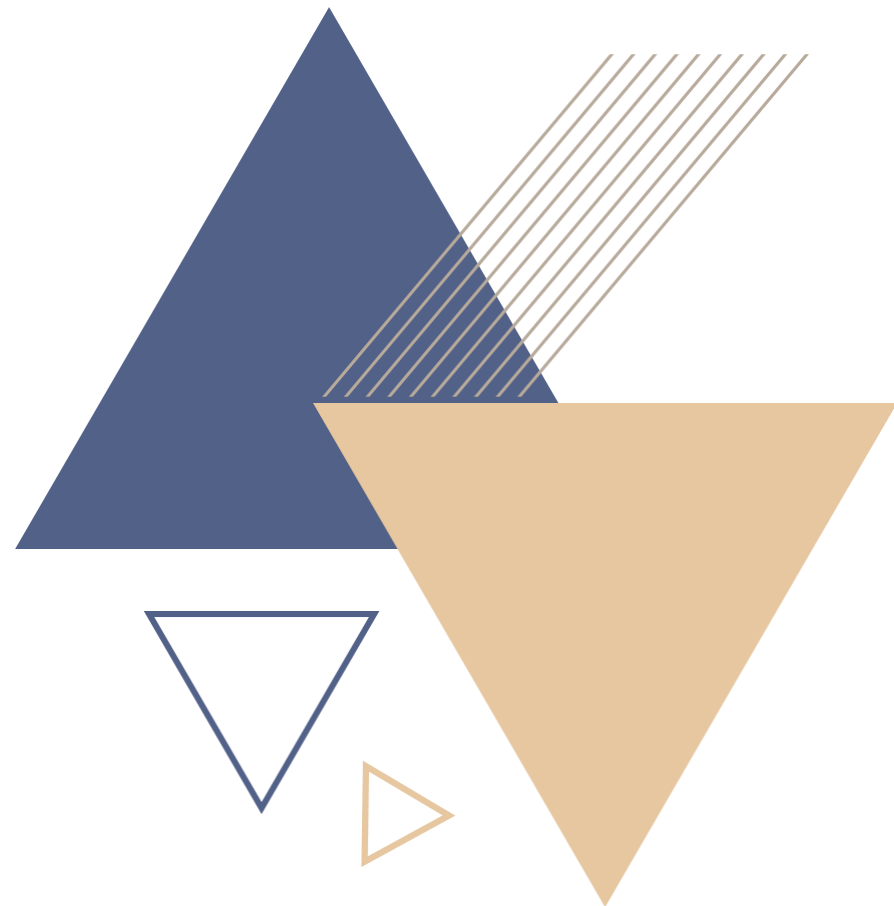


数据挖掘

第四讲· 数据预处理



Outline

1

数据清理

2

数据集成与变换

3

数据归约

4

数据离散化和概念分层产生

数据归约

- 为什么?

- 在海量数据上进行复杂的数据分析和挖掘需要很长的时间，使得这种分析不现实或者不可行，必须对数据进行归约。

- 怎么做?

- 数据归约技术可以用来得到数据集的归约表示，它小得多，但仍接近于保持源数据的完整性。

- 怎么样?

- 在归约以后的数据集上挖掘更加有效，并产生相同（或几乎相同）的分析结果。

数据归约的策略

维归约

减小所考虑的随机变量或属性的个数

- 小波变换
- 主成分分析
- 属性子集选择

数量归约

用替代的、较小的数据表示形式替换原数据

参数方法

- 回归
- 对数线性模型

非参数方法

- 直方图
- 聚类
- 抽样
- 数据立方体技术

数据压缩

使用变换获得压缩或者归约表示

- 有损压缩
- 无损压缩

维归约（1）：属性子集选择

- 属性子集选择，即：Attribute Subset Selection。
- 删除不相关的属性，减少数据量，通过找出最小属性集，使得数据类的概率尽可能接近全部的属性集。
- 对于属性子集选择，通常用压缩搜索空间的启发式算法，策略是局部最优选择。

维归约（1）：属性子集选择

■ 逐步向前选择 stepwise forward selection

- 由空属性开始，每次选一个最好的属性放入开始空属性集中，重复迭代；

■ 逐步向后删除 stepwise backward elimination

- 该过程由整个属性集开始，在每一步中，删除尚在属性集中的最坏的属性；

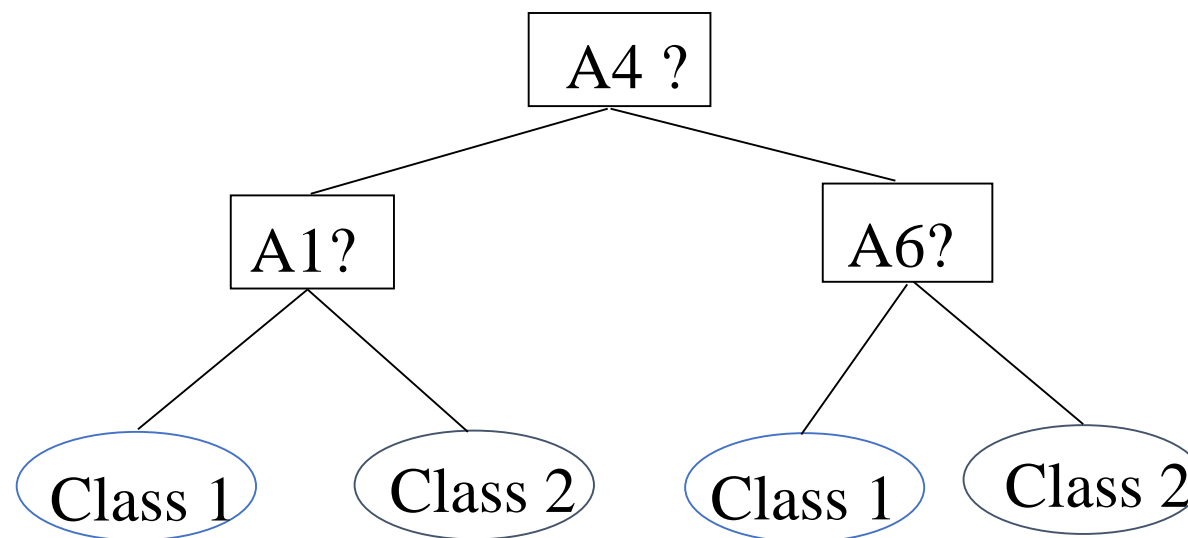
■ 逐步向前选择和逐步向后删除的结合

■ 决策树归纳 decision tree induction

- 构造一个类似于流程图的结构，其每个内部（非树叶）节点表示一个属性上的测试，每个分枝对应于测试的一个输出；每个外部（树叶）节点表示一个判定类
- 出现在树中的属性形成相关的属性集合。

维归约 (2) : 决策树归纳

初始属性: {A1, A2, A3, A4, A5, A6}



-----> 归约后的属性集: {A1, A4, A6}

维归约 (3) : 主成分分析

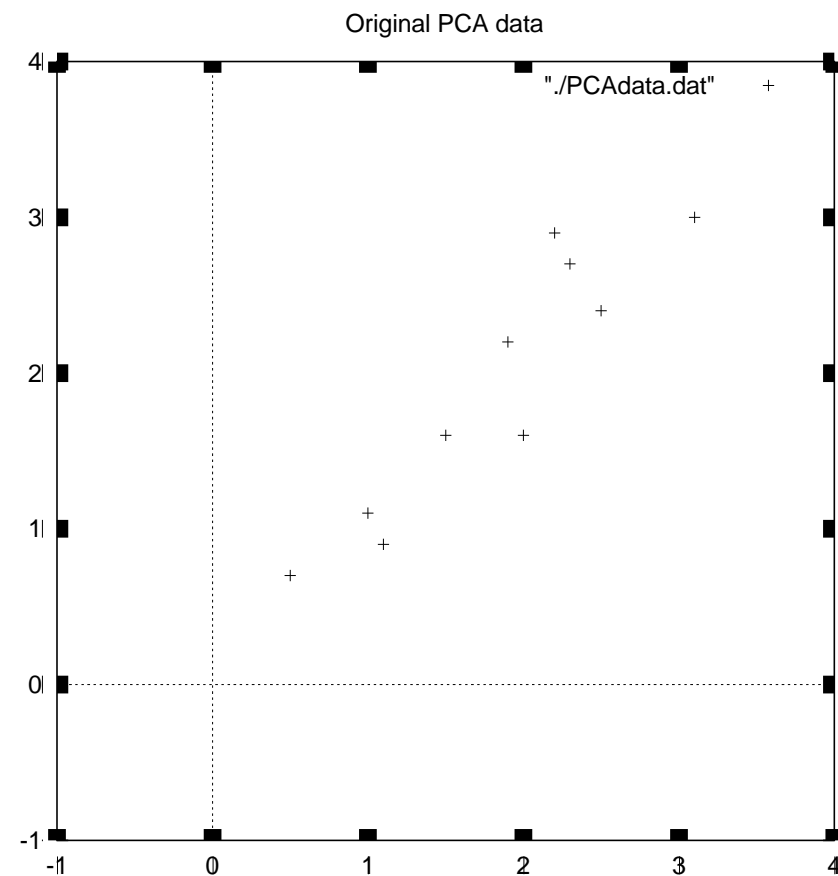
- Principal Component Analysis (PCA)

- 对于 n 维数据, PCA搜索 k 个最能代表数据的 n 维正交向量 ($k \leq n$) ;
- 减少数据集的维数, 同时保持数据集的对方差贡献最大的特征。保留低阶主成分, 忽略高阶主成分; 低阶成分往往能够保留住数据的最重要方面;
- 与属性子集选择通过保留原属性集的一个子集来减少属性集的大小不同的是, PCA通过创建一个替换的、较小的变量集合来“组合”属性的精华。

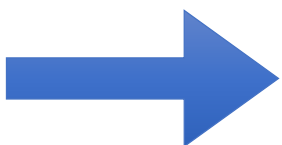
PCA举例

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9



PCA举例

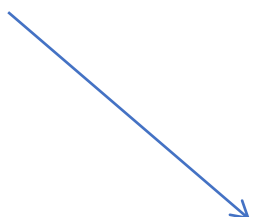
Data =	x	y	各维减去均值 	DataAdjust =	x	y
	2.5	2.4			.69	.49
	0.5	0.7			-1.31	-1.21
	2.2	2.9			.39	.99
	1.9	2.2			.09	.29
	3.1	3.0			1.29	1.09
	2.3	2.7			.49	.79
	2	1.6			.19	-.31
	1	1.1			-.81	-.81
	1.5	1.6			-.31	-.31
	1.1	0.9			-.71	-1.01

PCA举例

计算协方差矩阵

	x	y
	.69	.49
	-1.31	-1.21
	.39	.99
	.09	.29
DataAdjust =	1.29	1.09
	.49	.79
	.19	-.31
	-.81	-.81
	-.31	-.31
	-.71	-1.01

$$C = \begin{pmatrix} cov(x, x) & cov(x, y) & cov(x, z) \\ cov(y, x) & cov(y, y) & cov(y, z) \\ cov(z, x) & cov(z, y) & cov(z, z) \end{pmatrix}$$


$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

PCA举例

■ 计算协方差矩阵的特征值与特征向量

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$



$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

特征值和特征向量

定义 设 A 是 n 阶方阵, 若有数 λ 和非零向量 \mathbf{x} , 使得

$$A\mathbf{x} = \lambda\mathbf{x}$$

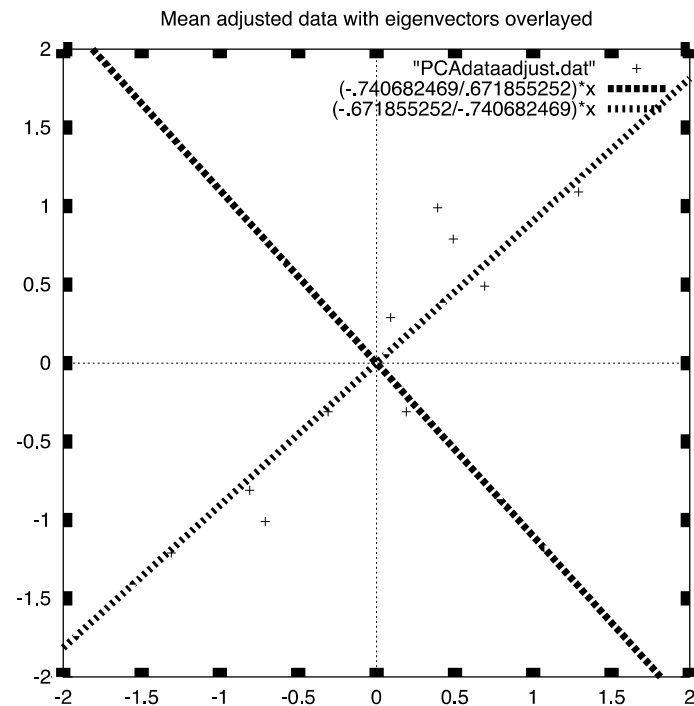
称数 λ 是 A 的特征值, 非零向量 \mathbf{x} 是 A 对应于特征值 λ 的特征向量。

例如 对 $A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$, 有 $\lambda = 3$ 及向量 $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, 使得 $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3 \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, 这说明 $\lambda = 3$ 是 A 的特征值, $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ 是 A 对应于 $\lambda = 3$ 的特征向量。

PCA举例

$$\text{FeatureVector} = (eig_1 \ eig_2 \ eig_3 \ \dots \ eig_n)$$

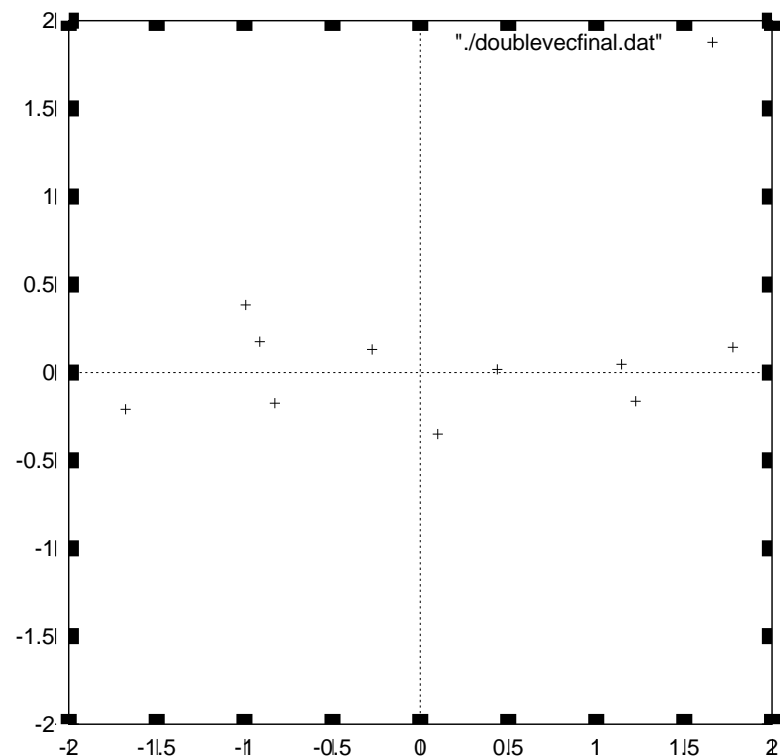
$$\text{FinalData} = \text{RowFeatureVector} \times \text{RowDataAdjust},$$



PCA举例

Transformed Data=

x	y
-.827970186	-.175115307
1.77758033	.142857227
-.992197494	.384374989
-.274210416	.130417207
-1.67580142	-.209498461
-.912949103	.175282444
.0991094375	-.349824698
1.14457216	.0464172582
.438046137	.0177646297
1.22382056	-.162675287



数量归约 (1) : 回归

■ 线性回归

- 对数据建模，使之拟合到一条直线

$$Y = w X + b,$$

- w, b 为回归系数;
- 回归系数可以用最小二乘法求解 the least-square method

■ 多元线性回归

- 允许响应变量 y 建模为两个或多个变量的线性函数

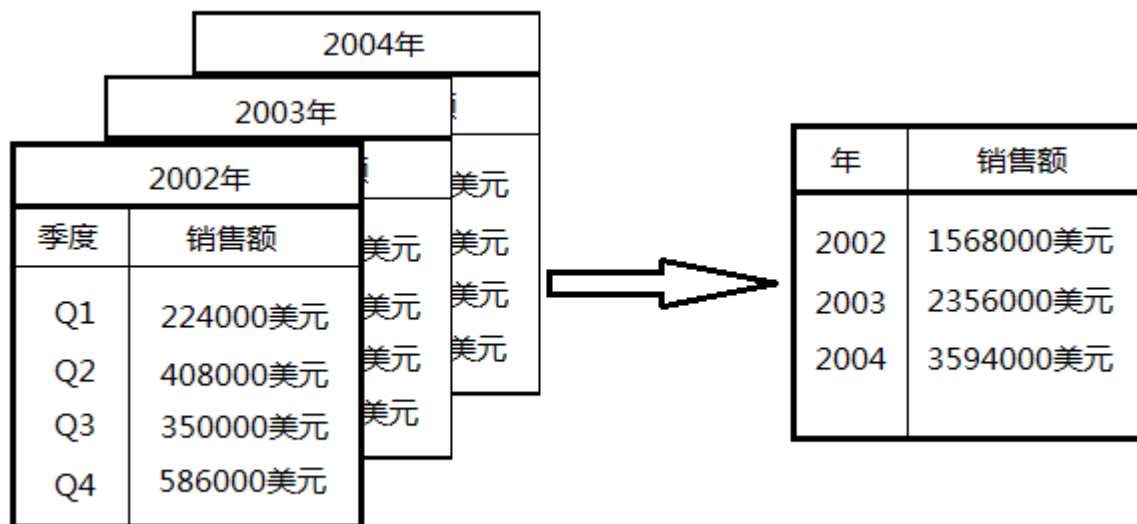
■ 对数线性模型 Log-linear model

- 近似离散的多维概率分布

数量归约 (2) : 数据立方体聚集

■ Data cube aggregation

- 对数据立方体用聚集操作，减少结果数据。也就是对数据进行汇总和聚集。



数量归约 (3) : 直方图

- 直方图使用分箱技术近似数据分布，属性A的直方图将A的数据划分为不相交的子集，或桶。
- 对于确定桶和属性值的划分，有如下划分规则：
 - 等宽：就是将桶的宽度区间设为一个常数，也就是横坐标；
 - 等深：就是将桶的频率设为一个常数（即每个桶含有相同个数的邻近样本），也就是纵坐标；
 - V-最优：给定桶个数，考虑所有可能的直方图，V-最优直方图是具有最小方差的直方图；直方图方差是每个桶代表的原来值的加权和，其中权等于桶中值的个数。
 - MaxDiff:考虑每对相邻值之间的差，桶的边界是具有 $n-1$ 个最大差的点。

数量归约 (4) : 抽样

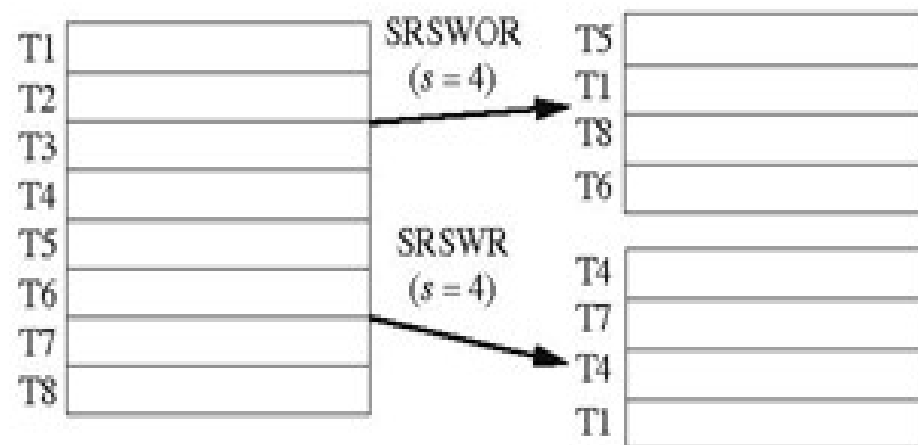
■ 用数据的较小随机样本（子集）表示大的数据集。

■ 无放回简单随机抽样 (SRSWOR)

■ 就是直接从 N 个元组中抽取 s 个样本，每个样本的概率为 $1/N$ 。

■ 有放回简单随机抽样 (SRSWR)

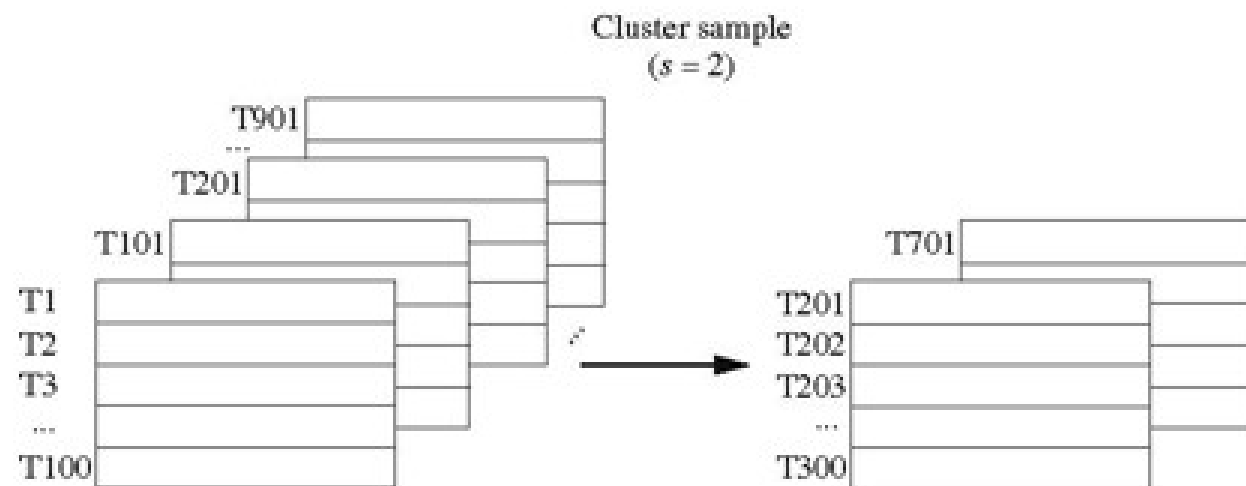
■ 就是一个元组被选择后，又把它放回去，以便它可以再次被选择。



数量归约 (4) : 抽样

- 聚类选样

- 如果D中的元组被分组放入M个不相交的“聚类”，则可以得到m个简单随机选样， $m < M$ 。例如，数据库中元组通常是一次取一页，这样每页可以看作一个聚类。



数量归约 (4) : 抽样

■ 分层选择

- 如果D被划分成互不相交的部分，称作层，则通过对每一层的简单随机选择就可以得到D的分层选择。
- 特别是当数据倾斜时，可以帮助确保样本的代表性。如对一个关于顾客的分层选择，可以按照年龄组创建分层，在每一层选样。

Stratified sample
(according to age)

T38	youth	T38	youth
T256	youth	T391	youth
T307	youth	T117	middle_aged
T391	youth	T138	middle_aged
T96	middle_aged	T290	middle_aged
T117	middle_aged	T326	middle_aged
T138	middle_aged	T69	senior
T263	middle_aged		
T290	middle_aged		
T308	middle_aged		
T326	middle_aged		
T387	middle_aged		
T69	senior		
T284	senior		

Outline

1

数据清理

2

数据集成与变换

3

数据归约

4

数据离散化和概念分层产生

离散化和概念分层生成

■ 离散化技术

- 将数值属性的值域划分为区间，用区间的标记替代实际的数据值，减少给定连续属性值的个数。

■ 概念分层

- 递归地收集标称数据高层的概念（如青年、中年或老年），并用它们替换较低层的概念（如年龄值）

自顶向下离散化与自底向上离散化

■ 自顶向下离散化

- 首先找出一点或几个点作为分列点，划分整个区间；
- 然后在结果区间上递归重复这个过程。
- 是从大区间开始逐渐变小的过程。

■ 自底向上离散化

- 首先将所有的连续值看成是可能的分列点，通过合并相邻值形成区间；
- 然后递归应用这一过程于结果区间。
- 是值合并为区间，小区间合并为大区间的过程。

离散化和概念分层生成主要方法

- 分箱
 - 自顶向下，非监督的
- 直方图分析
 - 自顶向下，非监督的
- 聚类分析
 - 自顶向下的划分策略或自底向上的合并策略，非监督的
- 基于熵的离散化
 - 自顶向下，监督的
- 基于聚类、决策树、相关分析的离散化
 - 自底向上，监督的
- 根据直观划分离散化：3-4-5规则
 - 自顶向下，非监督的

标称数据的概念分层生成

- 由用户或专家在模式级显式地说明属性的全序或部分序
 - 如数据库中可能包含如下属性组：street, city, province, country. 可以在模式级说明这些属性的全序，如
$$\text{street} < \text{city} < \text{province} < \text{country}$$
- 通过显式数据分组说明分层结构的一部分
 - $\{\text{Urbana, Champaign, Chicago}\} < \text{Illinois}$
- 说明属性集，但不说明它们的偏序
- 只说明部分属性集

自动概念分层

- 1) 根据每个属性的不同值个数，将属性按升序排列；
- 2) 按照排好的次序，自顶向下产生分层，第一个属性在最顶层，最后一个在最底层；
- 3) 用户考察所产生的分层，必要时，修改它以反映属性之间期望的语义联系。

小结

数据
清理

数据
归约

数据
集成

数据变换
与离散化

- 数据清理
 - 缺失值处理、噪声处理
- 数据集成与变换
 - 相同实体发现、冗余与冲突分析
 - 属性值规范化
- 数据归约
 - 维归约、数量归约
- 离散化与概念分层产生

尽管已经提出了一些数据预处理的方法，
数据预处理仍然是一个活跃的研究领域！

Data cleaning is one of the **three biggest problems** in data warehousing
—Ralph Kimball

- ◆ 预处理属于数据挖掘的基本步骤
- ◆ 涵盖基础概念和基本计算
- ◆ 建议认真领会，全面掌握。

