第一章：数据预处理总述

Recommend
火龙果软件

大数据项目开发流程

数据采集 → 数据预处理 → 数据存储 → 数据分析挖掘 → 数据可视化

目 录

01 ◆数据预处理背景

02 ◆数据预处理目的

03 ◆ 数据预处理的流程

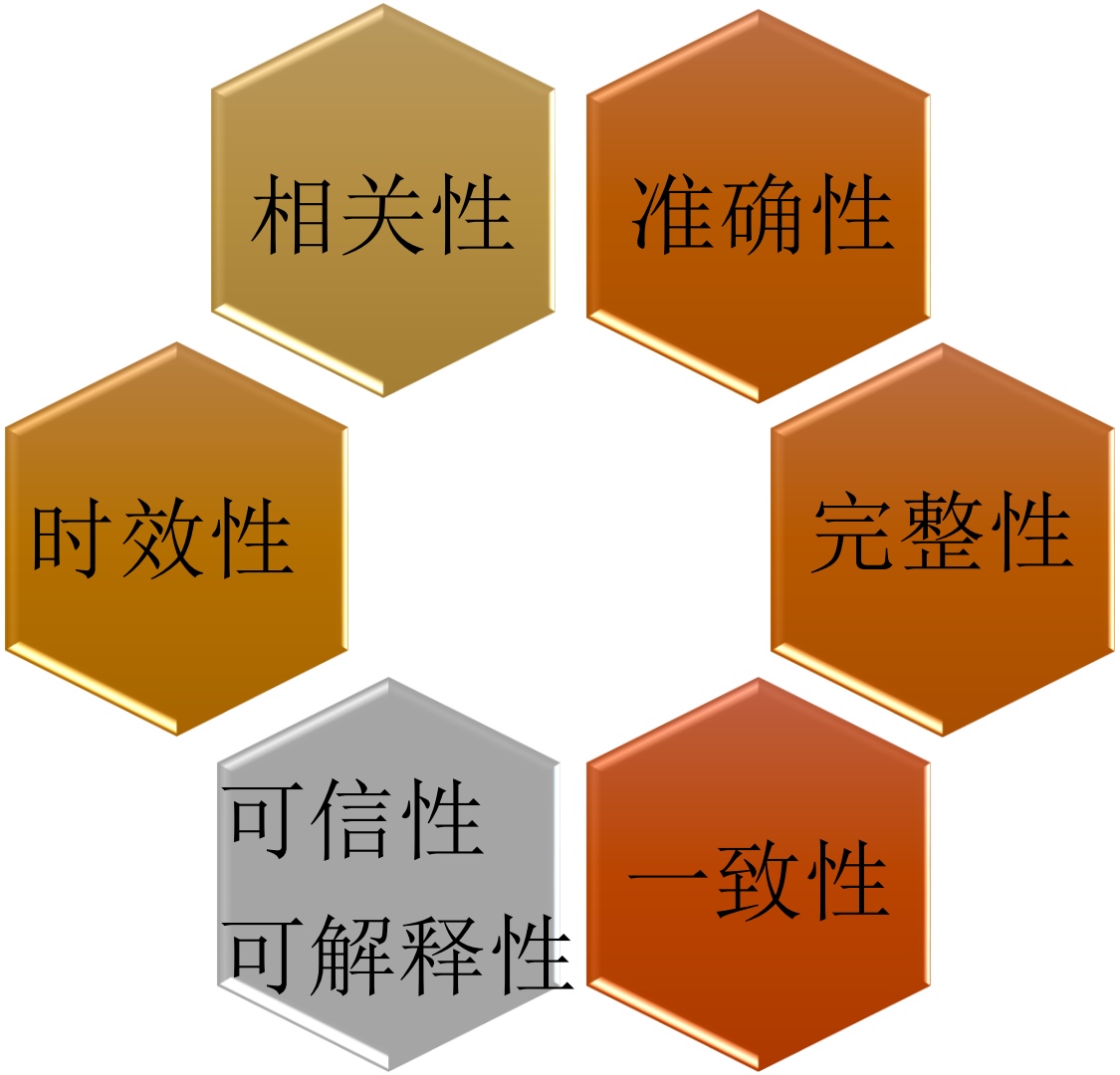
04 ◆ 数据预处理的工具

01

数据预处理背景

--数据的各种质量

数据质量问题



数据质量因素

准确性

- 数据是正确的，数据存储在数据库中的值对应于真实世界的值

造成数据不准确的原因

- 数据收集设备故障
- 数据输入错误
- 数据传输过程出错
- 命名约定、数据代码、输入字段的格式不一致

数据质量因素

完整性

指信息具有一个实体描述的所有必需的部分，在传统关系型数据库中，完整性通常与空值（NULL）有关。一般包括记录的缺失和记录属性的缺失。

造成数据不完整的原因：

- 涉及个人隐私，无法获取相关属性
- 数据输入时，由于人为的疏漏导致
- 数据输入或传输时，由于机器的故障导致

数据质量因素

一致性

在数据库中是指在不同地方存储和使用的同一数据应当是等价的，表示数据有相等的值和相同的含义

- 逻辑不一致

表 1 学生信息表

学号	姓名	性别	年龄	所在专业
95001	张晓云	女	18	M01
95002	刘一天	男	19	M02
95003	邓茹	女	18	M03
95004	王小刚	男	20	M15

表 2 专业信息表

专业号	专业名称	专业班级数	负责人
M01	计算机科学于技术	2	刘莉莉
M02	软件工程	3	朱晓波
M03	信息安全	2	李瑶
M04	通信工程	4	陈杨勇
M05	物联网	3	罗莉

数据质量因素

一致性

在数据库中是指在不同地方存储和使用的同一数据应当是等价的，表示数据有相等的值和相同的含义

- 记录规范不一致

示例

1933/10/7

类型(T):

*2012/3/14
*2012年3月14日
2012-03-14
二〇一二年三月十四日
二〇一二年三月
三月十四日
2012年3月14日

不同数据库对同一属性的命名规则不一致

A数据库 male=1 ， female=2

B数据库 male=‘男’ ， female=‘女’

C数据库 male=‘M’ , female=‘F’

数据质量因素

一致性

在数据库中是指在不同地方存储和使用的同一数据应当是等价的，表示数据有相等的值和相同的含义

- 数据冗余时数据内容不一致

表 3-1 学生信息表

学号	姓名	性别	年龄	所在专业
95001	张晓云	女	18	M01
95002	刘一天	男	19	M02
95003	邓茹	女	18	M03
95004	王小刚	男	20	M15

表 3-2 专业信息表

专业号	专业名称	专业班级数	负责人
M01	计算机科学于技术	2	刘莉莉
M02	软件工程	3	朱晓波
M03	信息安全	2	李瑶
M04	通信工程	4	陈杨勇
M05	物联网	3	罗莉

数据质量因素

相关性

数据的相关性是指数据与特定的应用和领域有关。

数据相关性的应用场景：

- 构造预测模型时，需要采集与模型相关的数据
- 相同的数据，在不同的应用领域中，相关性也是不一样的。

数据质量因素

时效性

是指数据仅在一定时间段内对决策具有价值的属性。
数据的时效性很大程度上制约着决策的客观效果。

数据时效性的应用场景：

- 商品推荐
- 城市交通

数据质量因素

可信性

数据来源的权威性、
数据的规范性、数据产生的时间

可解释性

反映数据是否容易理解

02

数据预处理目的

--提升数据质量

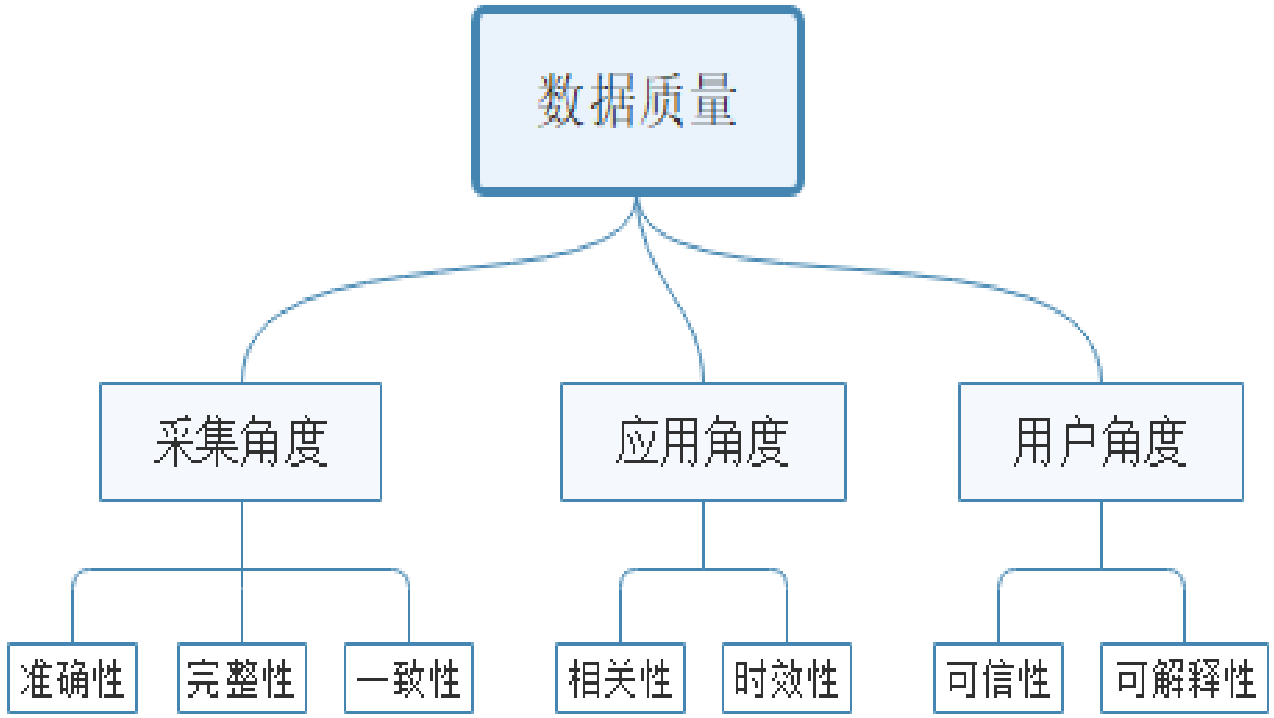
数据预处理的目的

数据采集 → 数据预处理 → 数据存储 → 数据分析挖掘 → 数据可视化

➢**重要性：**数据预处理是数据挖掘中必不可少的关键一步，更是进行数据挖掘前的准备工作。

➢**目的：**达到改进数据的质量，提高数据挖掘过程的准确率和效率。

- 保证数据挖掘的正确性和有效性。
- 通过对数据格式和内容的调整，使得数据更符合挖掘的需要。



目的

提升数据质量，符合数据挖掘的需求，保证数据挖掘的正确性和有效性

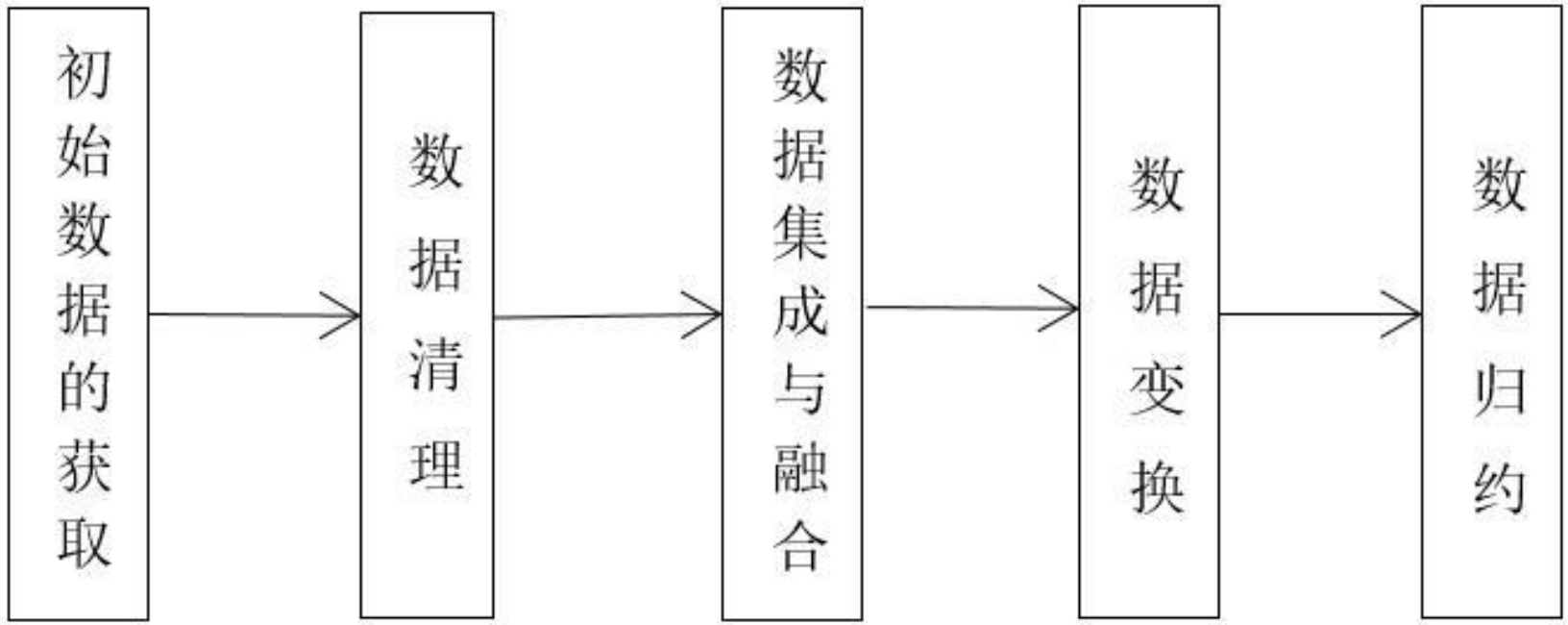
03

数据预处理流程

- 数据清理
- 数据集成
- 数据规约
- 数据变换

Recommend
火龙果软件

预处理流程



数据清理

脏数据：由于**重复录入**，**并发处理等不规范**的操作，导致产生**不完整**，**不准确**的，**无效**的数据。

越早处理脏数据，数据清理操作越简单

学号	姓名	性别	年龄	所在专业
95001	张 晓云	女	18	M01
95002	刘 一天	男	19	M02
95003	邓茹	女	180	M03
95004	王 小刚	男	20	M05

数据清理

学号	姓名	性别	年龄	所在专业
95001	张晓云	女	18	M01
95002	刘一天	男	19	M02
95003	邓茹	女	18	M03
95004	王小刚	男	Null	M05

数据清理试图填充空缺的值、识别孤立点、消除噪声，并纠正数据中的不一致性。

- 缺失值处理
- 噪声数据平滑

缺失值处理

缺失值：现有数据集中某个或某些属性的值是不完整的

客户信息表

客户编号	客户名称	……	风险等级	收入
0001	张三		3	8000
0002	李四		2	12000
0003	王五		2	11000
0004	赵六		1	20000
0005	李木		2	NULL
0006	王权		1	25000

缺失值处理方法：

- 忽略元组
- 人工填写
- 属性的中心度量来填充
- 使用于给定元组同一类的所有样本平均值
- 使用最可能的值

缺失值处理

客户信息表

客户编号	客户名称	……	风险等级	收入
0001	张三		3	8000
0002	李四		2	12000
0003	王五		2	11000
0004	赵六		1	20000
0005	李木		2	NULL
0006	王权		1	25000

- 元组
- 属性
- 中心度量
  - 平均值
  - 中位数

客户收入：8000、11000、12000、20000、25000

缺失值处理

忽略元组

客户编号	客户名称	……	风险等级	收入
0001	张三		3	8000
0002	李四		2	12000
0003	王五		2	11000
0004	赵六		1	20000
				NULL
0006	王权		1	25000

人工填写：人为填充，工作量大，不可行

全局变量：采用全局常量来替换空缺值

缺失值处理

使用属性的中心度量

收入
8000
12000
11000
20000
15200
25000

用同类样本属性的中心度量

客户编号	客户名称	……	风险等级	收入
0001	张三		3	8000
0002	李四		2	12000
0003	王五		2	11000
0004	赵六		1	20000
0005	李木		2	11500
0006	王权		1	25000

使用最可能的值：利用回归、贝叶斯计算公式推断最可能的值

噪声处理

噪声是被测量的变量的随机误差或偏差。

孤立点：不符合数据模型的数据。

噪声处理的目的：降低对数据分析和结果的影响

引起噪声数据的原因：

- 数据收集工具的问题
- 数据输入错误
- 数据传输错误
- 技术的限制
- 命名规则不一致

噪声处理的方法：

分箱法

回归

聚类

噪声处理

- 分箱：通过考察数据的**近邻**来光滑有序数据值。

举例：8、24、15、41、6、10、18、67、25等9个数。
排序：6、8、10、15、18、24、25、41、67。
分箱：箱1：6、8、10
箱2：15、18、24
箱3：25、41、67

分别用三种不同的分箱法求出平滑存储数据的值：

平均值：8；　箱1：8,8,8。

中值：18　箱2：18,18,18 。

边界值：　箱3：25,25,67，箱中的最大和最小值被视为箱边界。

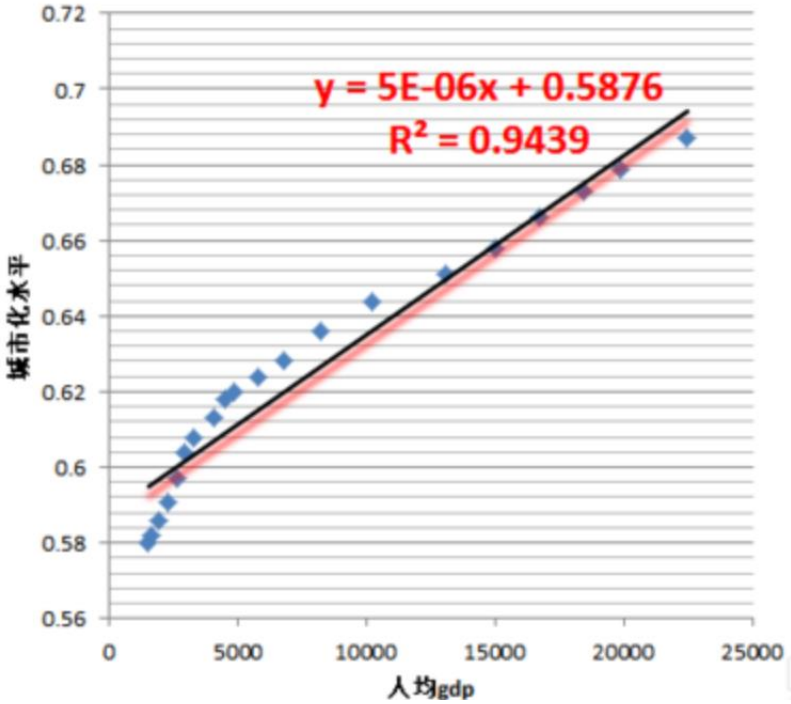
噪声处理

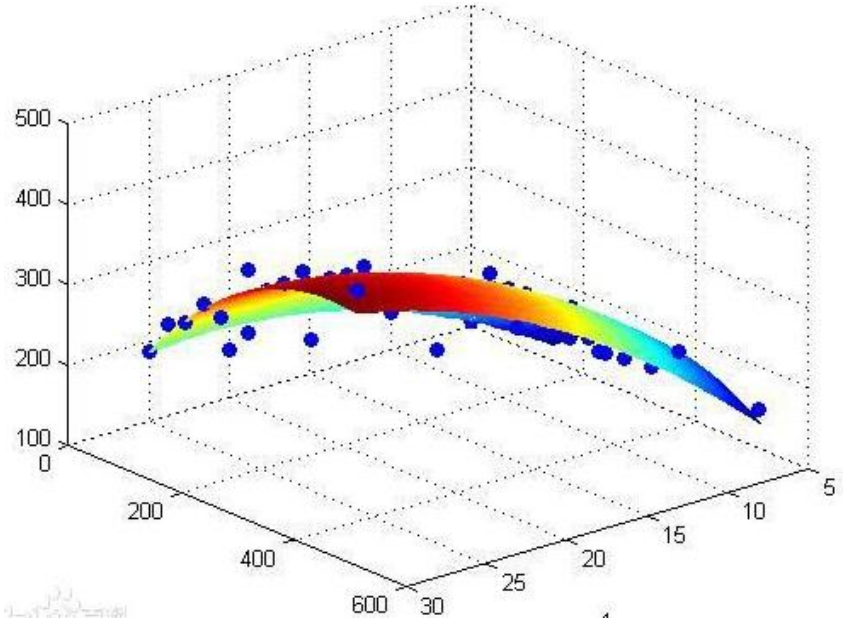
➢ 回归：采用一个函数拟合数据来光滑数据。

- 线性回归：找出拟合两个属性的“最佳”直线，帮助平滑数据及除去噪声。

- 多元线性回归：设计多个属性，拟合到一个多维的面

年份	城市化水平	人均GDP（元）
1981	0.58	1558
1982	0.582	1704
1983	0.586	1979
1984	0.591	2306
1985	0.597	2702
1986	0.604	2953
1987	0.608	3336
1988	0.613	4124
1989	0.618	4509
1990	0.62	4878
1991	0.624	5782
1992	0.628	6804
1993	0.636	8239
1994	0.644	10261
1995	0.651	13085
1996	0.658	15044
1997	0.666	16735
1998	0.673	18478
1999	0.679	19846
2000	0.687	22460





噪声处理

➢ 聚类：可以通过聚类算法来检测离群点，消除噪声

✓相似或相邻近的数据聚合在一起形成各个聚类集合，而那些位于聚类集合之外的数据对象，被视为离群点。

聚类算法：
K均值方法：
层次聚类方法
基于密度聚类方法
……

数据清理

- 数据清理
  - 缺失值处理
    - 忽略元组
    - 人工填写
    - 使用全局常量
    - 使用属性的中心度量
    - 使用同类样本属性的中心度量
    - 使用最可能的值　　回归、贝叶斯、决策树
  - 噪声处理
    - 分箱
    - 回归
    - 聚类

03

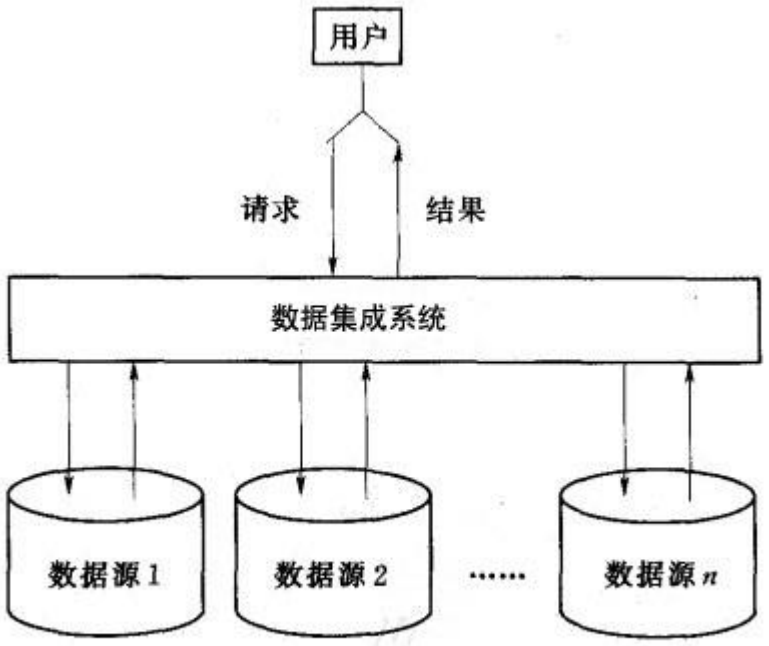
数据预处理流程

- 数据清理
- **数据集成**
- 数据变换
- 数据规约

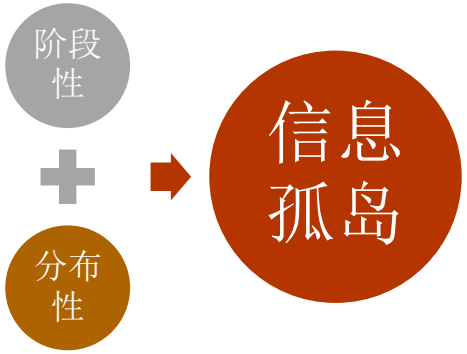
Recommend
火龙果软件

数据集成

定义：将互相关联的分布式异构数据源**集成**到一起，使用户能够以**透明的方式**访问这些数据源。



数据集成系统模型



信息孤岛：指不同软件间，尤其是不同部门间的数据信息不能共享，造成系统中存在大量冗余数据、垃圾数据，无法保证数据的一致性。

数据集成

数据集成的方法：联邦数据库，中间件集成，数据复制。

- 联邦数据库：将各数据源的数据**视图**集成为**全局模式**

- 中间件集成：通过统一的**全局数据模型**来访问异构的数据源

- 数据复制：将各个数据源的数据复制到同一处，即**数据仓库**

数据集成

- 集成过程中需要处理的问题

实体识别

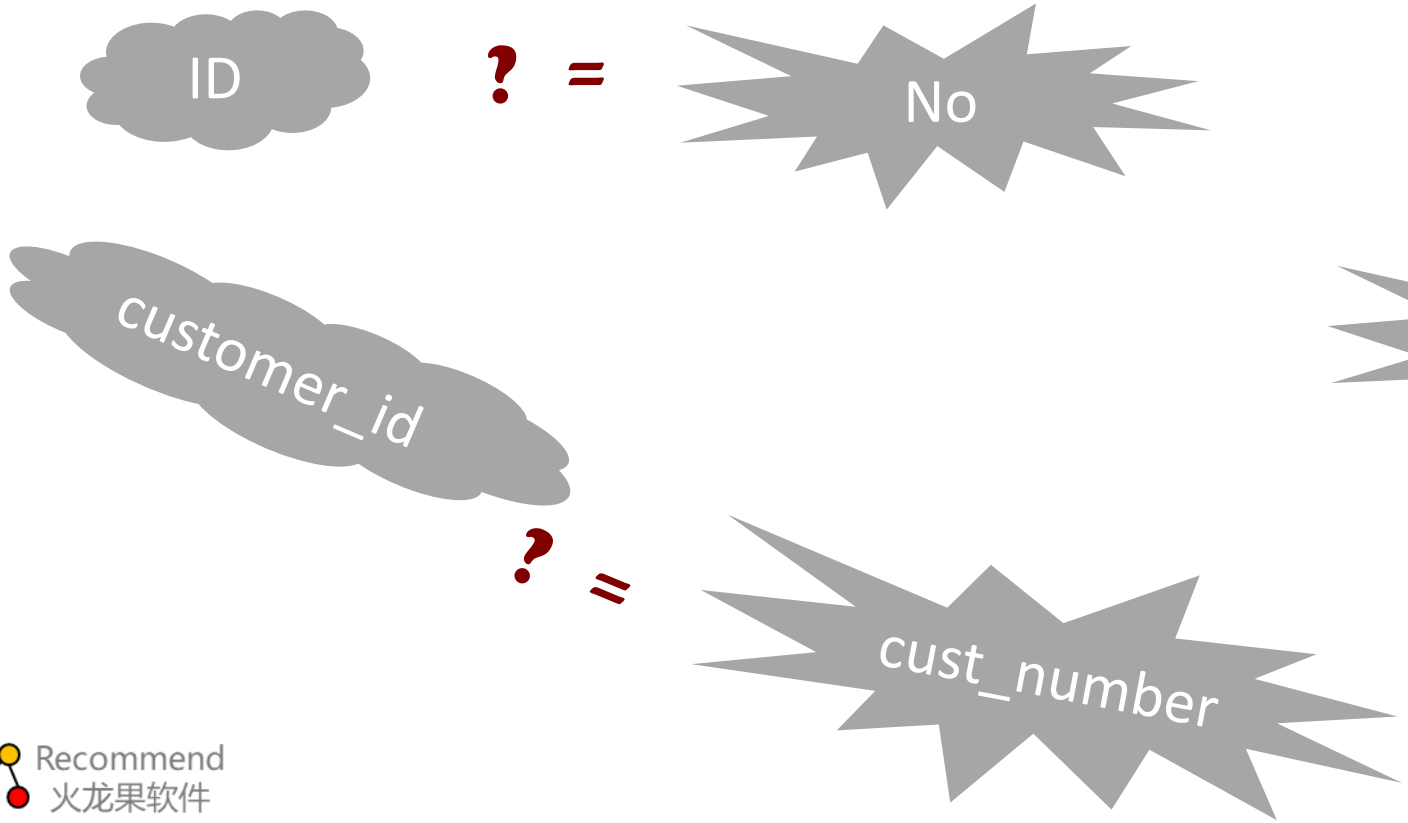
冗余与相关分析

数据冲突和检测

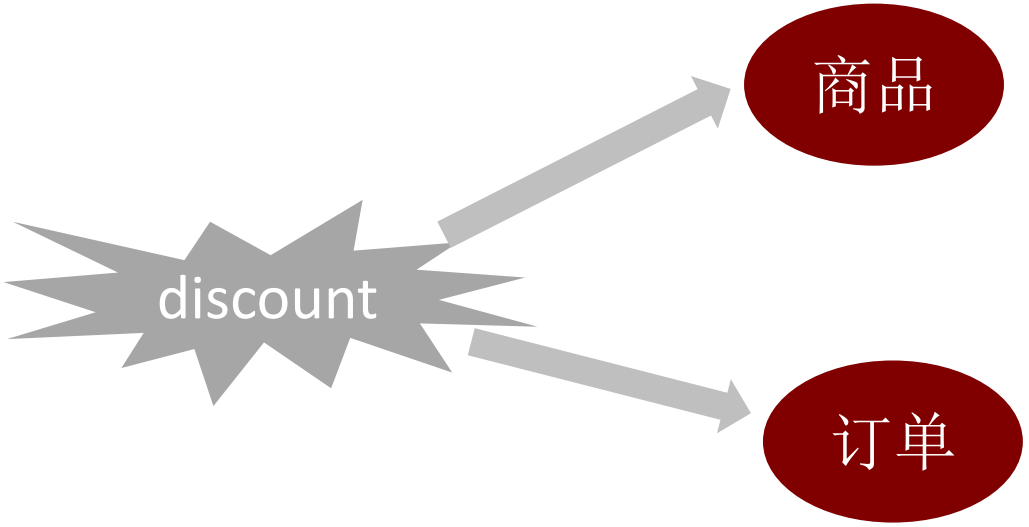
数据集成

➢ **实体识别：**匹配多个信息源在现实世界中的等价实体。

同义不同名



同名不同义



数据集成

➢ **冗余问题：**属性重复，属性相关冗余，元组重复。

属性重复

属性相关冗余

数值数据的相关系数

数值数据的协方差

客户编号	客户名称	……	female	性别	月薪	年收入
0001	张三		0	男	8	96
0002	李四		1	女	7	84
0003	王五		1	女	6	72
0004	赵刚		0	男	7.5	90
0005	赵刚		0	男	7.5	90

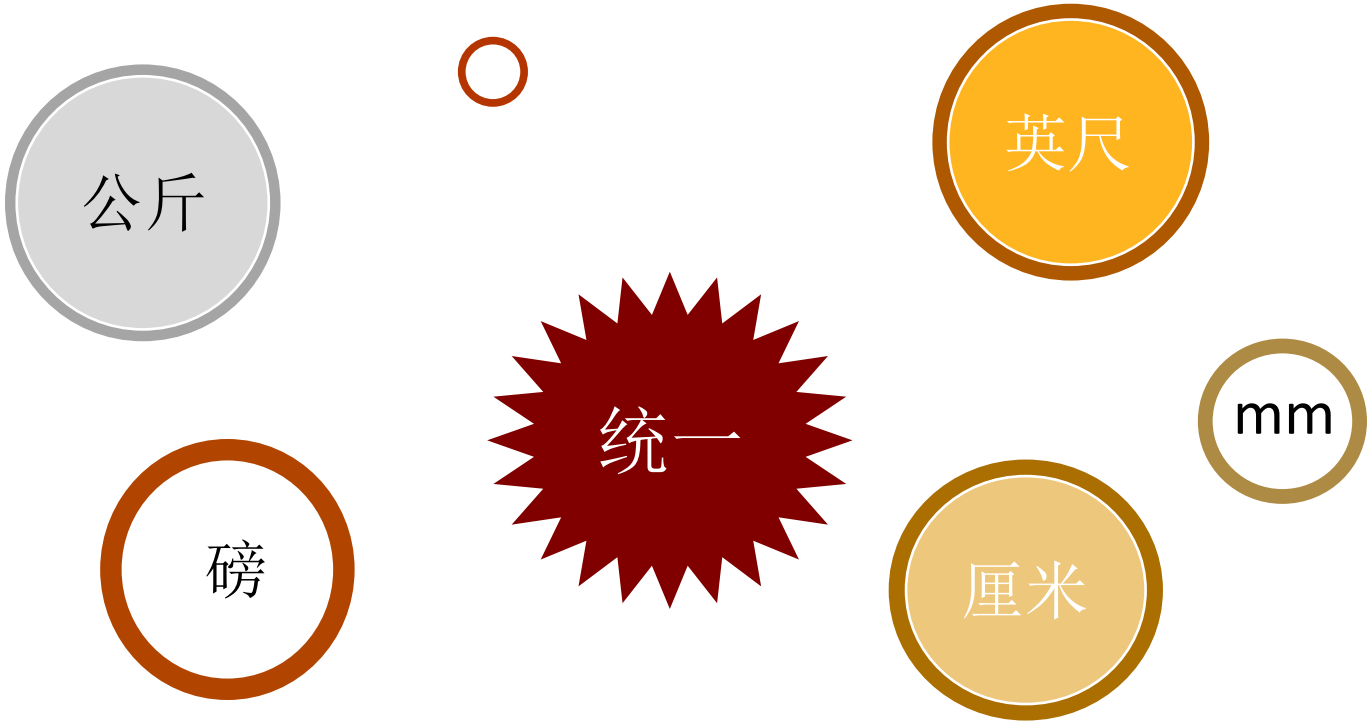
元组重复

数据集成

➢ 数据冲突的检测与处理

对现实世界的同一实体，来自不同数据源的属性定义不同。

原因：表示方法，度量单位、编码或比例的差异



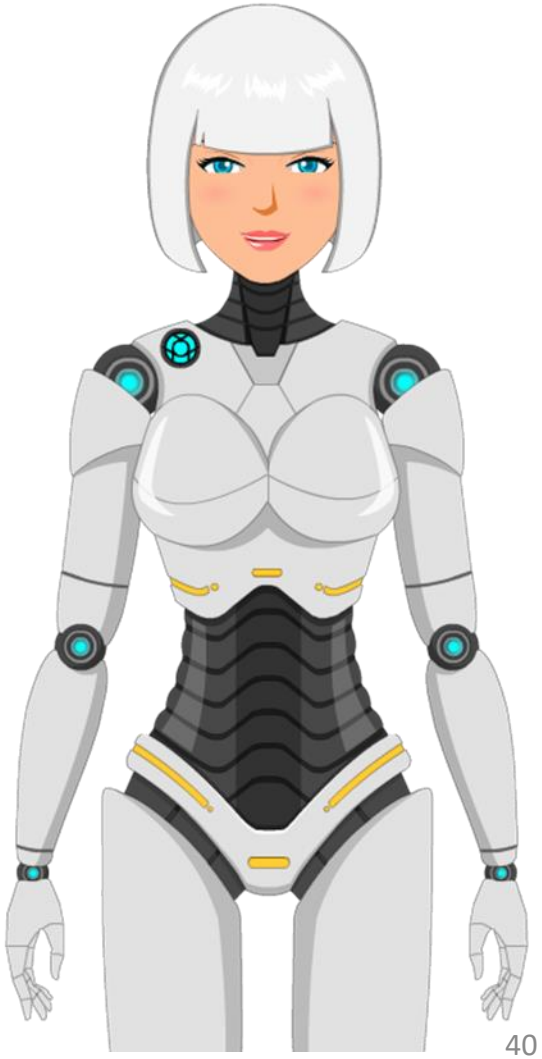
小结

- 数据集成
  - 实体识别
  - 冗余与相关分析
  - 数值冲突检测与处理

03

数据预处理

- 影响数据质量的因素
- 数据清理
- 数据集成
- **数据规约**
- 数据变换



数据规约

➢ 为什么要数据规约

在现实场景中，数据集是很庞大的，数据是海量的，在整个数据集上进行复杂的数据分析和挖掘需要花费很长的时间。

➢ 数据规约目的

用于帮助从原有庞大数据集中获得一个**精简**的数据集合，并使这一精简数据集**保持原有数据集的完整性**，这样在精简数据集上进行数据挖掘显然效率更高，并且挖掘出来的结果与使用原有数据集所获得结果是**基本相同**。

➢ 标准

- 用于规约的时间不应当超过或抵消在规约后的数据上挖掘节省的时间。
- 规约得到的数据比原数据小得多，但可以产生相同或几乎相同的分析结果

数据规约

- 维归约
- 数量归约
- 数据压缩

维归约-主成分分析

主成分分析:

主成分分析也称主分量分析，旨在利用降维的思想，把多指标转化为少数几个综合指标（即主成分），其中每个主成分都能够反映原始变量的大部分信息，且所含信息互不重复。

----这种方法在引进多方面变量的同时将复杂因素归结为几个主成分，使问题简单化，同时得到的结果更加科学有效的数据信息。

2004-2007 年行业效益变化

	2007 年 11 月	2006 年	2005 年	2004 年
		盈利能力		
销售毛利率	14.13	11.84	8.75	8.42
销售利润率	9.74	6.59	3.87	2.11
资产报酬率	5.04	4.66	3.21	1.98
		偿债能力		
负债率	81.27	79.59	81.06	81.46
亏损面	17.60	14.98	15.27	21.21
利息保障倍数	108.16	10.67	5.27	2.58
		发展能力		
应收帐款增长率	77.74	23.18	40.20	26.74
利润总额增长率	135.74	97.88	119.58	119.86
资产增长率	65.15	38.55	20.03	20.90
销售收入增长率	62.31	30.07	42.59	35.21
		营运能力		
应收帐款周转率	9.08	11.83	10.83	8.66
产成品周转率	23.56	39.04	41.30	36.40
流动资产周转率	0.88	1.10	1.07	0.95

维归约--属性子集选择

属性子集选择

- 用于检测并删除不相关、弱相关或冗余的属性。
- 目标是找出最小属性集，使得数据类的概率分布尽可能地接近使用所有属性得到的原分布

分析顾客是否愿意购买新的流行CD

顾客分类	年龄	喜好音乐类型	电话号码	住址

数量规约

数量归约：通过选择**替代的、较小的数据**表示形式来减少数据量。

（1）有参方法：回归和对数的线性模型

（2）无参方法：直方图、聚类、抽样和数据立方体聚集

数量规约

（1）有参方法：**回归和对数的线性模型**

通常使用一个参数模型来评估数据，该方法只需要存储参数，而不是实际数据，能大大减少数据量，但只对数值型数据有效。

如：线性回归方法：$Y=\alpha+\beta X$

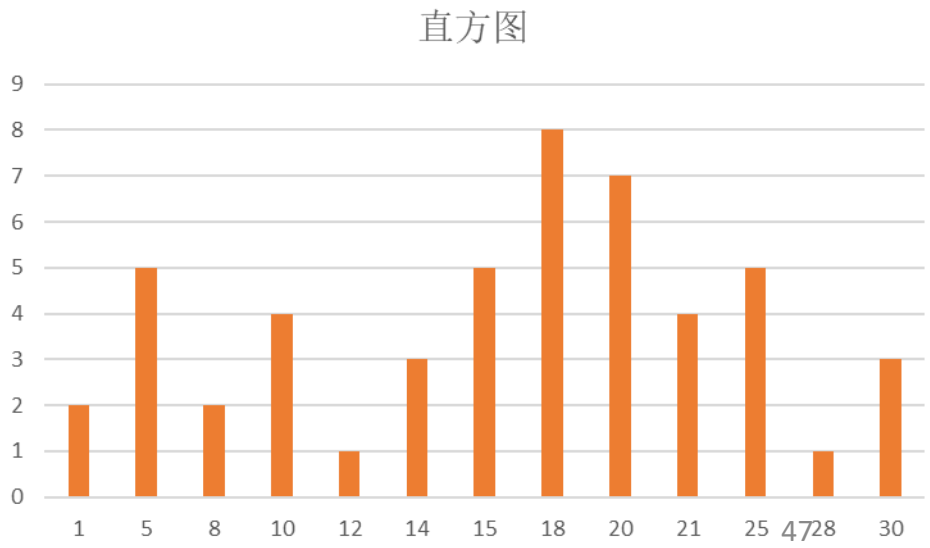
数量规约

（2）无参方法：

直方图：根据属性的数据分布将其分成若干不相交的区间，每个区间的高度与其出现的频率成正比。

例：下面的数据是某商店销售的商品的单价表（已排序）：
1,1,5,5,5,5,5,8,8,10,10,10,10,12,14,14,14,15,15,15,15,15,15,18,18, 18,18, 18,18, 18,18,20,20, 20,20, 20,20,20,21,21, 21,21,25,25, 25,25,25,28,28,30,30,30。
试用直方图表示，以压缩数据。

价格清单：1(2),5(5),8(2),10(4),12,14(4),15(5), 18(8),20(7),21(4),25(5),28,30(3)



Recommend
火龙果软件

数量规约

聚类：将原数据集划分成多个群或聚类。

◆原则：同类中的数据彼此相似；不同类中的数据彼此不相似。

◆相似：通常用空间距离度量

◆ 聚类的有效性依赖于实际数据的内在规律。

数量规约

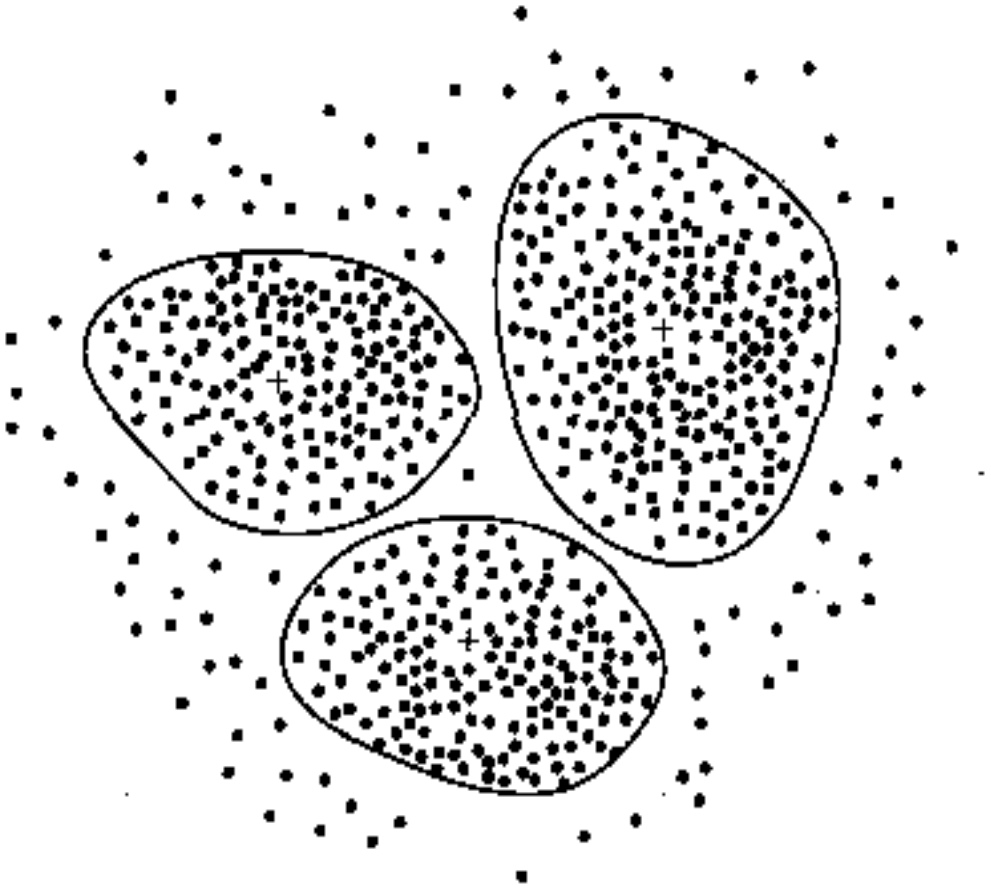


图3-3 孤立点可以被聚类分析检测

数量规约

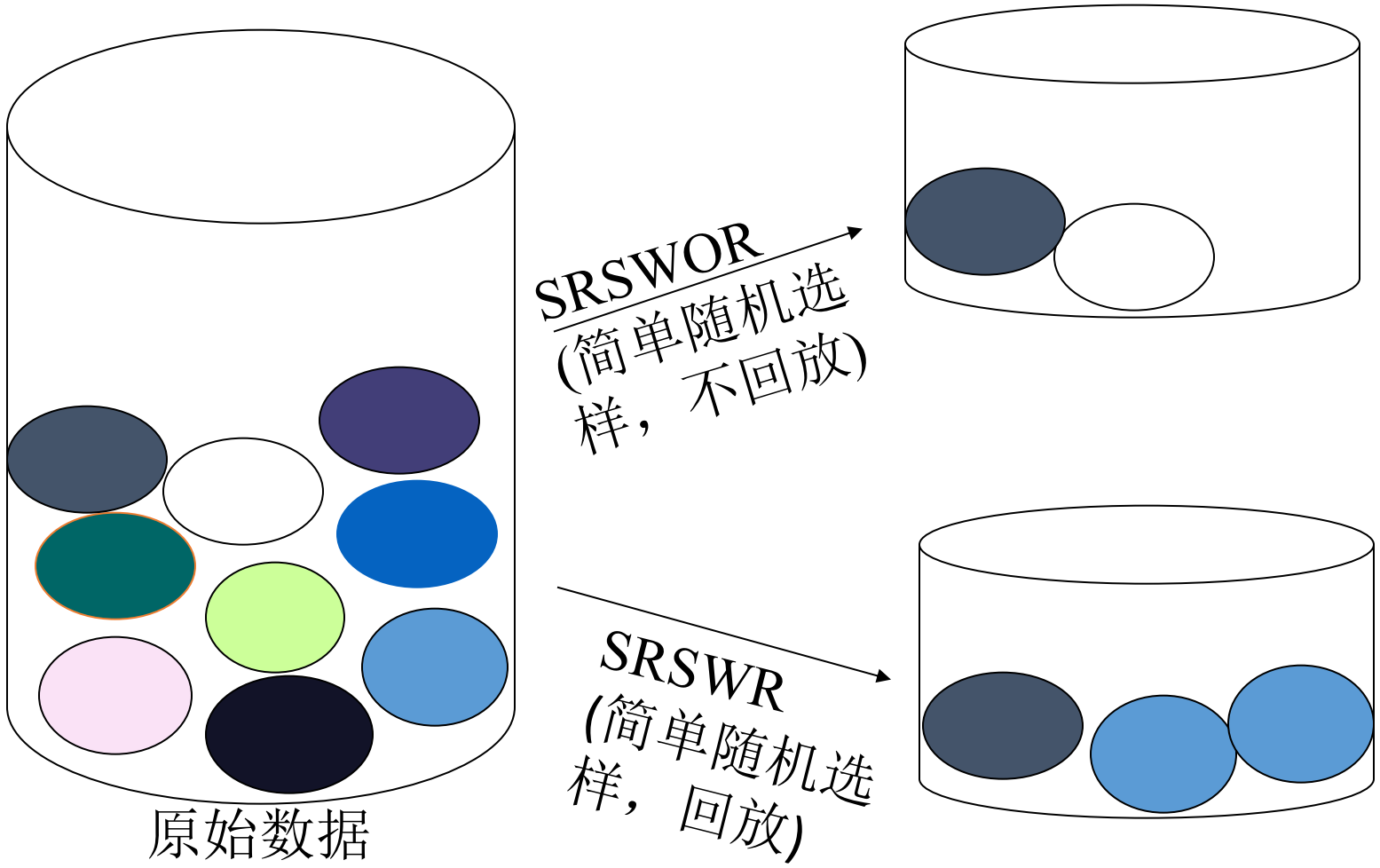
取样（抽样）：允许用数据的较小随机样本（子集）表示大的数据集。

取样方法：

- 不放回简单随机取样 （Simple Random Sampling Without Replacement, SRSWOR）
- 放回简单随机取样（Simple Random Sampling With Replacement, SRSWR）
- 聚类取样：先聚类，后取样
- 分层取样：先分层，后取样。

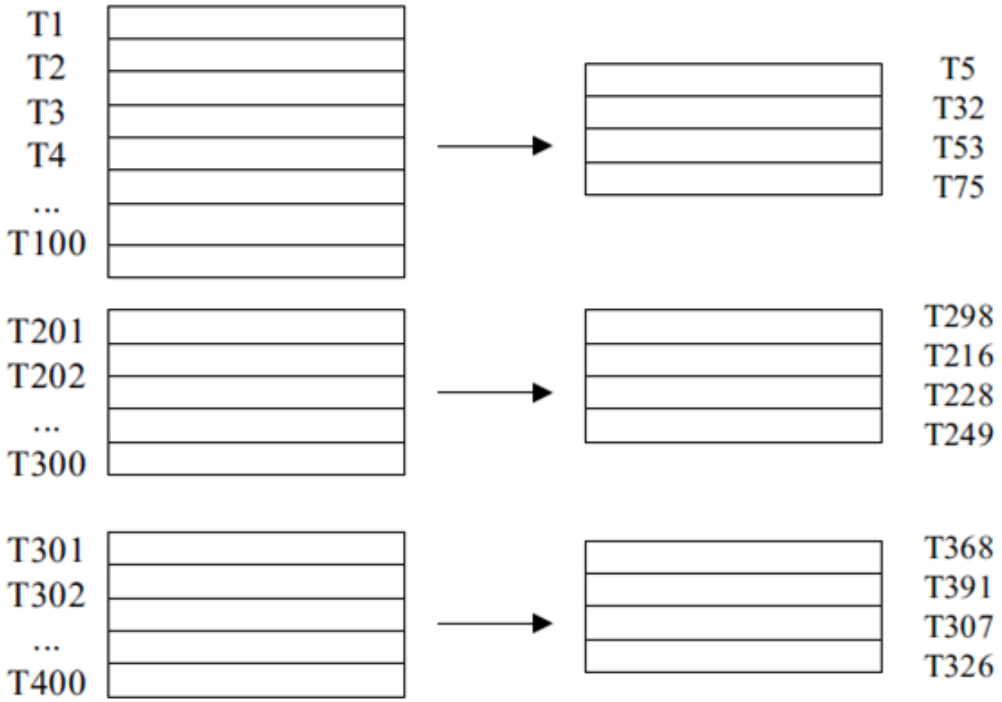
Recommend 火龙果软件

数量规约： 取样——SRS



数量规约

聚类采样： 首先将大数据集D划分为M个互不相交的聚类，然后再从M个类中的数据对象分别进行随机抽取，可最终获得聚类采样的数据子集。

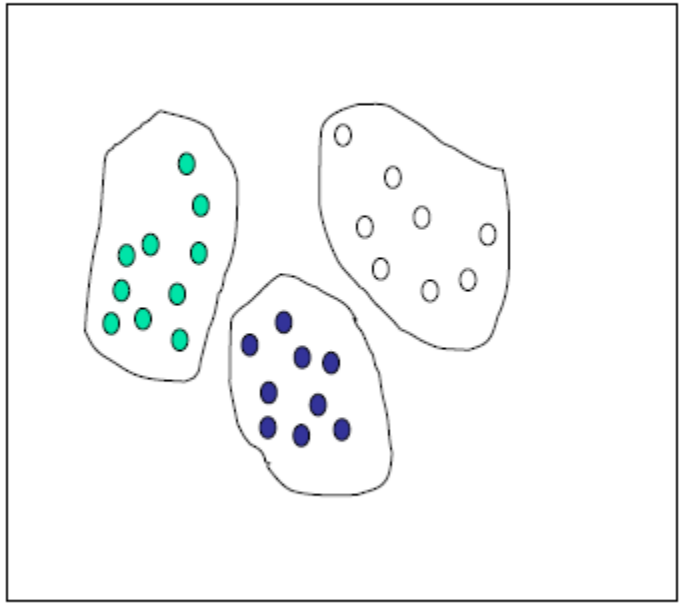


聚类采样方法示意图

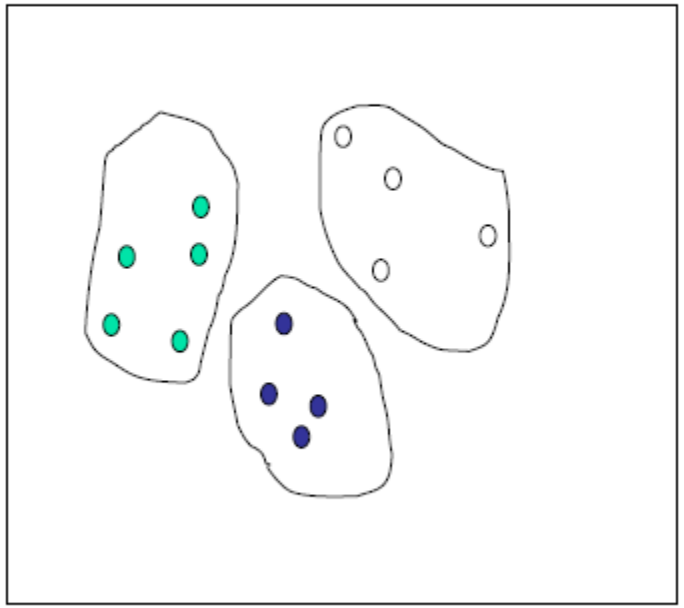
数量规约

聚类取样图示

Raw Data



Cluster/Stratified Sample

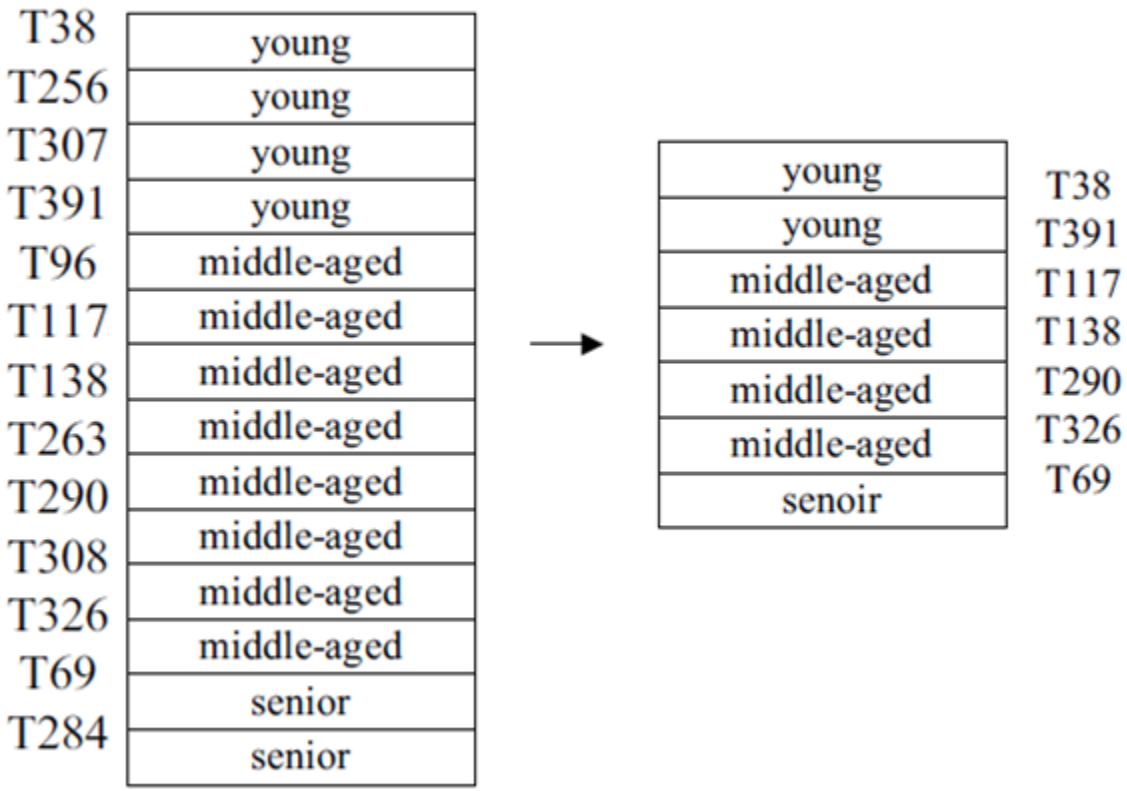


数量规约

分层取样： 首先将大数据集D划分为互不相交的层，然后对每一层简单随机选样得到D的分层选样。

数量规约

分层取样图示



分层采样方法示意图

根据顾客的年龄组进行分层，然后再在每个年龄组中进行随机选样，从而确保了最终获得分层采样数据子集中的年龄分布具有代表性。

数量规约

数据立方体聚集

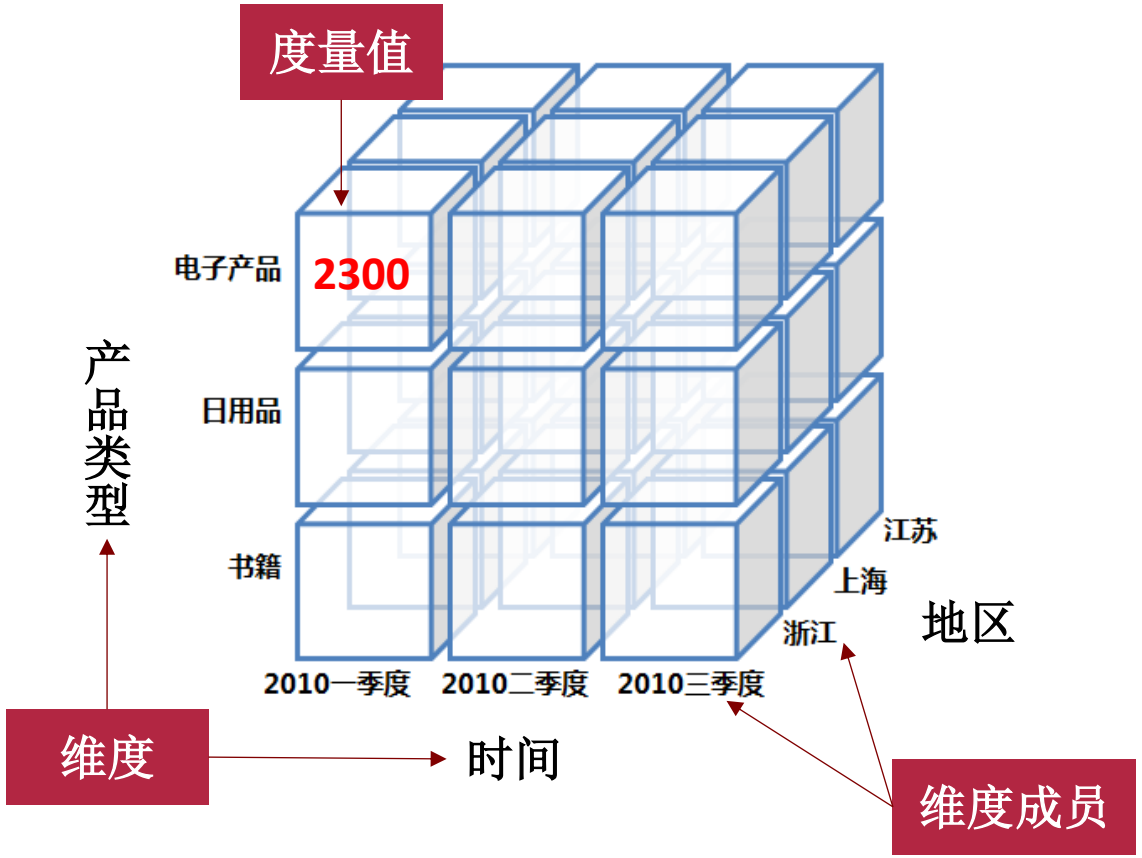
➢ 数据立方体概念

数据立方体是数据的**多维建模和表示**，由维度、维度成员和度量值组成。

- **维度**:观察数据的角度
- **维度成员**:维的取值
- **度量值**:实际意义的数值

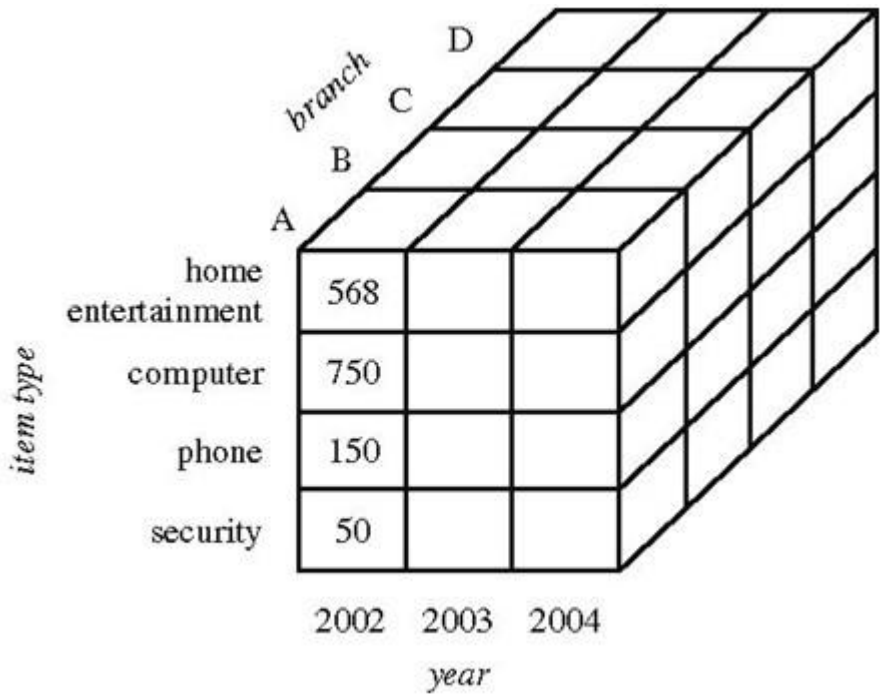
➢ 数据立方体聚集定义

- 将**n维**数据立方体聚集成**n-1维**的数据立方体



数据规约

数据立方体聚集



	2002	2003	2004
Home entertanment	2033	3000	4000
computer	3006	4555	6000
phone	600	1000	4000
security	200	500	1000

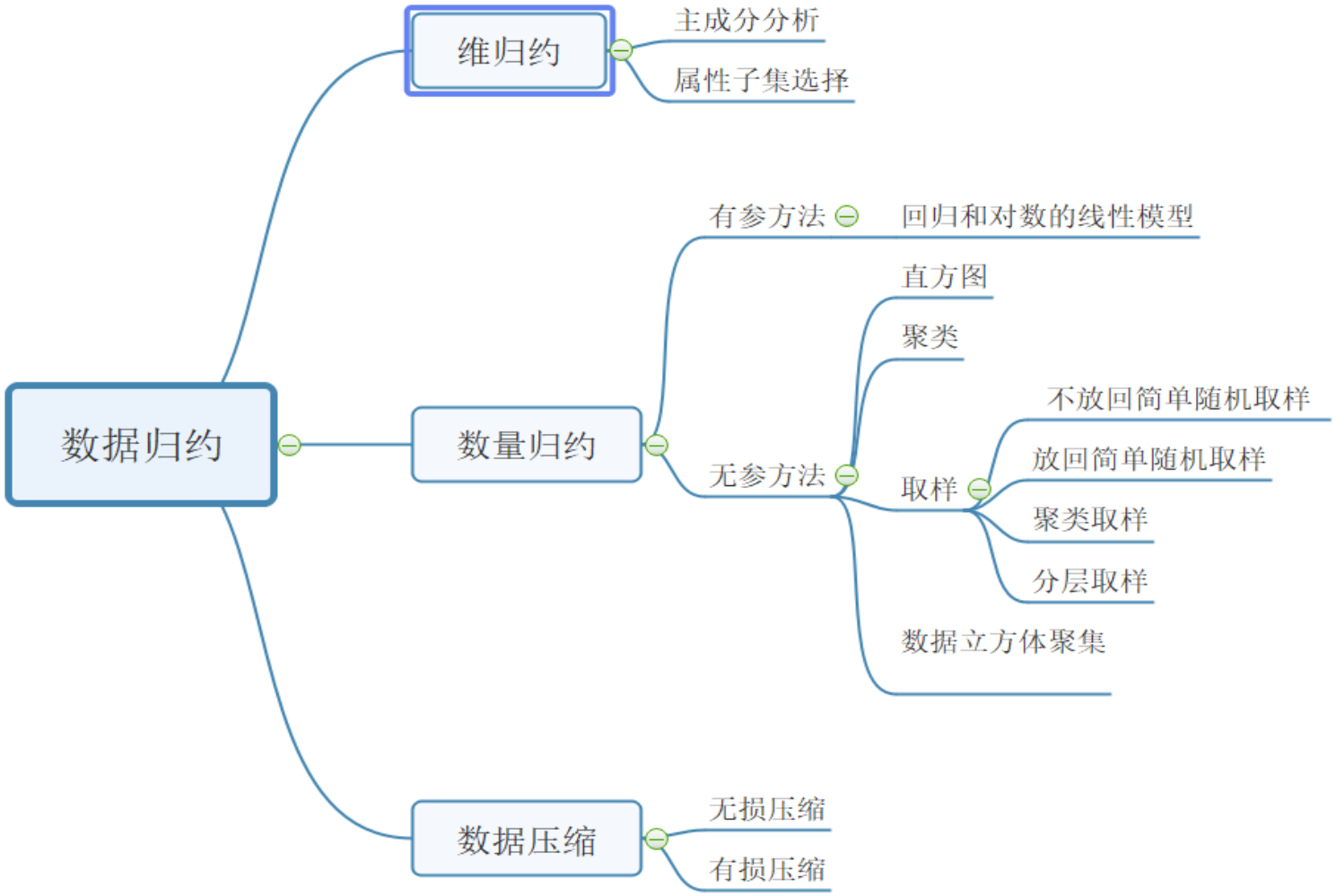
数据归约

数据压缩：利用数据编码或数据转换将原来的数据集合**压缩**为一个**较小规模**的数据集合。

- 无损压缩（loseless）：可以不丢失任何信息地**还原**压缩数据；如：字符串压缩，压缩格式：Zip或RAR
- 有损压缩（lossy）：只能重新构造原数据的**近似**表示；如：音频/视频压缩。

- 音频能够在没有察觉的质量下降情况下实现 10:1 的压缩比，
- 视频能够在稍微观察质量下降的情况下实现如 300:1 这样非常大的压缩比。

数据规约小结


数据归约
维归约
主成分分析
属性子集选择
数量归约
有参方法
回归和对数的线性模型
无参方法
直方图
聚类
取样
不放回简单随机取样
放回简单随机取样
聚类取样
分层取样
数据立方体聚集
数据压缩
无损压缩
有损压缩

Recommend
火龙果软件

03

数据预处理

- 影响数据质量的因素
- 数据清理
- 数据集成
- 数据规约
- **数据变换**

数据变换

目的：将数据转换或统一成**易于进行数据挖掘**的数据存储形式，使得挖掘过程可能更有效。

方法策略：

- 光滑：去掉数据中的噪音。
- 属性构造：由给定的属性**构造新的属性**并添加到属性集中，帮助数据分析和挖掘。
- 聚集：对数据进行**汇总**或**聚集**
- 规范化：将属性数据按**比例缩放**，使之落入一个小的特定区间
- 离散化：数值属性用**区间标签**或**概念标签**替换。
- 由标称数据产生概念分层：属性，如street，可以泛化到较高的概念层，如city或country。

数据变换

规范化：将数据按比例进行缩放，使之落入一个特定的区域。

常用的算法：

- 最小-最大规范化；
- 零-均值规范化（z-score规范化）；
- 小数定标规范化。

数据变换

数据规范化：将数据按比例缩放至一个小的特定区间：

1）**最小—最大规范化**：假定$\min_A$和$\max_A$分别为属性A的最小和最大值，则通过下面公式将A的值映射到区间[new_minA，$new_\max_A$]中：

$$v' = \frac{v - \min_A}{\max_A - min_A}(new_\max_A - new_min_A) + new_min_A$$

例：假定属性income的最小与最大值分别为$12000和$98000，可根据最小—最大规范化方法将其范围映射到[0,1]：

如：属性值$**73600**将变换为：

[(**73600**-12000)/(98000-12000)]*(1-0)+0=**0.716**

数据变换

2）z-score规范化（零均值规范化）：

- 将属性A的值根据其**平均值**和**标准差**进行规范化；
- 常用于属性最大值与最小值未知，或使用最小最大规范化方法会出现异常数据的情况。

$$v' = \frac{v - mean_A}{standard_dev_A}$$

其中，$mean_A$、$standard_devA$分别为属性A取值的**均值**和**标准差**。

例： 假定属性income的平均值与标准差分别为$**54000**和$**16000**，使用z-score规范化，则属性值$73600将变换为：

(73600-**54000**）/**16000**=1.225

Recommend
火龙果软件

数据变换

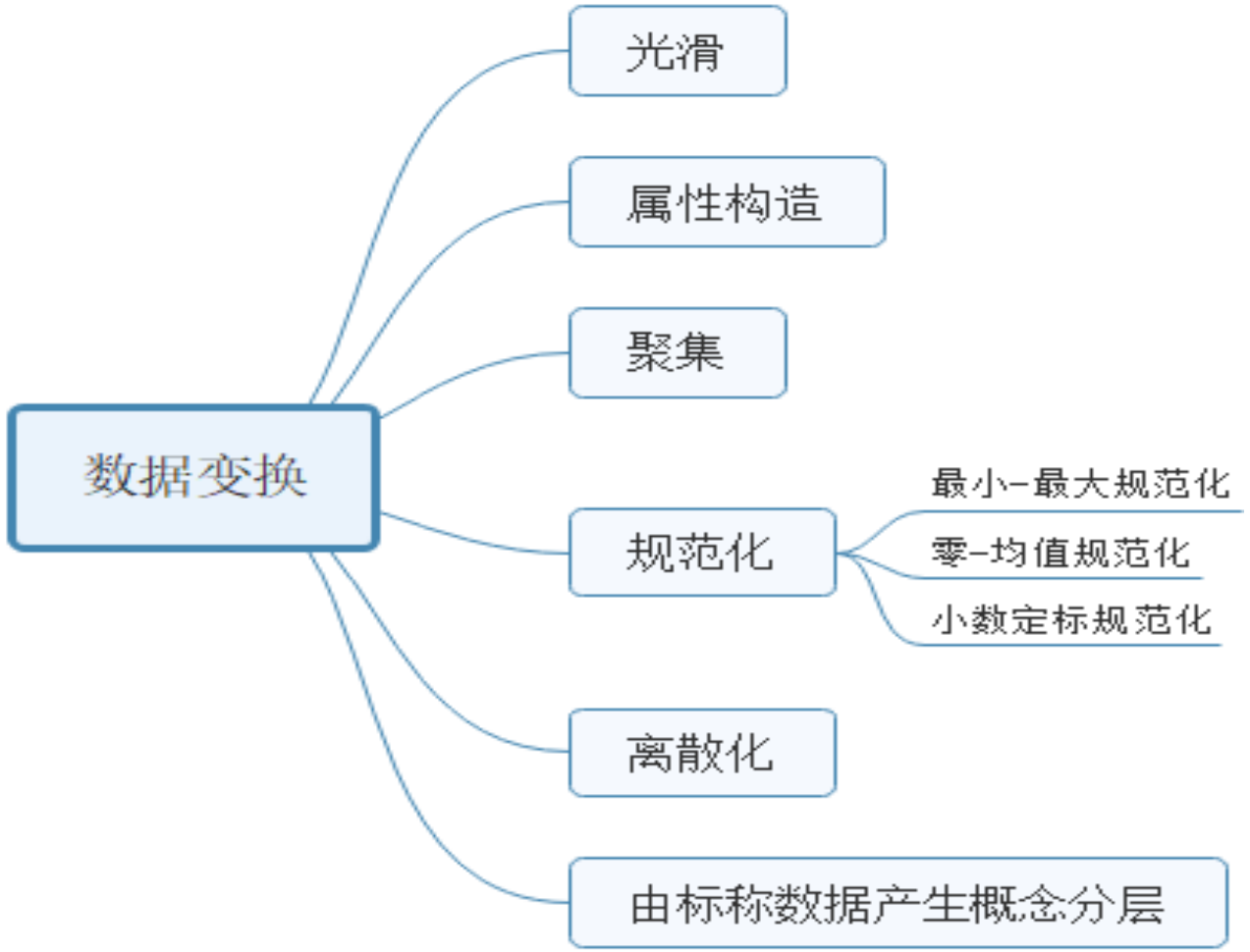
3）**小数定标规范化**：通过移动属性A的小数点位置进行规范化，小数点的移动依赖于A的最大绝对值：

$$v' = \frac{v}{10^j}$$

其中，j是使 Max(| v' |)<1的最小整数

例：假定A的取值范围[-986, 917]，则A的最大绝对值为986，为使用小数定标规范化，用1000（即j=3）除每个值，这样-986被规范化为-0.986。

数据变换总结



数据预处理注意项

◆数据预处理的上述流程并不是完全分开的，在某种场景下是可以一起使用的。

◆应该针对具体所要研究的问题通过详细分析后再进行预处理方案的选择。

◆整个预处理过程要尽量人机结合，尤其要注重和客户以及专家多交流。

◆预处理后，若挖掘结果显示和实际差异较大，在排除源数据的问题后，则有必要考虑数据的二次预处理。

04

数据预处理的工具

--推荐Kettle&python

数据预处理工具-Kettle

◆开源，且支持可视化编程。

◆支持各种数据源。

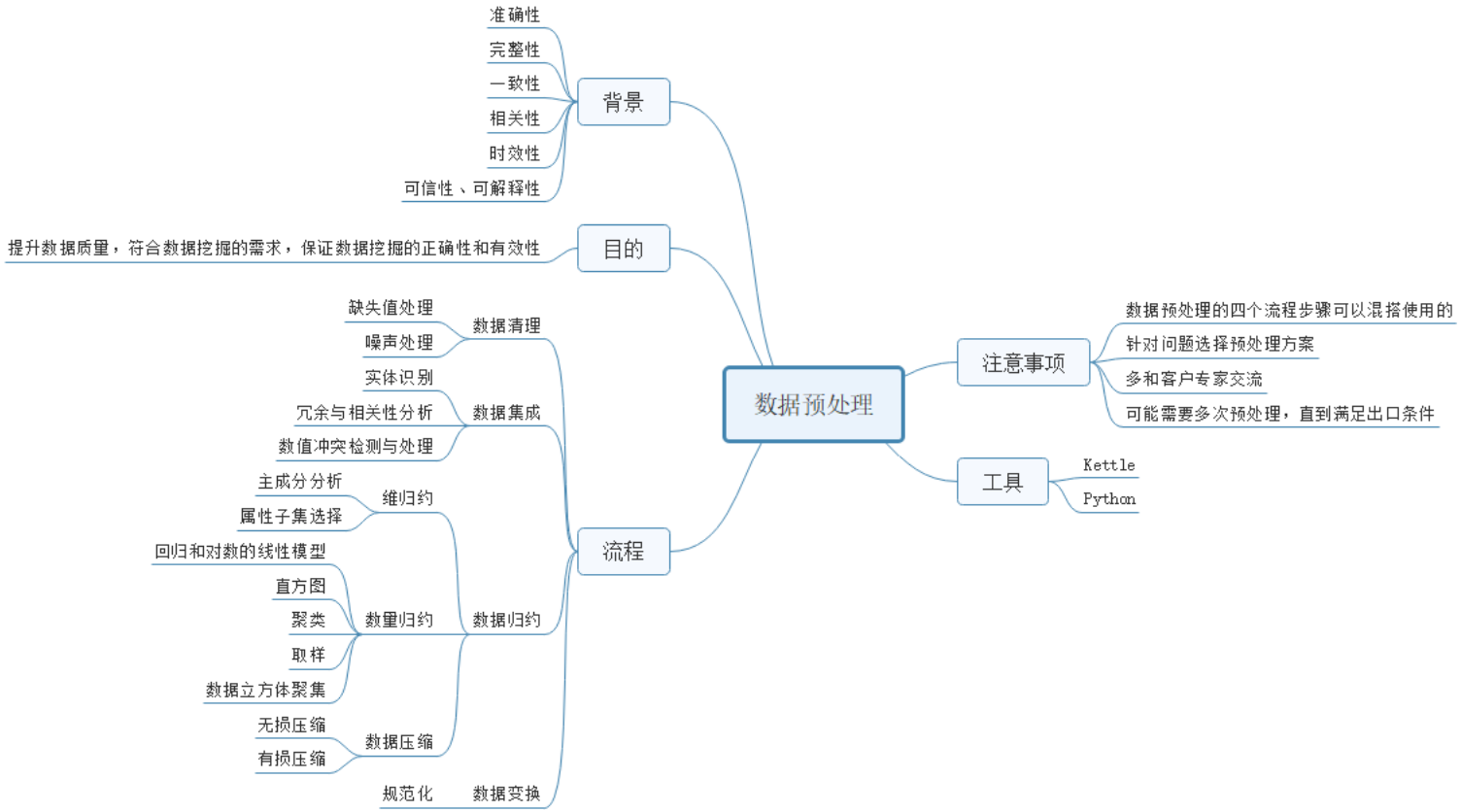
◆数据处理功能很强大。

数据预处理工具-Python

◆可以胜任很多领域的工作，是人工智能和大数据时代的明星语言。

◆是一种面向对象的解释型计算机程序设计语言，具有丰富和强大的库。

本章总结


数据预处理
背景
准确性
完整性
一致性
相关性
时效性
可信性、可解释性
目的
提升数据质量，符合数据挖掘的需求，保证数据挖掘的正确性和有效性
流程
数据清理
缺失值处理
噪声处理
数据集成
实体识别
冗余与相关性分析
数值冲突检测与处理
数据归约
维归约
主成分分析
属性子集选择
数量归约
回归和对数的线性模型
直方图
聚类
取样
数据立方体聚集
数据压缩
无损压缩
有损压缩
数据变换
规范化
注意事项
数据预处理的四个流程步骤可以混搭使用的
针对问题选择预处理方案
多和客户专家交流
可能需要多次预处理，直到满足出口条件
工具
Kettle
Python

感谢您的观赏

THANK YOU FOR WATCHING

Recommend
火龙果软件