

所谓数据挖掘，就是在众多繁杂的数据中通过系统的方法来找寻有意义的信息。自从 20 世纪 90 年代数据挖掘概念被提出以来，其发展到现在可以说是最具有代表性的技术之一。本章主要讲述数据挖掘基本流程中的第一步，即如何收集数据并进行管理、常见的数据挖掘软件有哪些、不同软件平台的特点及适用性等。

1.1 数据收集及管理

1.1.1 数据的类型

数据指的是能够通过观察、计算或者做实验得出来的结果，也就是数值，按照不同的分类法可以得到不同的数据类型。通常可进行如下分类。

1. 按观测结果的特征分类

1) 定性数据

定性数据 (qualitative data) 是用来描述事物的性质及其类别的一组文字表述型数据，因而不能将其量化，只能将其定性。换句话说，定性数据就是诸如性别、国家、动植物分类等数据。例如，现在将某个班级的学生按照户籍分为城市、乡镇两类，那么在这里，户籍就构成了一个定性变量。

如户籍：城市，乡镇，乡镇，城市，城市，城市，乡镇，城市，……，乡镇，乡镇，城市，乡镇，城市，乡镇。

2) 定量数据

与定性数据用来描述事物的性质和类别不同，定量数据 (quantitative data) 指的是以数量的形式而存在的数据，可以对其进行衡量等操作。定量数据类似于平常所说的年龄、身高、体重、时间、质量、距离等数值。例如，现在统计某个班级里每个学生的年龄，那么年龄在这里就成为一个定量变量。

如年龄 (岁)：17, 18, 17, 18, 18, 19, 16, 18, 16, 17, 19, ……，16, 15, 18, 17, 16, 19, 18, 17, 17。

2. 按数据收集方法分类

1) 观测数据

观测数据 (observational data) 指的是可以通过观察以及调查而直接得到的数据，所以其又被称为原始数据。在这里举个小例子来说明何为观测数据：现在给某个班级的同学每人分发一份调查问卷，共取得 35 份问卷，那么收集到的这些问卷数据就成为几组观

测数据。

2) 实验数据

实验数据 (experimental data) 指的是在实验中通过对相关变量, 如实验时间、过程、对象等进行有效的控制而得到的数据。比如, 对一起栽培的植物施加不同的光照条件, 以检验不同的光照条件对植物生长的影响, 所得到的结果就是实验数据。

3. 按描述的现象与时间关系分类

1) 截面数据

截面数据 (cross-sectional data) 是指在同—一个时间点或者时间段上不同变量的数据的集合, 是静态数据, 又被称为横截面数据, 其反映了某一个事物或者现象在某个时间点上的变化情况。例如, 2020 年我国各个省份的 GDP (国内生产总值), 这些数据就是一个截面数据:

省份	广东	江苏	山东	……	海南	宁夏	青海	西藏
GDP/亿元	110 760.94	102 719	73 129	……	5 532.39	3 920.55	3 005.92	1 902.74

2) 时间序列数据

时间序列数据 (time series data) 是指在不同时间点上按照一定的时间间隔收集来的数据, 其反映了某一个事物或者现象随着时间变化而变化的情况, 所以又被称为动态数据。例如, 现在收集海南 2010—2020 年的 GDP 数据, 那么收集而来的这些数据就是一个时间序列数据:

时间	2010 年	2011 年	2012 年	……	2017 年	2018 年	2019 年	2020 年
GDP/亿元	2 064.50	2 522.66	2 855.54	……	4 462.54	4 832.05	5 308.93	5 532.39

1.1.2 数据的收集

在平常的工作和生活中, 人们难免会跟各式各样的数据打交道, 在打交道的过程中有时候需要根据数据的类型不同来作出对应的判断和决定。为了让判断和决定更加有效和准确, 人们就开始进行有目的的数据收集, 根据数据来挖掘其相关的信息。数据收集指的是按照自身需求来收集相关的数据。在准备进行数据收集的时候, 往往需要进行一系列的工作, 其步骤如下。

- (1) 确定调查的对象, 明确调查的目的。
- (2) 根据调查对象不同选择相对应的调查方法。
- (3) 收集数据。
- (4) 用合适的方式整理好收集来的数据。
- (5) 分析数据, 得出结果。

不过需要注意的是, 在进行数据收集的时候, 一定要确保采样得来的数据具有代表性, 能够代表总体, 同时样本的分布也要较为均匀。除此之外, 数据的收集还需要有一定的格式, 如基础格式、文字标签等, 下面简单介绍这两种格式。

原始数据格式是人们平时最常见的基础格式, 如图 1-1 所示。

	A	B	C	D	E	F	G	H	I	J
1	员工编号	性别	学历	年龄	本司工作年限	英语水平	身高	体重	年收入	所在部门
2	1	2	3	23	1	2	158	65	1	2
3	2	1	2	34	3	1	168	75	2	5
4	3	2	4	32	2	2	174	65	2	4
5	4	1	2	37	3	1	185	78	3	4

图 1-1 原始数据格式

这种原始数据的特点是：一行代表一个样本，一列代表一个样本属性，有多少个样本就需要录入多少行数据。

而文字标签则是让结果显示成具体的标签而不是数字，如图 1-2 所示。

	A	B	C	D	E
1	标题	数字	标签		
2	性别	1	男		
3		2	女		
4	学历	2	大学本科		
5		3	硕士研究生		
6		4	博士研究生		
7	本司工作年限	1	一年以下		
8		2	一年至三年		
9		3	三年至五年		
10	英语水平	1	四级		
11		2	六级		
12	年收入	1	10万元以下		
13		2	10万~20万元		
14		3	20万~30万元		
15	所在部门	2	销售部		
16		4	人事部		
17		5	财务部		

图 1-2 文字标签格式

【例 1-1】调查数据

为了了解某企业 50 名员工的基本状况，现在收集了这些员工的 10 项指标：

员工编号：定性变量，可以简单记为工号或者 ID（身份标识号）。

员工性别：定性变量，简单记为性别，或者记为 gender。

员工学历：定性变量，简单记为学历，或者记为 education。

员工年龄：定量变量，单位为岁，简单记为年龄，或者记为 age。

员工在公司的工作年限：定量变量，单位为年，简单记为工作年限，或者记为 year。

员工的英语水平：定性变量，简单记为英语水平，或者记为 level。

员工的身高：定量变量，单位为 cm，简单记为身高，或者记为 height。

员工的体重：定量变量，单位为 kg，简单记为体重，或者记为 weight。

员工的年收入：定量变量，单位为万元，简单记为年收入，或者记为 income。

员工的部门：定性变量，简单记为部门，或者记为 department。

数据由其变量和属性组成。在本例子中共有 10 个变量：工号、性别、学历、年龄、工作年限、英语水平、身高、体重、年收入、部门。

表 1-1 是该企业 50 名员工的个人基本情况，按照基础数据格式，每一行代表一个对象，也就是样本，每一列就是该列所在的行样本的属性。这样就可以得到表 1-1 的数据集。

表 1-1 50 名员工的个人基本信息

工号	性别	学历	年龄	工作年限	英语水平	身高	体重	年收入	部门
10248008	女	硕士研究生	33	一年以下	专业八级	158	65	26.0	销售部
10229019	男	大学本科	36	一年至三年	专业四级	174	75	10.4	销售部
12108019	女	大学本科	36	三年至五年	专业四级	167	65	21.0	财务部
12332010	男	博士研究生	31	三年至五年	四级	171	75	24.9	人事部
12331015	男	大学本科	25	一年以下	四级	175	74	25.9	后勤部
16248014	男	硕士研究生	26	三年至五年	六级	178	67	18.6	销售部
16352030	女	硕士研究生	31	三年至五年	六级	164	56	9.1	销售部
16171019	女	博士研究生	37	三年至五年	四级	165	50	12.5	后勤部
16391008	女	硕士研究生	27	一年以下	四级	158	55	35.6	后勤部
20395019	男	大学本科	29	一年以下	专业四级	170	71	22.8	财务部
20100029	男	大学本科	30	三年至五年	六级	182	71	10.3	销售部
20324035	男	大学本科	26	一年以下	六级	172	76	13.0	销售部
22186005	男	大学本科	33	三年至五年	四级	174	80	9.8	后勤部
22160006	女	硕士研究生	33	五年以上	专业四级	159	52	35.3	人事部
22274026	女	硕士研究生	38	三年至五年	六级	167	45	20.5	销售部
23376027	男	硕士研究生	41	三年至五年	四级	182	64	34.1	销售部
23368030	女	大学本科	42	三年至五年	专业四级	161	62	20.6	财务部
24225006	男	大学本科	36	五年以上	四级	168	68	15.8	财务部
24105026	女	大学本科	29	三年至五年	四级	165	53	29.4	销售部
24286013	男	硕士研究生	31	五年以上	六级	171	76	14.8	销售部
25235027	女	博士研究生	39	三年至五年	六级	164	54	8.2	销售部
25352033	男	硕士研究生	35	三年至五年	四级	178	74	11.1	销售部
26177005	男	大学本科	42	五年以上	四级	184	79	15.8	人事部
26196010	男	硕士研究生	41	五年以上	六级	177	68	19.8	后勤部
27173011	女	硕士研究生	30	三年至五年	专业四级	162	60	11.5	销售部
27237032	女	硕士研究生	29	三年至五年	六级	160	53	19.4	销售部
27289024	男	博士研究生	28	一年以下	六级	173	75	10.8	销售部
29107020	男	硕士研究生	28	三年至五年	专业四级	182	75	22.9	销售部
29365032	男	大学本科	34	五年以上	四级	176	74	10.4	人事部
30273031	男	大学本科	37	三年至五年	六级	179	68	35.6	财务部
30243029	男	硕士研究生	25	一年以下	专业四级	169	53	9.5	人事部
31364037	女	大学本科	28	三年至五年	四级	162	56	7.3	销售部
31316038	女	大学本科	26	三年至五年	六级	159	45	22.8	销售部
32304031	女	博士研究生	30	三年至五年	六级	167	63	17.9	财务部
32208040	男	博士研究生	34	三年至五年	专业四级	181	78	35.5	人事部
32292012	男	大学本科	34	三年至五年	六级	170	71	28.4	财务部
32185004	女	大学本科	27	一年以下	四级	158	55	13.4	销售部
33219013	女	大学本科	31	三年至五年	六级	163	49	11.1	销售部

续表

工号	性别	学历	年龄	工作年限	英语水平	身高	体重	年收入	部门
33384028	男	硕士研究生	40	五年以上	专业八级	174	72	8.1	人事部
33172017	女	大学本科	27	三年至五年	四级	161	53	27.2	销售部
37288004	女	大学本科	39	三年至五年	六级	166	48	19.1	销售部
37359035	女	博士研究生	28	一年以下	专业四级	160	54	17.6	财务部
38391022	女	硕士研究生	34	五年以上	六级	159	45	10.3	财务部
38399025	男	大学本科	42	五年以上	四级	173	57	9.5	人事部
38120022	男	大学本科	27	一年以下	专业四级	162	63	35.6	人事部
38319004	男	硕士研究生	32	三年至五年	专业八级	176	55	14.4	销售部
38254010	女	大学本科	35	五年以上	四级	165	52	25.3	销售部
40294017	女	大学本科	32	三年至五年	专业八级	168	50	21.4	人事部
40365026	女	硕士研究生	30	三年至五年	六级	166	62	15.5	销售部
40388036	女	博士研究生	32	三年至五年	四级	164	51	36.8	销售部

1.1.3 数据的管理

当人们将收集的数据整理好以便分析的时候，就不得不面临一个问题：要怎样保存这些数据以便日后访问呢？这时候就需要对收集的数据进行管理工作，也就是数据管理。

数据管理在计算机应用领域当中是一个非常重要的模块，它指的是针对不同的数据进行收集、整理、组织、存储、加工、传输和检索的各个过程，目的是能够更加充分且有效地发挥数据的作用。随着技术的进步，数据管理经历了从人工管理到文件系统，再到数据库系统这三个发展阶段。如今市场上可以用于数据管理的软件很多，不过对于一般的数据分析来说，电子表格已经足以胜任其所需的数据管理工作了。

1. 电子表格管理数据

如果要对数据量不是很大的数据集进行管理，并且不要求多人同时访问，则可采用电子表格来管理数据。目前市面上常用的电子表格有 Excel 和 WPS，这两款软件各有千秋，可以根据自身习惯和喜好来选择使用哪款来管理数据。图 1-3 是电子表格管理数据的界面。

本书将例 1-1 收集的数据存放在 BaPy_data.xlsx 文件的 BAdata 工作表中。

2. 数据库管理数据

当数据量比较大的时候，使用电子表格软件来进行管理就比较困难了，面对这种情况，可以采用数据库来管理这些数据。数据库是一个仓库，能够按照一定的规则来保存大量的数据。目前常用的数据库管理软件有 Navicat、DBeaver、Workbench、SQLyog 以及 phpMyAdmin 等，这些软件可以说是各有特色，当需要用到数据库来管理数据的时候，可以按需选择来使用。

	A	B	C	D	E	F	G	H	I	J
1	员工编号	性别	学历	年龄	本司工作年限	英语水平	身高	体重	年收入	所在部门
2	10248008	女	硕士研究生	33	一年以下	专业八级	158	65	26.0	销售部
3	10229019	男	大学本科	36	一年至三年	专业四级	174	75	10.4	销售部
4	12108019	女	大学本科	36	三年至五年	专业四级	167	65	21.0	财务部
5	12332010	男	博士研究生	31	三年至五年	四级	171	75	24.9	人事部
6	12331015	男	大学本科	25	一年以下	四级	175	74	25.9	后勤部
7	16248014	男	硕士研究生	26	三年至五年	六级	178	67	18.6	销售部
8	16352030	女	硕士研究生	31	三年至五年	六级	164	56	9.1	销售部
9	16171019	女	博士研究生	37	三年至五年	四级	165	50	12.5	后勤部
10	16391008	女	硕士研究生	27	一年以下	四级	158	55	35.6	后勤部
11	20395019	男	大学本科	29	一年以下	专业四级	170	71	22.8	财务部
12	20100029	男	大学本科	30	三年至五年	六级	182	71	10.3	销售部
13	20324035	男	大学本科	26	一年以下	六级	172	76	13.0	销售部
14	22186005	男	大学本科	33	三年至五年	四级	174	80	9.8	后勤部
15	22160006	女	硕士研究生	33	五年以上	专业四级	159	52	35.3	人事部
16	22274026	女	硕士研究生	38	三年至五年	六级	167	45	20.5	销售部
17	23376027	男	硕士研究生	41	三年至五年	四级	182	64	34.1	销售部
18	23368030	女	大学本科	42	三年至五年	专业四级	161	62	20.6	财务部
19	24225006	男	大学本科	36	五年以上	四级	168	68	15.8	财务部
20	24105026	女	大学本科	29	三年至五年	四级	165	53	29.4	销售部
21	24286013	男	硕士研究生	31	五年以上	六级	171	76	14.8	销售部
22	25235027	女	博士研究生	39	三年至五年	六级	164	54	8.2	销售部
23	25352033	男	硕士研究生	35	三年至五年	四级	178	74	11.1	销售部
24	26177005	男	大学本科	42	五年以上	四级	184	79	15.8	人事部
25	26196010	男	硕士研究生	41	五年以上	六级	177	68	19.8	后勤部
26	27173011	女	硕士研究生	30	三年至五年	专业四级	162	60	11.5	销售部
27	27237032	女	硕士研究生	29	三年至五年	六级	160	53	19.4	销售部
28	27289024	男	博士研究生	28	一年以下	六级	173	75	10.8	销售部
29	29107020	男	硕士研究生	28	三年至五年	专业四级	182	75	22.9	销售部
30	29314037	男	硕士研究生	40	五年以上	四级	168	69	15.1	销售部

图 1-3 电子表格管理数据的界面

1.2 数据挖掘软件

1.2.1 常见统计分析软件

数据分析指的是对数据进行控制、处理、整理、分析的过程。目前市面上的统计分析软件比较多,如 Excel、SAS(Statistical Analysis System)、SPSS(Statistical Product and Service Solutions)、R 语言、EViews (Econometrics Views)、Stata 等,下面本节简单地介绍常见的统计分析软件。

(1) Excel。Excel 是由微软公司所开发的一款电子表格软件,与金山的 WPS 电子表格一样,Excel 作为工作中常用的办公软件,既是进行数据管理的工具,也是进行数据分析的基本工具。Excel 具有许多强大的功能,比如合并计算、快速填充和可视化功能等,这极大地提高了人们工作时的效率。但是 Excel 也有一些缺点,比如不能很好地处理大型的数据集,统计方法也不太全面。

(2) SAS。SAS 是一个模块化、集成化的大型应用软件系统。得益于其广泛使用,SAS 被人们誉为统计分析的标准软件。与此同时,它也是世界上功能最为强大的统计分析软件,有着十分完善的数据分析和管理工作。不过由于 SAS 包含了太多功能,在使用时需要使用者掌握许多知识,所以 SAS 一般更适合专业人士。

(3) SPSS。SPSS 是如今最著名的统计分析软件之一，翻译成中文就是社会学统计软件包的意思。SPSS 具有完整的数据输入、编辑、统计分析、报表、图形制作等功能，广泛地应用于社会科学、自然科学的各个领域。与 SAS 适合专业人士使用不同，SPSS 无论是初学者、熟练者抑或是专业者都适用。

(4) R 语言。R 语言是用于统计分析、绘图的语言和操作环境，来自 S 语言，是 S 语言的一个变种。R 是一个自由、免费、源代码开放的软件，是一个非常优秀的统计分析工具。

(5) EViews。EViews 一般称为计量经济学软件包，是如今最为流行的计量经济学软件之一。EViews 的前身是 1981 年第 1 版的 Micro TSP，与 SPSS 偏向于统计学领域不同，EViews 更偏向于计量经济学领域。EViews 更重视建立计量模型，也就是模型回归估计，是完成计量经济学研究的得力工具。

(6) Stata。Stata 和 SPSS、SAS 并称为当今三大统计软件，它是一套完整、集成的统计分析软件包。Stata 与 SAS 相比不但体积小巧，而且功能也十分强大并且简单易懂。Stata 在新版本中采用了具有亲和力的窗口接口，当使用者创建程序时，Stata 能提供具有直接命令式的语法。同时，Stata 提供完整的使用手册，包含统计样本建立、解释、模型与语法、文献等。

虽然用于数据分析的工具具有许多种，但是其实总结起来万变不离其宗，无非是数据收集、数据管理、数据分析、数据展示等几个方面。由于 Python 功能强大并且语法简单，可拓展性也强，所以本书选择了 Python 作为分析工具。

1.2.2 Python 语言

Python 作为一种面向对象的解释型语言，同时也是一种优雅、简单明了的编程语言。Python 是数据分析的一把利器，作为一门新兴的编程语言，Python 相比其他语言更加容易入门。与此同时，Python 是完全免费开源的，它拥有非常多用于数据分析或者其他领域的优秀的第三方库。由于 Python 自身有许多的优点，这使得它成为如今最受欢迎的程序设计语言之一。目前国内外许多公司都在使用 Python，如 YouTube、Google、阿里云等。

下面本书介绍常见的 Python 数据分析领域相关库。

(1) NumPy (Numerical Python)：NumPy 是 Python 中科学计算领域的一个基础包，可用来存储和处理大型矩阵，支持大量的维度数组与矩阵运算，此外针对数组运算 NumPy 提供了大量的数学函数库。

(2) Matplotlib：Matplotlib 是 Python 的一个 2D (二维) 绘图库，作为 Python 最流行的可视化模块之一，Matplotlib 不但功能十分强大，而且用起来也较为简便。

(3) Pandas：Pandas 是为了解决数据分析任务而创建的，它基于 NumPy，是 Python 的一个数据分析包，同时也是一个完全开源的 Python 库。

(4) SciPy：SciPy 是 Python 的一个开源科学分析包，可以用来进行信号处理、图像处理、常微分方程数值的求解等问题的处理。由于 SciPy 能够有效地计算 NumPy 矩阵，所以在进行数据分析的同时使用 NumPy 和 SciPy 可以更加高效地解决问题。

(5) Statsmodels: Statsmodels 是 Python 的一个软件包,其可用来拟合多种统计模型,执行统计测试以及可视化,为 SciPy 提供了补充。

(6) scikit-learn: scikit-learn 简称 sklearn,其基于 NumPy 和 Matplotlib 库而构建,为 SciPy 的扩展。sklearn 支持包括分类、回归、降维和聚类 (clustering) 四大机器学习算法,同时还包括了模型评估、数据处理和特征提取这三大模块。

相比前面提到过的其他数据分析软件,Python 是综合能力最为强大的一个数据分析工具,下面将 Python 和上述软件进行对比。

(1) Python 是通过调用各种库和模块来处理 Excel 表格的,与 Excel 相比,Python 在处理更大的数据集时效率更高,能够很轻松地实现自动化,同时也能够搭建一些算法模型。

(2) 与 SPSS 相比,由于 SPSS 是个统计软件,所以其不太适合做一些贴近实际应用场景的数据分析,更加适合在科学研究方面做实验数据的分析。而 Python 则不一样,它能够处理复杂的数据逻辑,因此在一些实际的应用场景使用上会比 SPSS 更有优势。

(3) 与 R 语言相比,Python 的速度更快,语法更清晰,更加简单易学,更适合初学编程的人入门学习。

(4) 同时相比上述的这些工具,Python 在大数据分析以及机器学习这些领域的应用中处于领先地位。如今使用 Python 可以实现很多数据科学方面的应用,如数据收集、整理、清洗、可视化、机器学习以及开发等,所以,只使用 Python 就可以做到全套服务。

1.3 习 题

一、填空题

1. 按观测结果的特征分类,数据可分为_____和_____。
2. 观测数据又称_____。
3. 常见的电子表格软件有_____和_____。
4. 数据管理是对不同的数据进行_____、_____、组织、存储、_____、传输和检索的各个过程。
5. Python 是一种面向对象的_____语言。
6. EViews 一般称为_____。
7. scikit-learn 简称为_____。
8. Stata 与_____、_____并称当今三大统计软件。
9. 当数据量不是很大的时候可以使用_____来管理数据。
10. 最常见的基础格式是_____。

二、判断题

1. 截面数据指的是按照一定的时间间隔在不同时间上收集到的数据。()
2. 定性数据是一组文字表述型数据。()

3. Excel 可以处理大型的数据集。()
4. 数据是通过观察、计算或者实验得出来的结果。()
5. R 语言是功能性语言。()
6. 数据管理经历了人工管理、文件系统以及数据库系统这三个发展阶段。()
7. 实验数据指可以直接通过调查或观察收集得来的数据。()
8. SPSS 一般称为计量经济学软件包。()
9. 数据挖掘是指从大量的数据中通过算法搜索隐藏于其中信息的过程。()
10. 原始数据的特点是一行代表一个样本属性，一列就代表一个样本。()
11. 文字标签就是让最后的结果显示成具体的标签而不是数字。()

三、简答题

1. 按照不同的分类法可以得到哪些不同的数据类型?
2. 说出三个常见的统计分析软件。
3. 进行数据收集的时候需要进行哪些工作?
4. 简述 Python 相比其他数据分析软件的优势。
5. 常见的 Python 数据分析领域相关库有哪些?