量"，并以表格形式返回的处理结果：

书名	作者	库存量
计算思维导论	陈国良	15
计算思维导论	唐培和	12

这两本图书满足库存量大于10且定价低于50的条件。

图 5-5　大模型工具按库存量和定价筛选后的结果

5.2.3　数值操作

数值操作主要是对数据进行替换、排序等。本节仍以之前的数据为例，进行数值操作。

【例题 5-5】　基于 5.2.1 节处理后的图书管理数据集（如图 5-3 所示），依次执行以下操作，并以表格形式返回大模型工具的处理最终结果。

（1）将库存量小于 12 的值替换成 20；

（2）将数据集记录按照库存量进行降序排列。

【解】 模型回复如图 5-6 所示。

根据您提供的图书管理数据集，依次执行以下操作后的最终处理结果如下：

(1) 将库存量小于12的值替换成20；

(2) 将数据集记录按照库存量进行降序排列。

处理后的数据集表格如下：

图书ID	书名	作者	出版社	出版日期	定价	库存量
20220001	大学计算机	王移芝	高等教育出版社	2022.9	43	30
20240002	人工智能引论	吴飞	高等教育出版社	2024.4	56	30
20220002	人工智能通识教程	王万良	清华大学出版社	2022.6	59.8	22
20240001	计算机科学导论	战德臣	高等教育出版社	2024.8	82	20
20200001	人工智能导论	廉师友	清华大学出版社	2020.10	69.8	20
20120002	计算思维导论	陈国良	高等教育出版社	2012.10	29	15
20120003	计算思维导论	唐培和	广西师范大学出版社	2012.10	45	12
20240003	人工智能导论	莫宏伟	人民邮电出版社	2024.1	69.8	20
20120001	计算机导论	蔡平	电子工业出版社	2012.8	35	20

图 5-6 大模型工具进行替换和排序操作后的结果

读者可以思考，若将库存量为 20 的数据记录参与排序，应该如何修改提示词（参考思维链提示）。

5.2.4 数值运算

数值运算主要是对数据进行算术运算、比较运算和汇总运算等，这些运算能够帮助用户从原始数据中提取出有价值的信息，为后续的数据分析和决策提供支持。本节仍以之前的数据为例进行数值运算。

【例题 5-6】 基于 5.2.1 节处理后的图书管理数据集（如图 5-3 所示），依次执行以下操作，并以表格形式返回大模型工具的处理结果。

（1）将“定价”和“库存量”进行相乘，用相乘结果生成“小计”列，并删除“出版日期”列；

（2）对数据记录进行计数，并对“小计”列求和，返回计算结果。

【解】 模型回复如图 5-7 和图 5-8 所示。

5.2.5 数据分组

数据分组是按照某些维度先对数据进行分组，再对分组后的数据进行汇总运算等操作。本节仍以之前的数据为例进行数据分组。

【例题 5-7】 基于 5.2.1 节处理后的图书管理数据集（如图 5-3 所示），依次执行以下操作，并以表格形式返回大模型工具的处理结果。

根据您的要求，以下是对图书管理数据集进行处理后的结果：

处理步骤：

1. **将“定价”和“库存量”进行相乘，生成“小计”列，并删除“出版日期”列。**

图书ID	书名	作者	出版社	定价	库存量	小计
20120001	计算机导论	蔡平	电子工业出版社	35	8	280
20120002	计算思维导论	陈国良	高等教育出版社	29	15	435
20120003	计算思维导论	唐培和	广西师范大学出版社	45	12	540
20220001	大学计算机	王移芝	高等教育出版社	43	30	1290
20240001	计算机科学导论	战德臣	高等教育出版社	82	20	1640
20200001	人工智能导论	廉师友	清华大学出版社	69.8	20	1396
20220002	人工智能通识教程	王万良	清华大学出版社	59.8	22	1315.6
20240002	人工智能引论	吴飞	高等教育出版社	56	30	1680
20240003	人工智能导论	莫宏伟	人民邮电出版社	69.8	10	698

图 5-7　大模型工具完成增加列、删除列操作后的结果

2. **对数据记录进行计数，并对“小计”列求和。**

- 数据记录计数：9条
- “小计”列求和：9274.6

以上就是按照您的要求处理后的图书管理数据集及相关的计算结果。

图 5-8　大模型工具完成计数和求和操作后的结果

（1）以出版社名称进行分组，返回除高等教育出版社以外的分组数据；

（2）统计各出版社出版图书的数量。

【解】　模型回复如图 5-9 和图 5-10 所示。

5.2.6　时间序列分析

时间序列分析主要有时间格式互换、时间索引和时间运算等。本节仍以之前的数据为例进行时间序列分析。

【例题 5-8】　基于 5.2.1 节处理后的图书管理数据集（如图 5-3 所示），依次执行以下操作，并以表格形式返回大模型工具的处理结果。

（1）删除“定价”“库存量”两列，同时删除高等教育出版社出版的图书记录；

（2）提取“出版日期”中的“月”部分，存储在“月份”新列中；

（3）将清华大学出版社所出版图书的出版日期推迟 30 天，存储在“备注”新列中。

【解】　模型回复如图 5-11、图 5-12 和图 5-13 所示。