



目 录

CONTENTS

01

数据对象与属性类型

Data Objects and Attribute Types

02

数据的基本统计描述

Basic Statistical Descriptions of Data

03

数据可视化

Data Visualization

04

度量数据的相似性和相异性

Measuring Similarity and Dissimilarity

数据对象Data Objects

- 数据集由数据对象组成。
- 一个数据对象代表一个**实体(entity)**。
 - 销售数据库: 顾客, 商品, 销售
 - 医疗数据库: 患者
 - 大学数据库: 学生、教授、课程
- 数据对象又称为**样本、实例、数据点、或对象**。
- 数据对象用**属性(attribute)**描述。
- 若把数据集看作是数据库中的一张表, 数据行对应数据对象; 列对应属性。

属性Attributes

- 属性(attribute)是一个数据字段，表示数据对象的一个特征。
 - 如: *customer_ID, name, address*
- 类型:
 - 标称属性(nominal)
 - 二元属性(binary)
 - 序数属性(ordinal)
 - 数值属性(numeric)
 - 区间标度属性(interval-scaled)
 - 比率标度属性(ratio-scaled)

属性类型Attribute Types

- 标称属性(nominal attribute)
 - 其值是一些符号或者事物的名称。
 - 头发颜色= { 黑色, 棕色, 灰色, 白色 }
- 二元属性(binary attribute)
 - 又叫布尔 (bool) 属性
 - 是一种标称属性, 只有两个状态: 0或1。
 - 对称的(symmetric): 两种状态具有同等价值, 携带相同权重。
 - 如: 性别
 - 非对称的(asymmetric): 其状态的结果不是同样重要。
 - 如: 艾滋病毒的阳性和阴性结果。
 - 对重要的结果用1编码, 另一个用0编码。

属性类型Attribute Types

■ 序数属性(ordinal attribute)

- 其可能的值之间具有有意义的序或者秩评定(ranking), 但是相继值之间的差是未知的。
 - 成绩={优, 良, 中, 差}
- 其中心趋势可以用它的众数和中位数表示, 但不能定义均值。

■ 注意

- 标称、二元和序数属性都是定性的, 即只描述对象的特征, 不给出实际的大小。

属性类型Attribute Types

- 数值属性(numeric attribute)
 - 区间标度(interval-scaled)属性
 - 使用相等的单位尺度度量。
 - 值有序，可以评估值之间的差，不能评估倍数。
 - 没有绝对的零点。
 - 如:摄氏温度，华氏温度，日期
 - 比率标度(ratio-scaled)属性
 - 具有固定零点的数值属性。
 - 值有序，可以评估值之间的差，也可以说一个值是另一个的倍数。
 - 如：开式温温标(K)，重量，高度，速度

属性类型Attribute Types

- 属性的另一种分类方式
- 离散属性(discrete Attribute)
 - 具有有限或者无限可数个值。
 - 如： 邮编、省份数目具有有限个值， customer_Id是无限可数的。
 - 可以用或者不用整数表示。
- 连续属性(Continuous Attribute)
 - 属性值为实数。
 - 一般用浮点变量表示。



目 录

CONTENTS

01

数据对象与属性类型

Data Objects and Attribute Types

02

数据的基本统计描述

Basic Statistical Descriptions of Data

03

数据可视化

Data Visualization

04

度量数据的相似性和相异性

Measuring Similarity and Dissimilarity

概述

- 目的
 - 更好地识别数据的性质，把握数据全貌：中心趋势度量，数据散布
- 中心趋势度量(measures of central tendency)
 - 均值、中位数、众数、中列数
- 数据的散布(dispersion of the data)
 - 极差、四分位数极差、五数概括、盒图
- 数据可视化(graphic displays of basic statistical descriptions)
 - 分位数图、分位数-分位数图、直方图、散点图

中心趋势度量

■ 均值 (mean)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\bar{x} = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i}$$

■ 加权算术平均：每个值 x_i 与一个权重 w_i 相关联

■ 截尾均值：丢弃高低端极值

■ 中位数 (median)

■ 有序数值的中间值

■ 数据集的中位数可以通过插值(interpolation)估算

$$median = L_1 + \left(\frac{N / 2 - (\sum freq)_l}{freq_{median}} \right) \times width$$

L_1 : 中位数区间下界

N : 整个数据集中值的个数

$width$: 中位数区间的宽度

$(\sum freq)_l$: 低于中位数区间的所有区间频率和

$freq_{median}$: 中位数区间的频率

练习

- 设给定的数据集已经分组到区间，这些区间和对应频率如图。

计算该数据的近似中位数

- 确定中位数所在组

$$(\sum freq)_l / 2 = 3194 / 2 = 1597$$

$$950 < 1597 < 950 + 1500$$

因此中位数在21~50组

- 计算中位数

<i>age</i>	<i>frequency</i>
1 ~ 5	200
6 ~ 15	450
16 ~ 20	300
21 ~ 50	1500
51 ~ 80	700
81 ~ 110	44

$$median = 21 + \frac{3194 / 2 - 950}{1500} \times 29 = 33.508 \approx 34$$

中心趋势度量

■ 众数(mode)

- 数据集中出现频率最高的值
- 最高频率对应多个峰值，分为单峰的(unimodal), 双峰的(bimodal), 三峰的(trimodal)

<i>age</i>	<i>frequency</i>
1-5	200
6-15	450
16-20	300
21-50	1500
51-80	700
81-110	44

- 经验公式（单峰）：

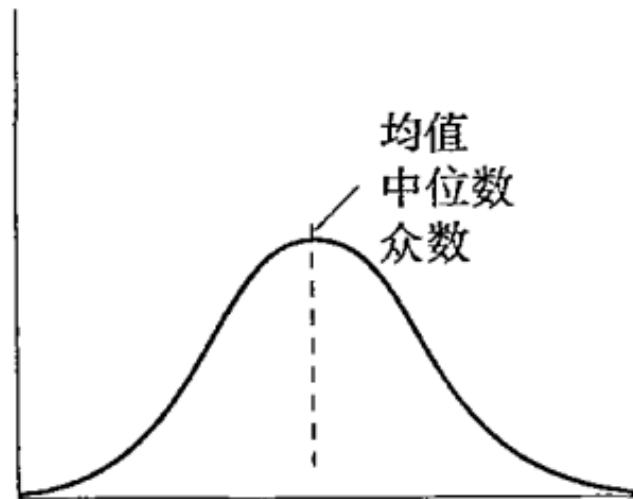
$$median - mode = 3 \times (mean - median)$$

■ 中列数(midrange)

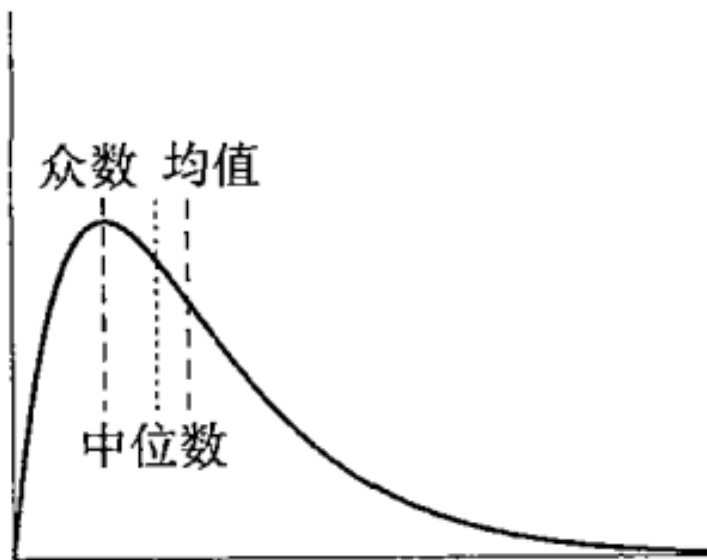
- 最大数和最小数的平均值

对称数据和非对称数据

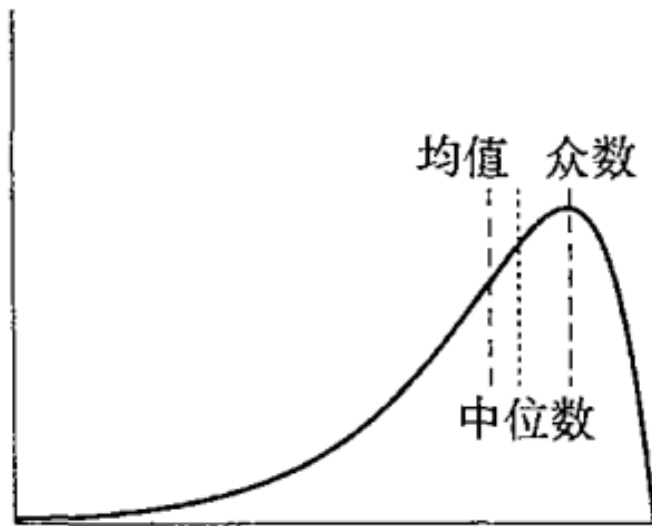
- 对称、正倾斜、负倾斜数据的中位数、均值和众数



a) 对称数据



b) 正倾斜数据



c) 负倾斜数据

度量数据的分散性

■ 分位数，离群点、盒图

- 四分位数 **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
- 四位分数极差 **Inter-quartile range**: $IQR = Q_3 - Q_1$
- 五数概括 **Five number summary**: min, Q_1 , median, Q_3 , max
- 盒图 **Boxplot**: 分布直观表示，体现五数概括
- 离群点 **Outlier**: 第三个四分位数之上或者第一个四分位数之下至少 $1.5 \times IQR$ 的值

■ 方差和标准差

- 方差 **Variance**:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right]$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

- 标准差 **Standard deviation** 方差的平方根

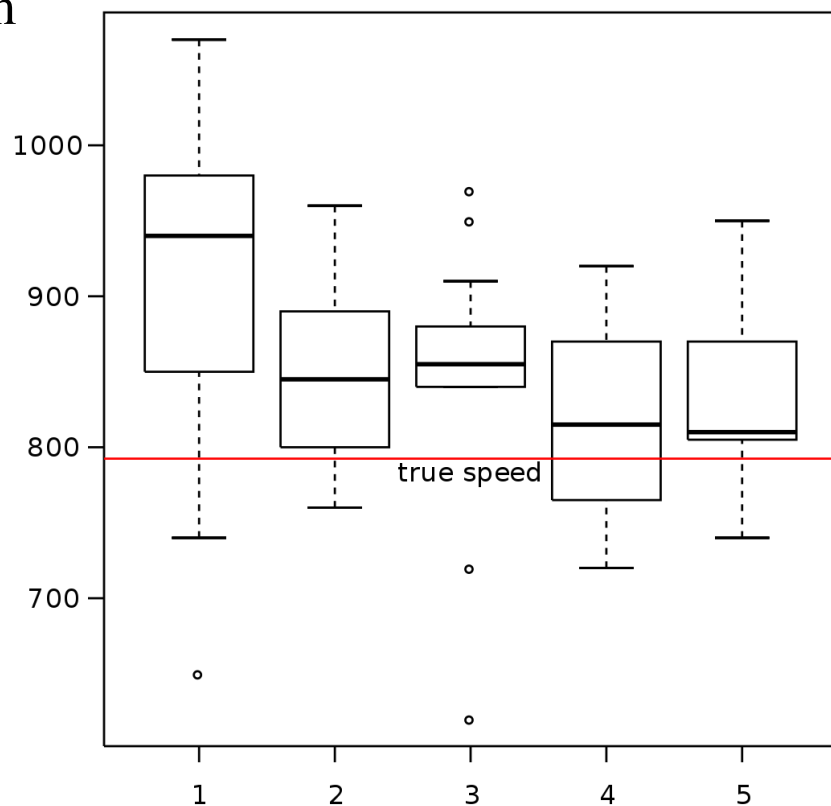
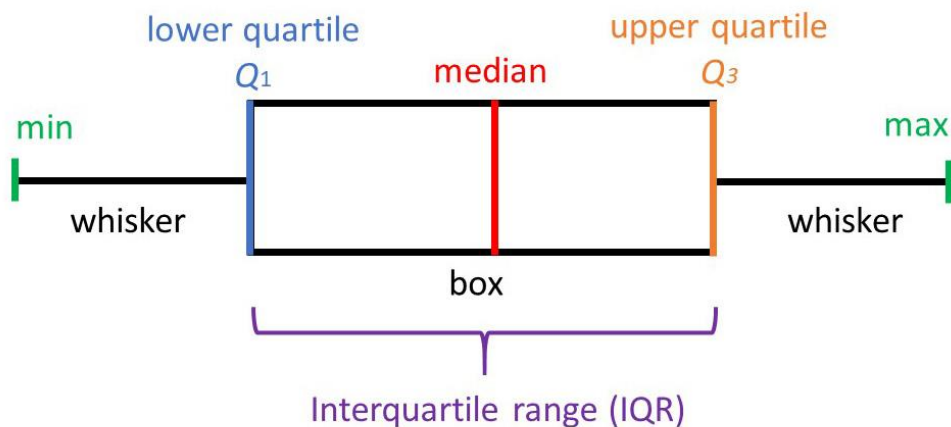
盒图

■ 五数概括

- Minimum, Q1, Median, Q3, Maximum

■ 盒图

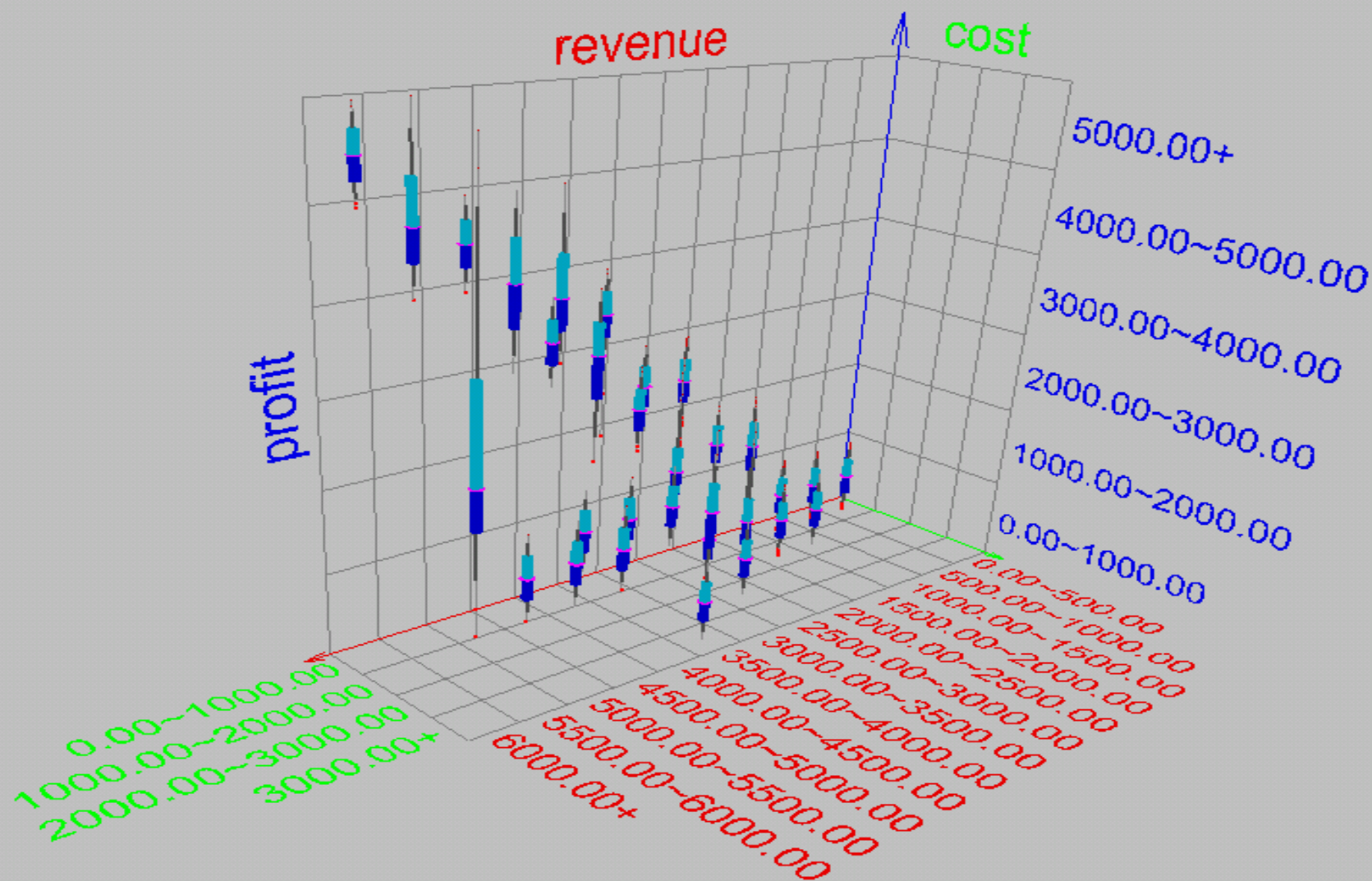
- 盒的端点在四分位数上，使得盒长度为四分位数极差IQR
- 中位数用盒内线标记
- 盒外线延伸到最小和最大的观测值



» 2 数据的基本统计描述



3-D盒图

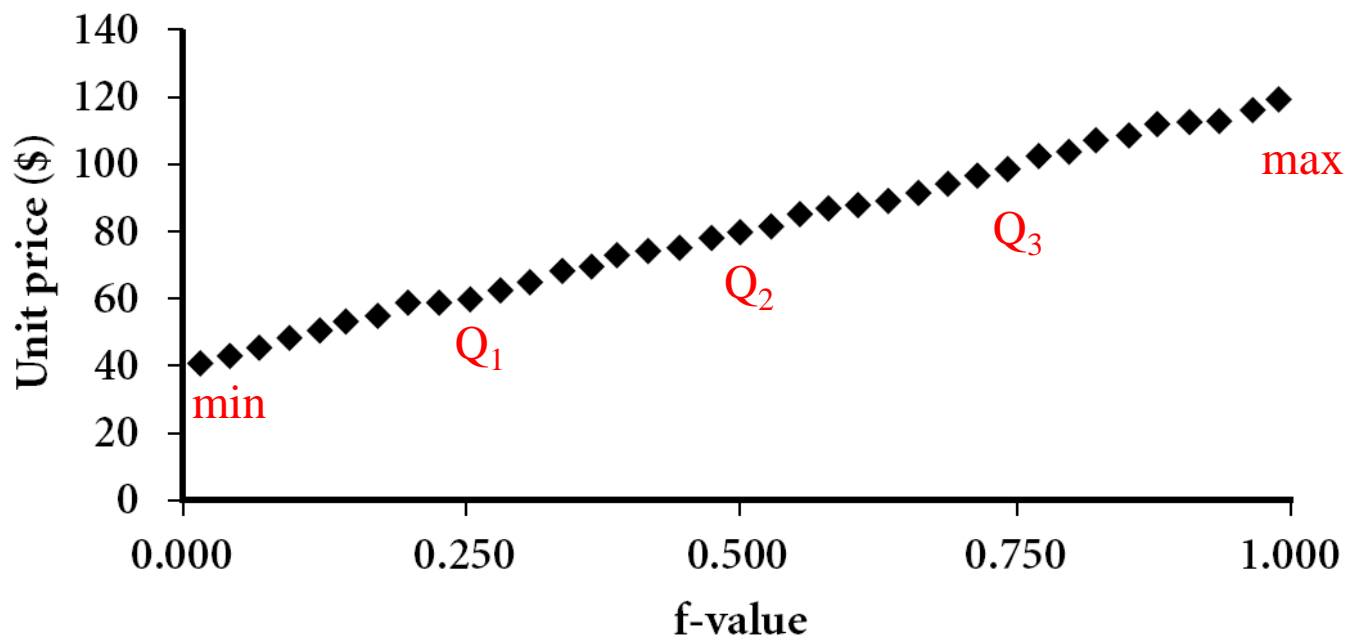


基本统计图

- 盒图 **Boxplot**: 五数概括
- 直方图 **Histogram**: x-axis 数值大小, y-axis 频率
- 分位数图 **Quantile plot**: 观测单变量数据分布, x_1 最小 x_n 最大
- 分位数-分位数图 **Quantile-quantile (q-q) plot**: 两个观测集, 观察一个分布到另一个分布是否漂移
- 散点图 **Scatter plot**: 每个值视作一个坐标对, 作为一个点画在平面上

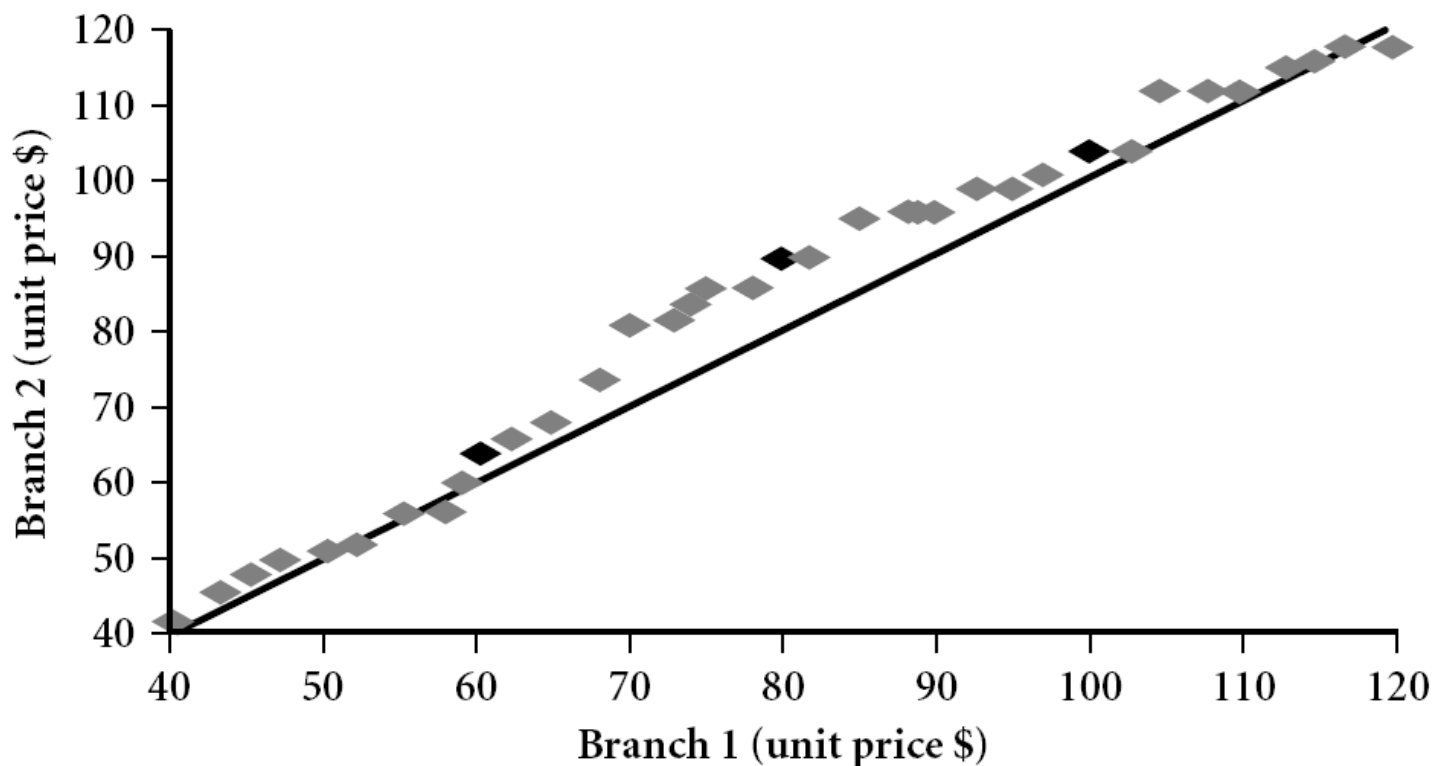
分位数图

- 显示给定属性所有数据
- 绘制分位数信息
- 增序排列，每个观测值 x_i 与一个百分数 f_i 配对，百分比0.5对应中位数，0.75对应Q3



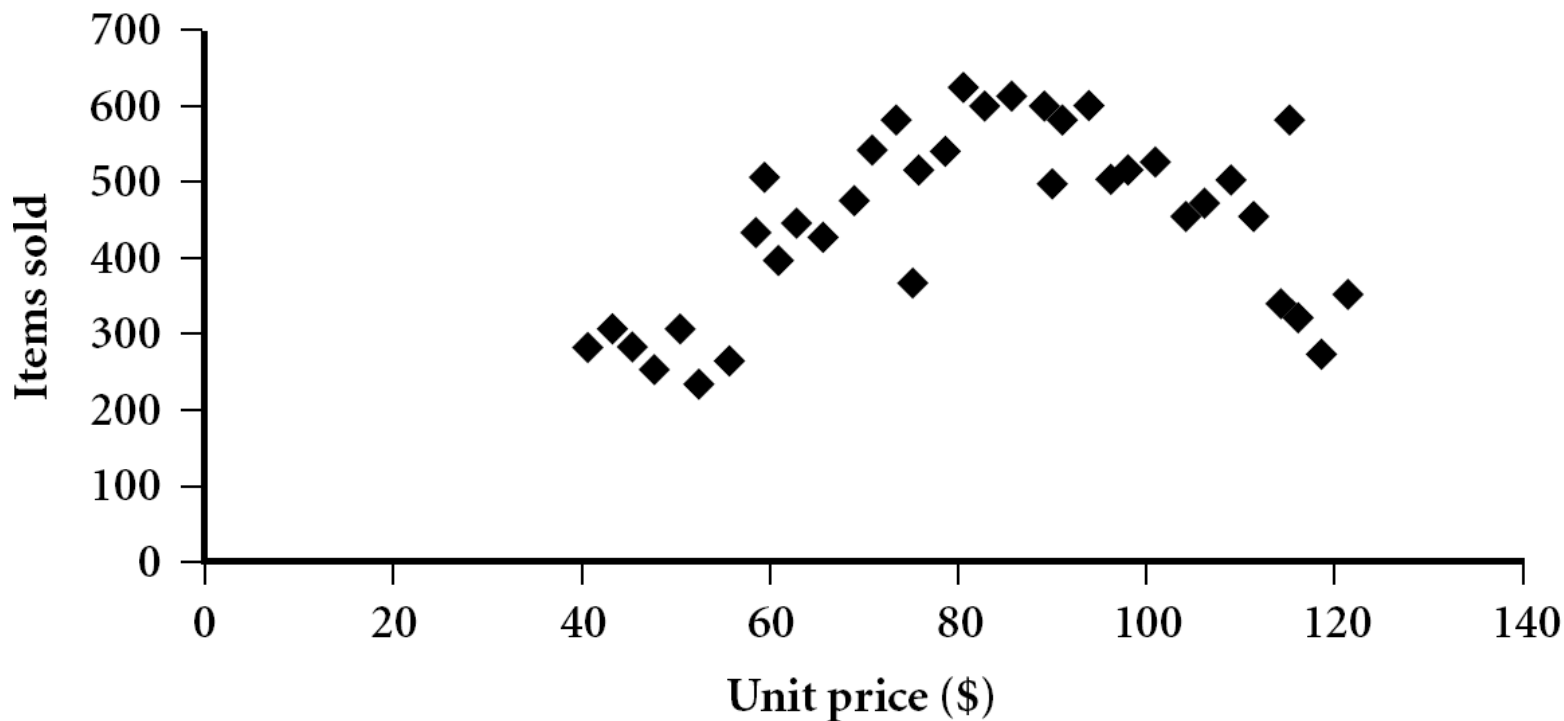
分位数-分位数图 (q-q图)

- 对着另一个对应的分位数，绘制一个单变量分布的分位数
- 使得用户可以观测从一个分布到另一个分布
- X, Y轴分别代表不同的观测集，存在两个观测集的值个数不一致时，不是所有的值都被表示

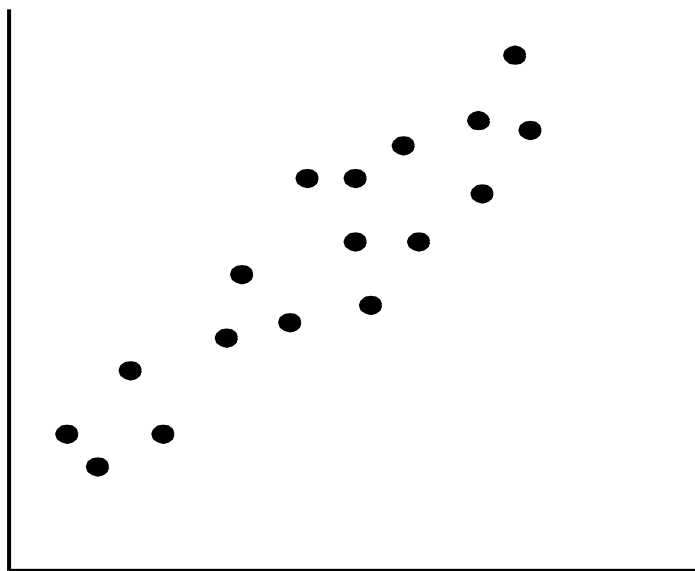


散点图

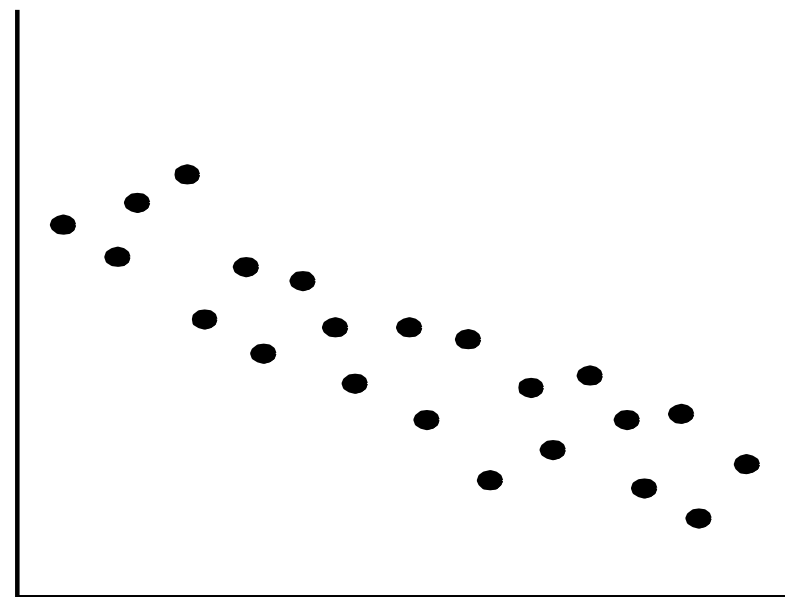
- 确定两个数值变量之间看上去是否存在联系
- 观察双变量数据的有用的方法



正相关和负相关的散点图



(a) 正相关

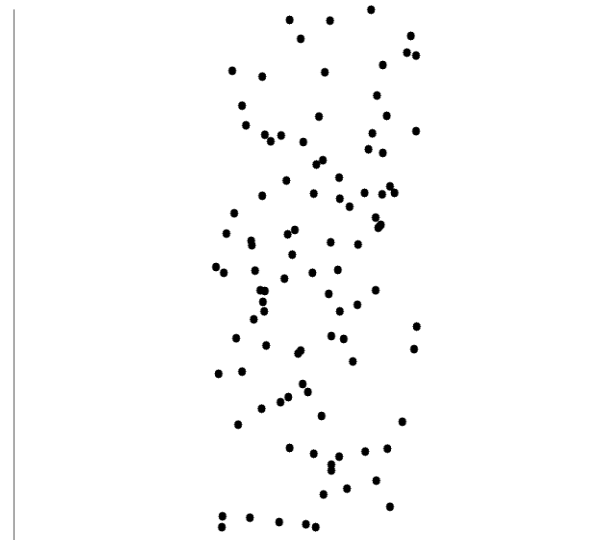
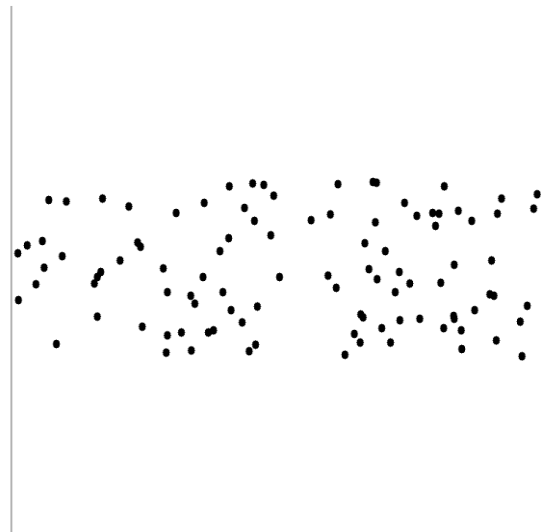
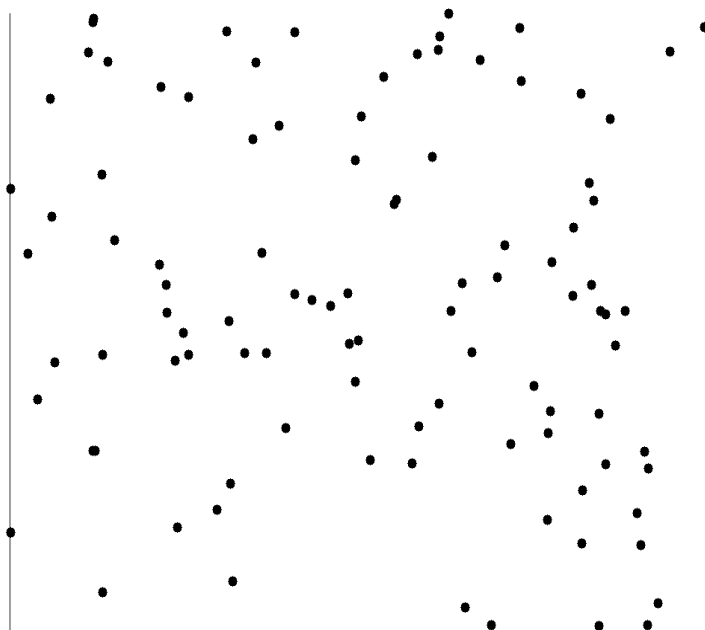


(b) 负相关

» 2 数据的基本统计描述



不相关的散点图





目 录

CONTENTS

01

数据对象与属性类型

Data Objects and Attribute Types

02

数据的基本统计描述

Basic Statistical Descriptions of Data

03

数据可视化

Data Visualization

04

度量数据的相似性和相异性

Measuring Similarity and Dissimilarity

数据可视化概述

■ 数据可视化意义

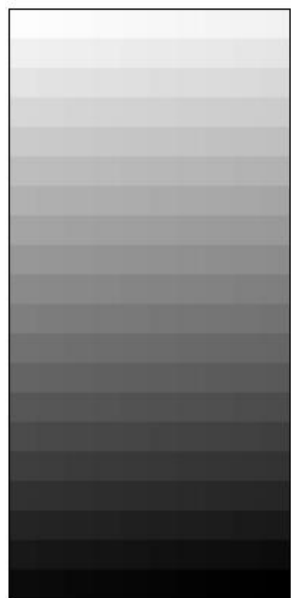
- 通过将数据映射在图元（graphical primitives）上来表示数据，便于深刻理解数据信息
- 便于对大型数据集进行定性描述（ qualitative overview ）
- 便于搜索数据间的模式（patterns），倾向（patterns），结构（structure），不规则性（structure）与联系性（relationships）
- 为进一步的定量分析找到合适的区间与变量

■ 数据可视化的技术

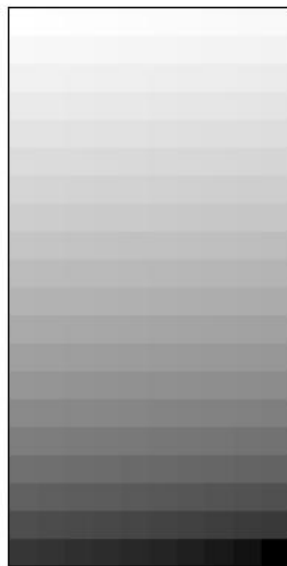
- 基于像素
- 几何投影
- 基于图符
- 层次可视化
- 可视化复杂对象与关系

基于像素可视化技术

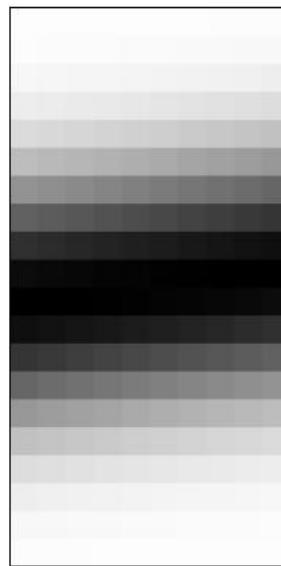
- 对于一个 m 维的数据集，在屏幕上创建 m 个窗口，每个窗口代表一个维度
- 记录的 m 个维值映射到这些窗口对应位置上的 m 个像素
- 像素的颜色反映相对应的值（ corresponding values ）



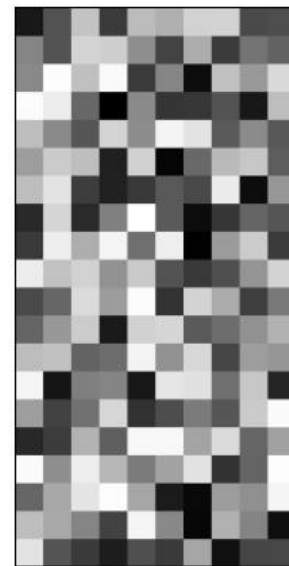
(a) Income



(b) Credit Limit



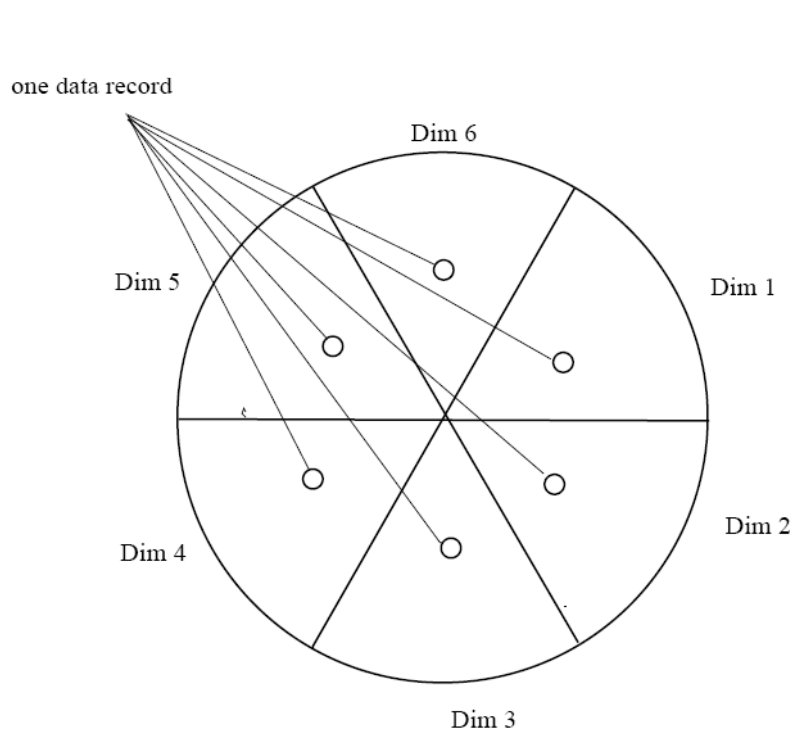
(c) transaction volume



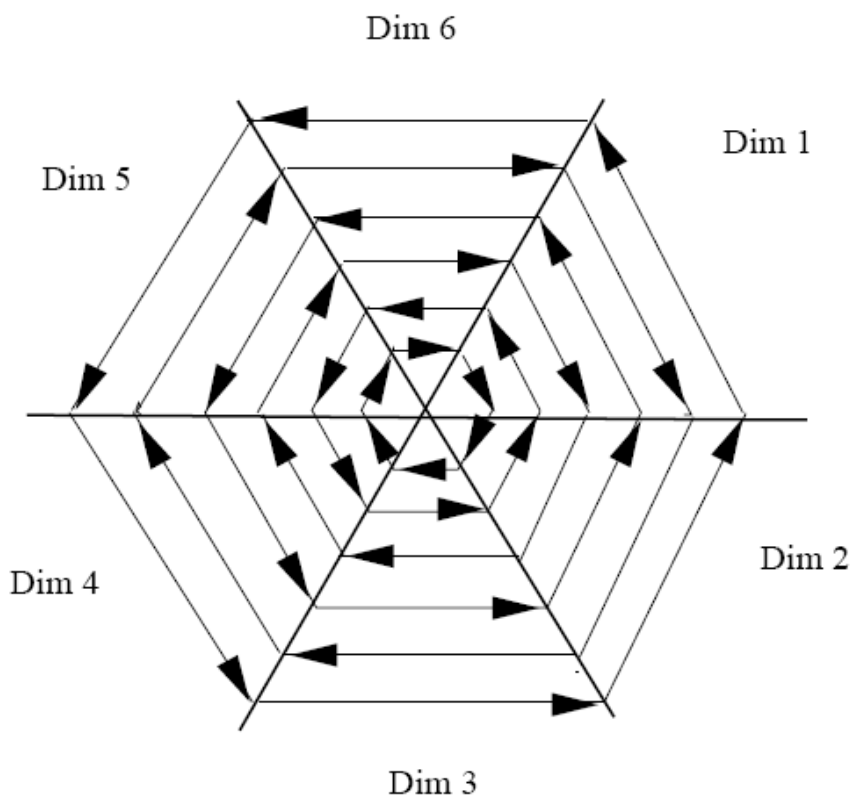
(d) age

圆弓分割技术

- 圆弓分割 (Circle Segment Technique) 是一种节约空间且简明扼要展示多维间关系的方法



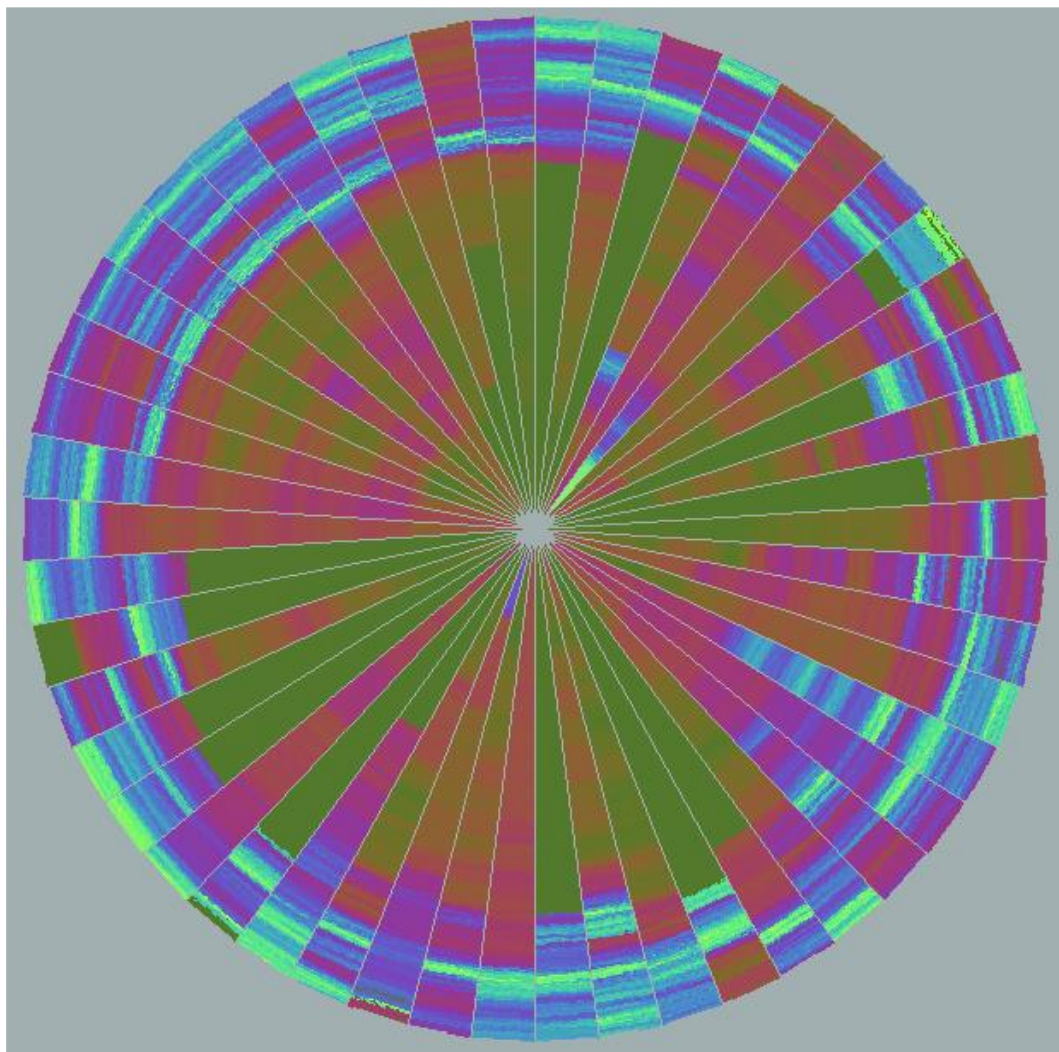
(a) 在圆弓内表示一个数据记录



(b) 在圆弓内安排像素

圆弓分割技术——示例

对265,000个50维的数据点进行可视化



图片来源：Ankerst, Mihael, Daniel A. Keim, and Hans-Peter Kriegel. "Circle segments: A technique for visually exploring large multidimensional data sets." In *Visualization*. 1996.

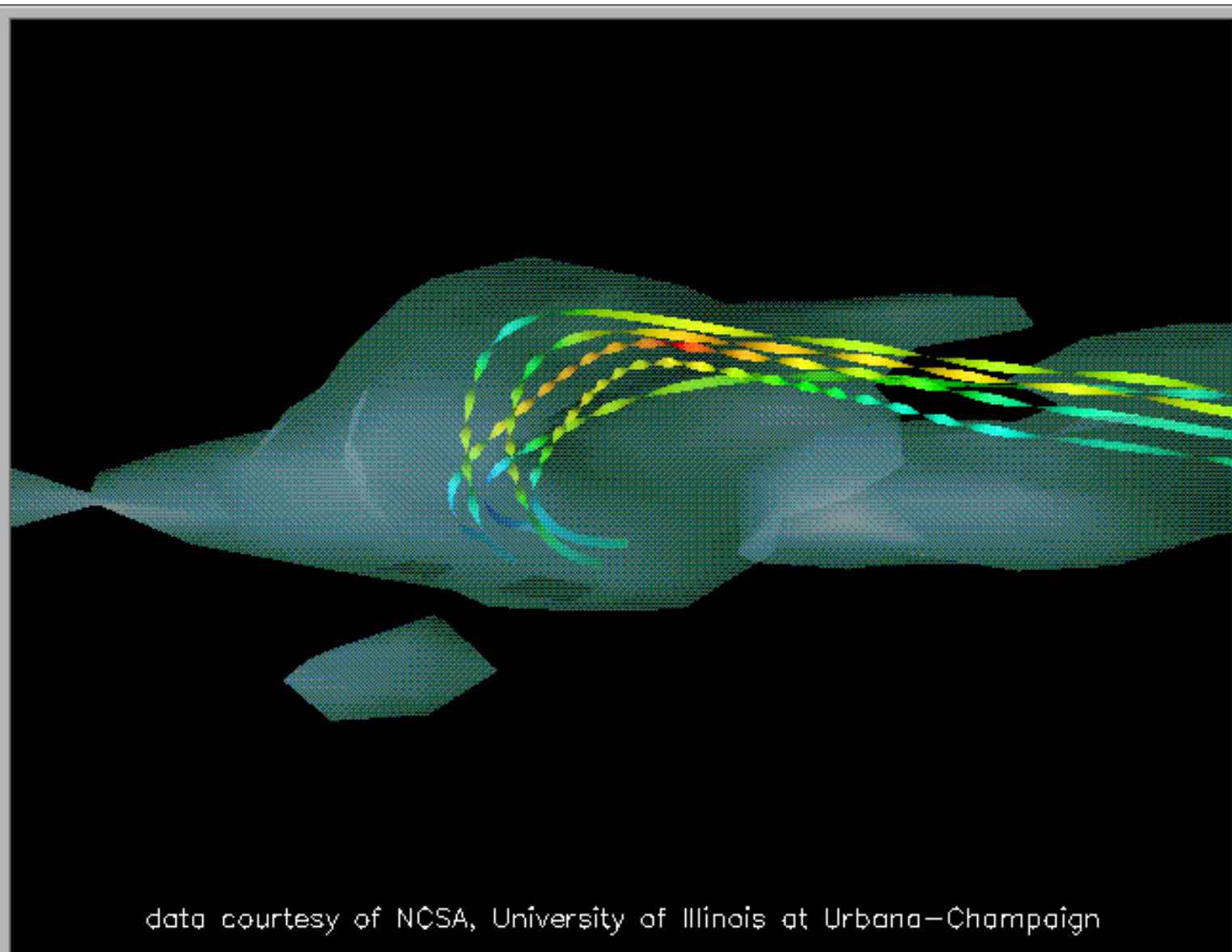
几何投影可视化

- 基于像素：对理解多维空间数据分布帮助不大，例如：不能显示在多维子空间是否存在稠密区域
- 将数据几何化，帮助用户发现多维数据在高维空间上的投影
- 技术
 - 直接投影
 - 散点图或散点图矩阵
 - 平行坐标

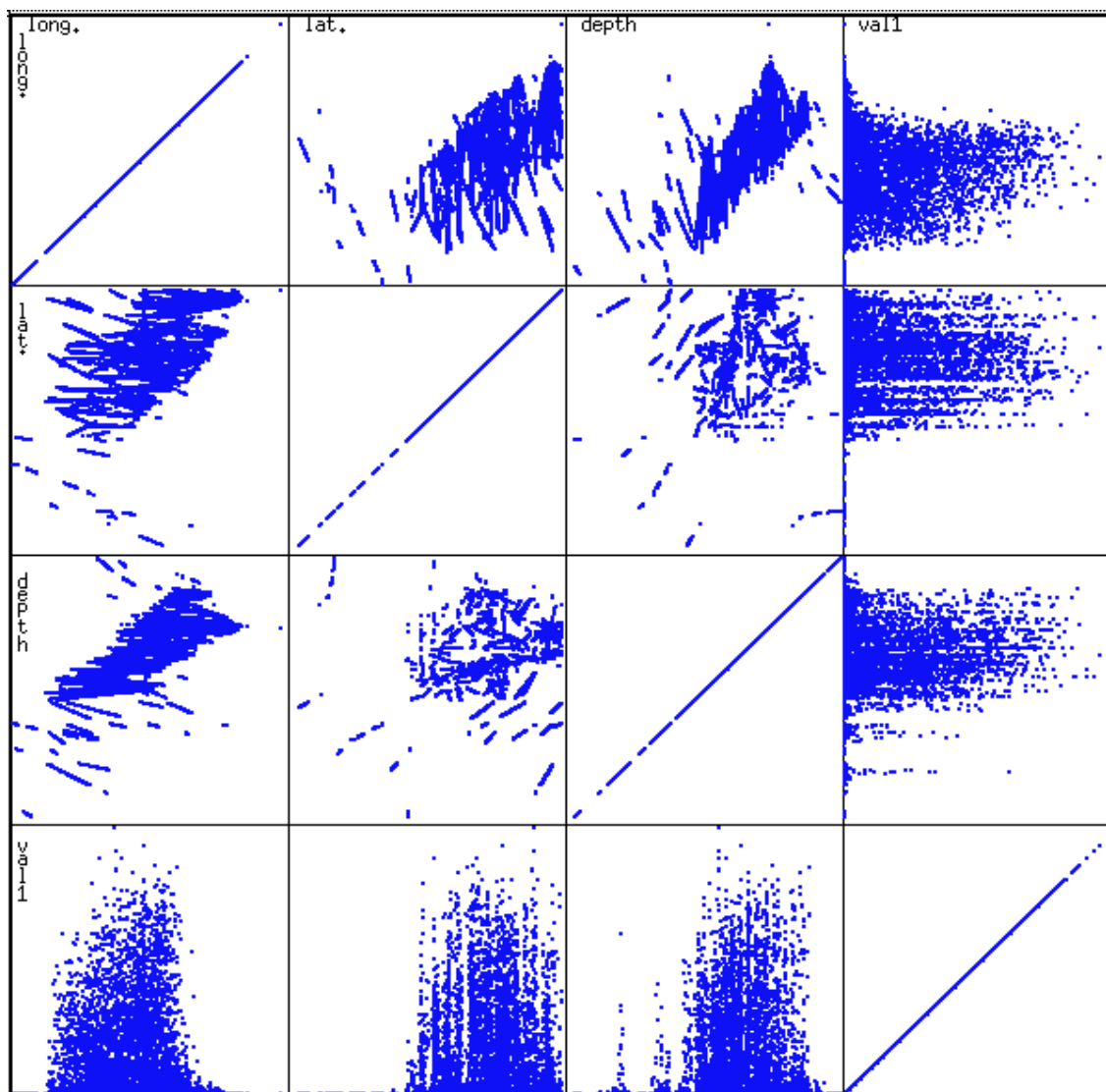
» 3 数据可视化

直接投影

Ribbons with Twists Based on Vorticity



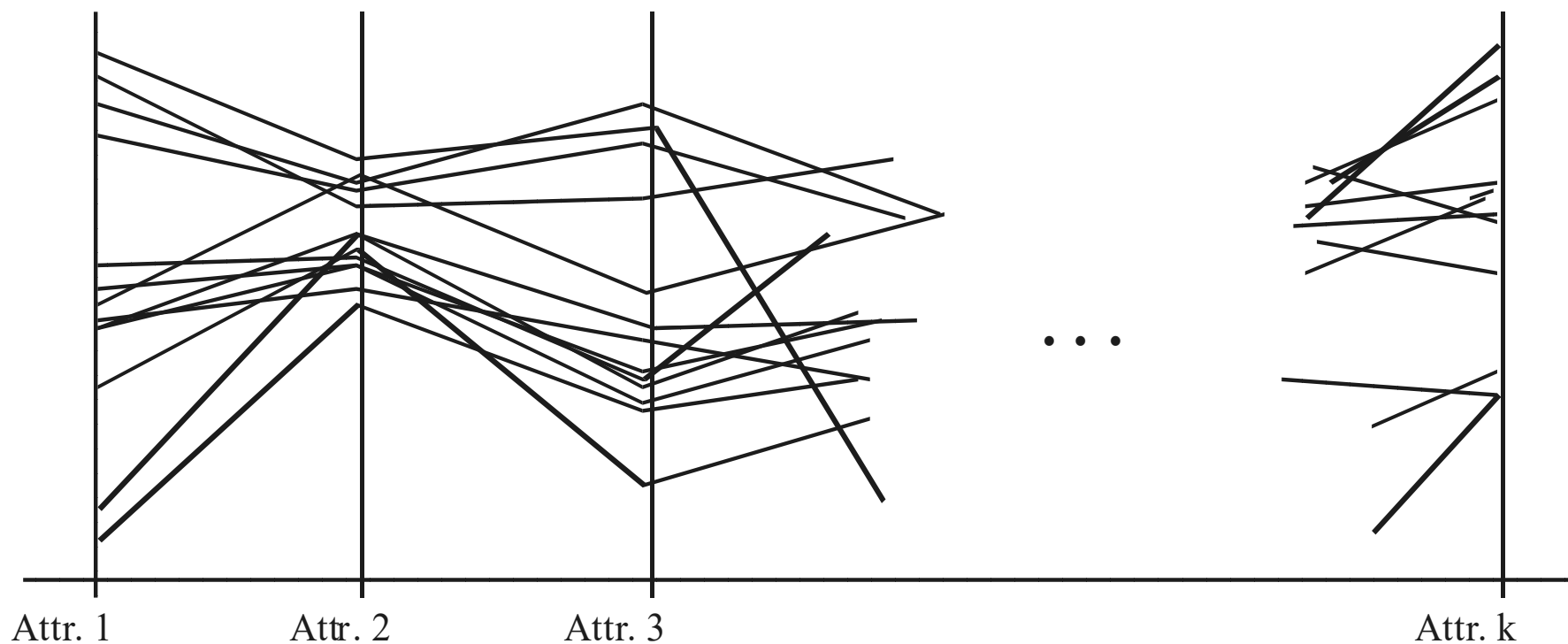
散点图矩阵



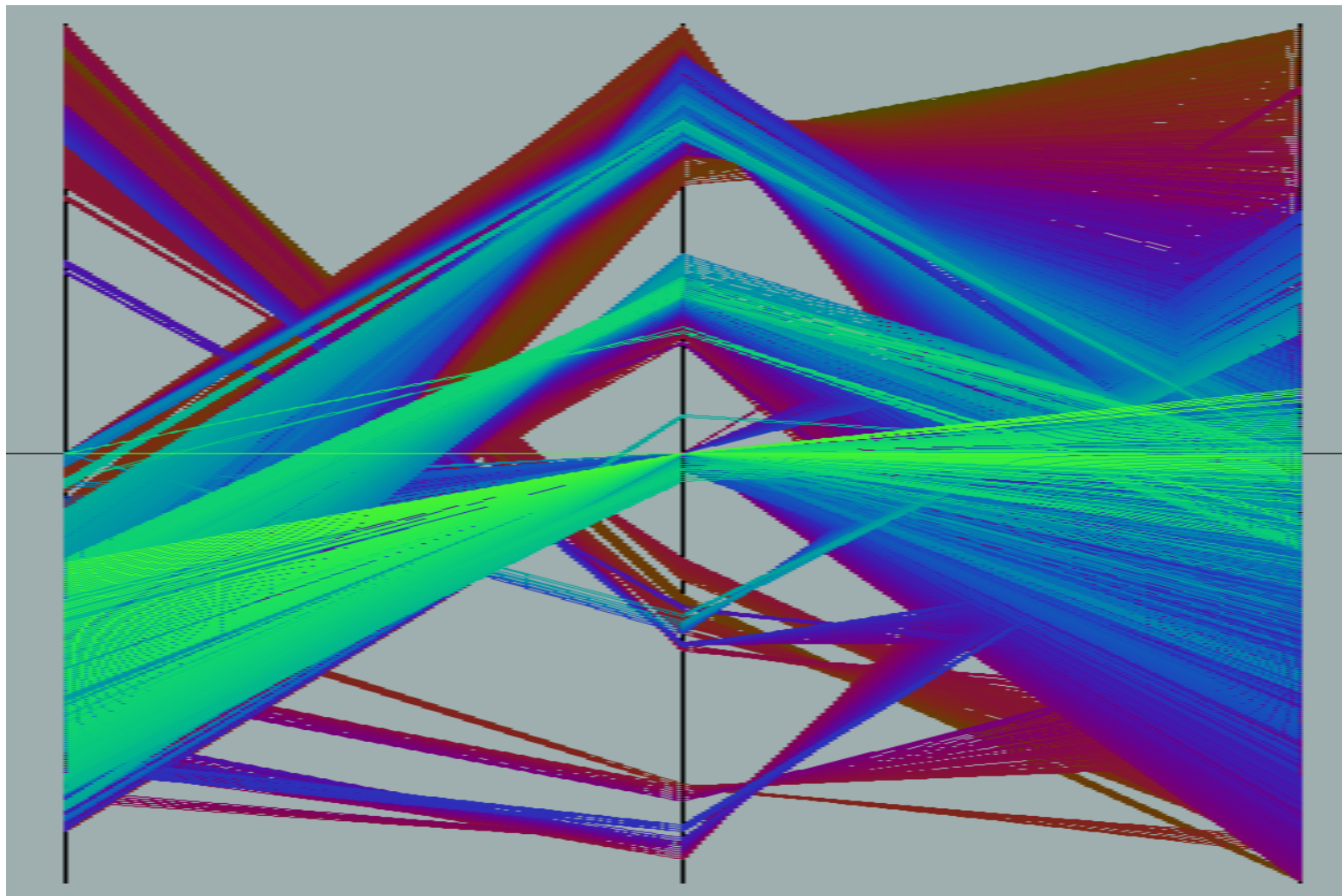
Matrix of scatterplots (x-y-diagrams) of the k -dim. data [total of $k*(k-1)$ scatterplots]

平行坐标

- 绘制 n 个等距离，相互平行的轴，每个代表一个维
- 数据记录用折线表示，与每个轴在对应相应维值的点上相交



平行坐标

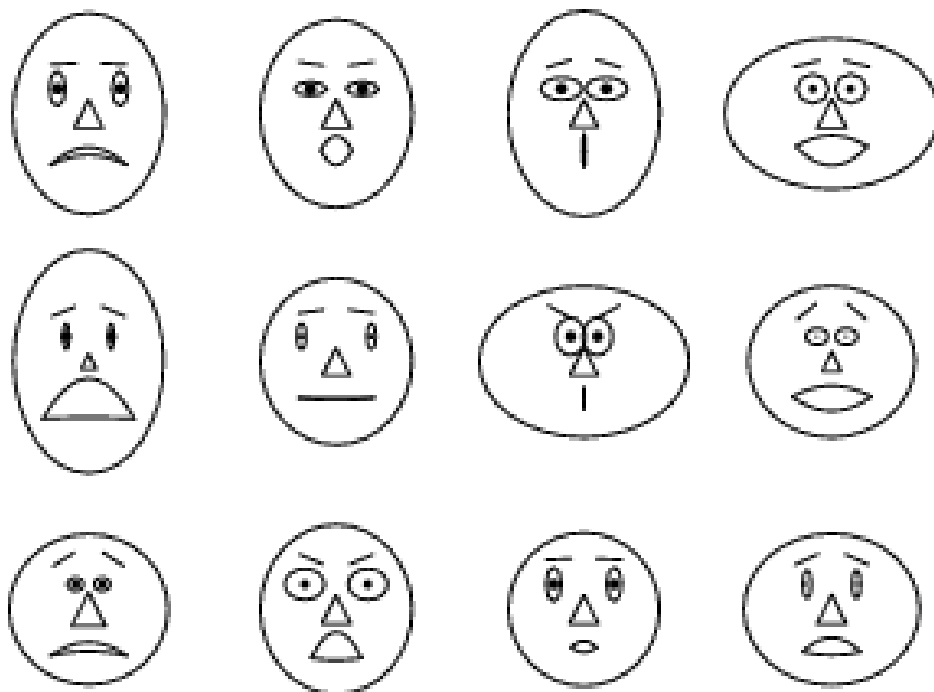


基于图符可视化技术

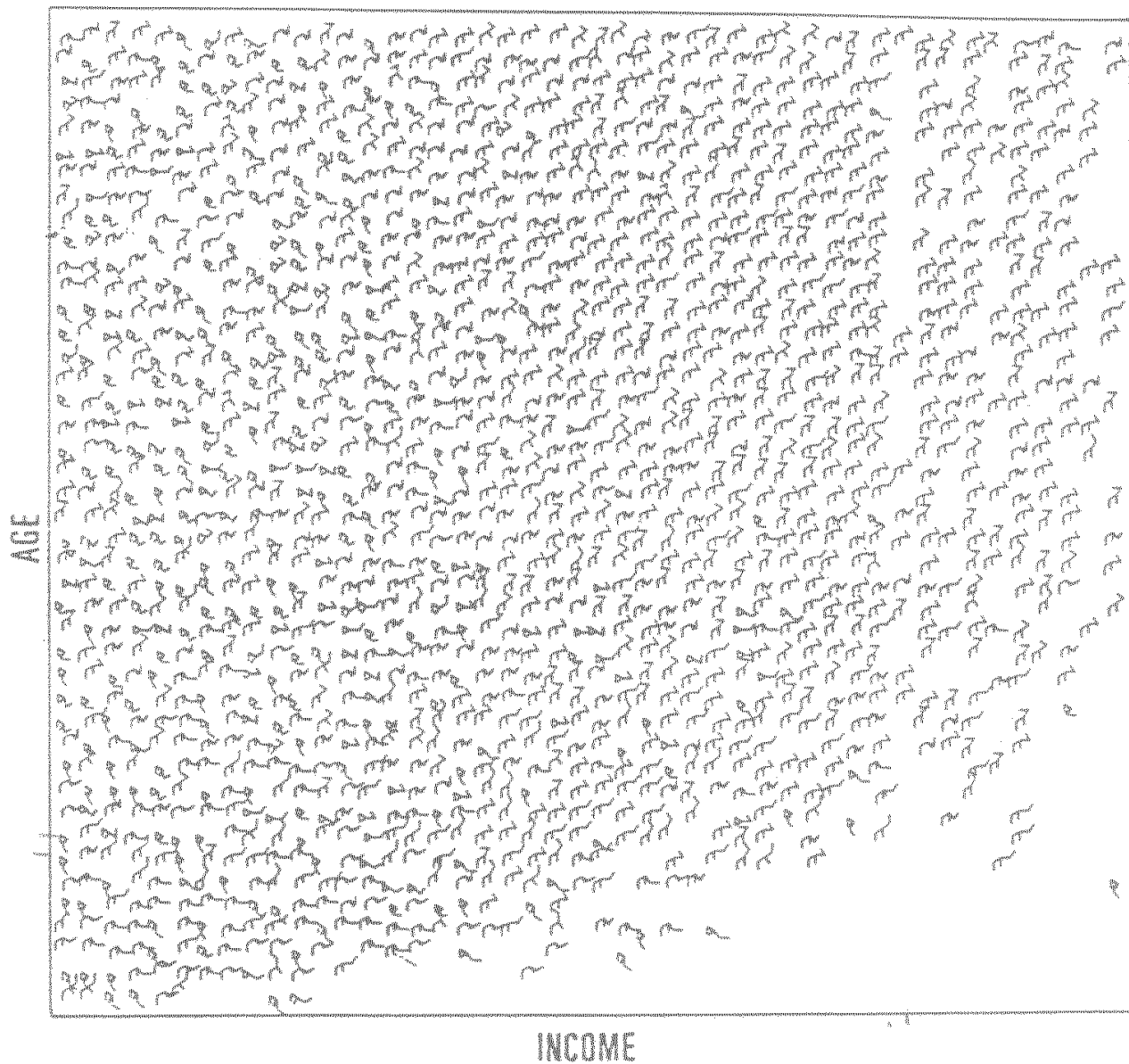
- 将数据值可视化为有不同特征的图符
- 代表技术
 - 切尔诺夫脸
 - 人物线条画

基于图符可视化技术

- 用二维的脸表示18维的多维数据（赫尔曼·切尔诺夫）
- 切尔诺夫脸利用脸的眼耳口鼻等要素的不同形状，大小，位置和方向代表维的值。利用人的思维能力，识别面部特征的微小差异来理解许多面部特征，有助于数据的规律性和不规则性的可视化。



人物线条画



X和Y轴映射两个维

用五段人物线条画表示其他维

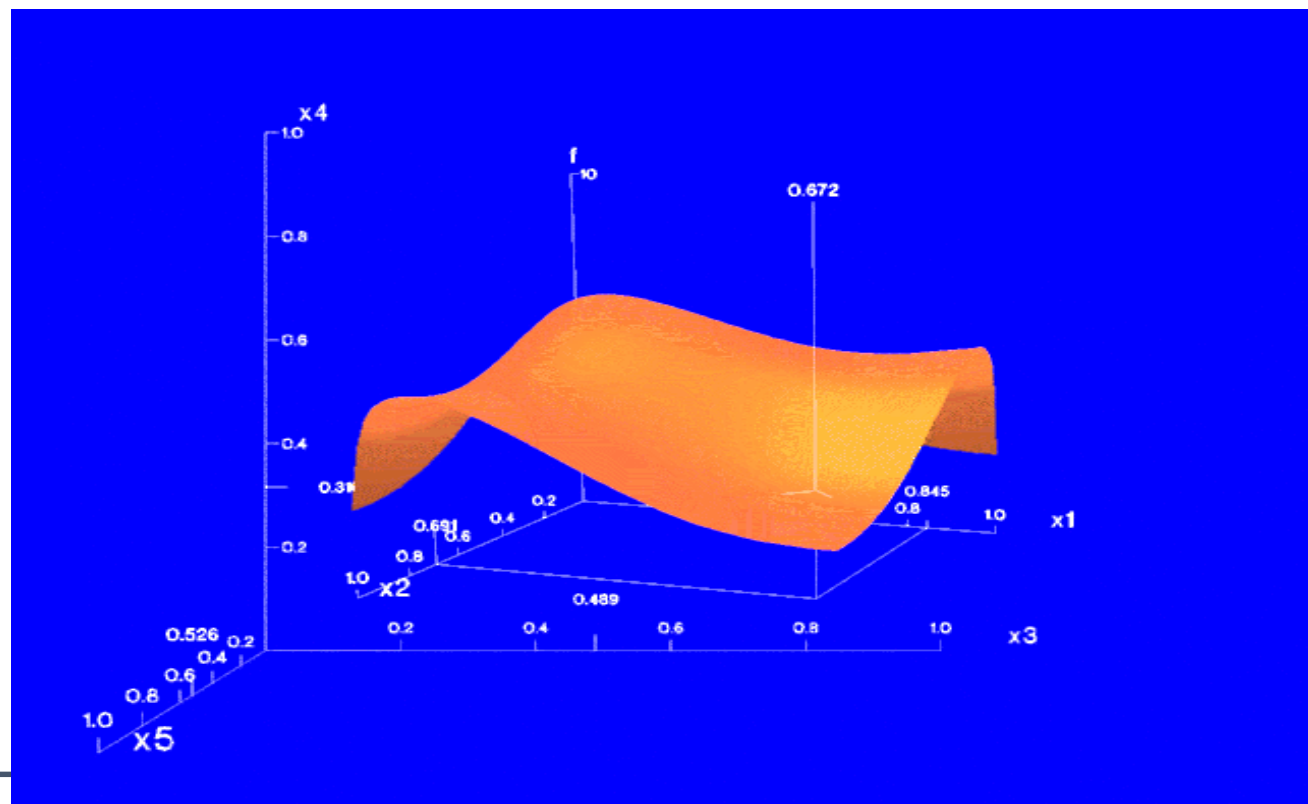
人口
统计
数据

层次可视化技术

- 把所有维划分成子集（子空间），子空间按层次可视化
- 方法
 - 世界中的世界（Worlds-within-Worlds）
 - 树图（Tree-map）

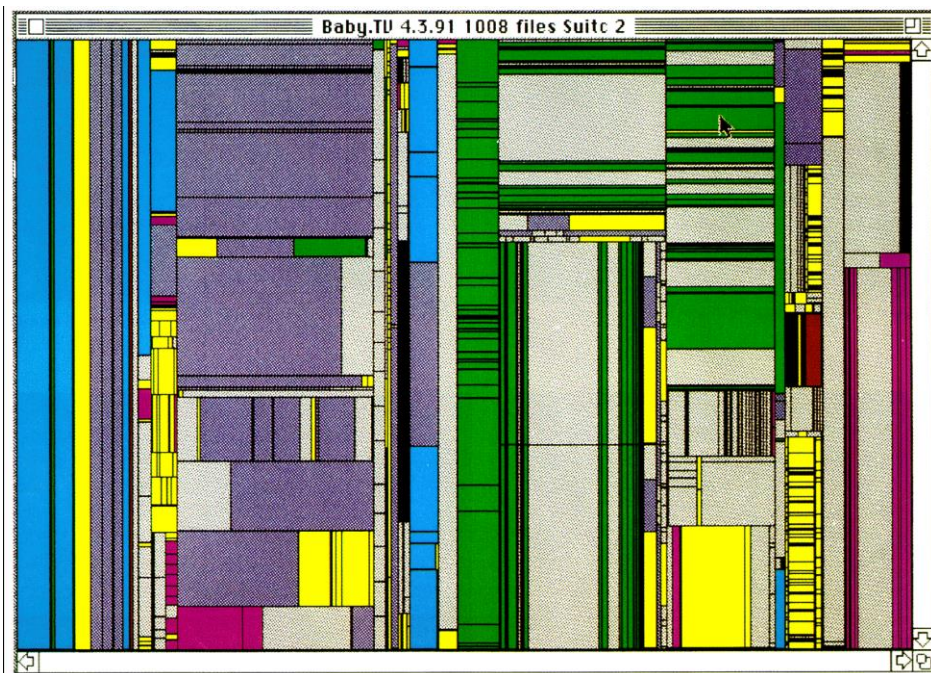
世界中的世界 (Worlds-within-Worlds)

- 世界中的世界(又称 n -Vision)
- 对六维数据集 (F, X_1, \dots, X_5) 可视化
- 把 X_3, X_4, X_5 作为固定值, 例如 (c_3, c_4, c_5) , 对另外三维可视化, 内世界的点位于外世界 (c_3, c_4, c_5) 处, 外世界是另一个三维图

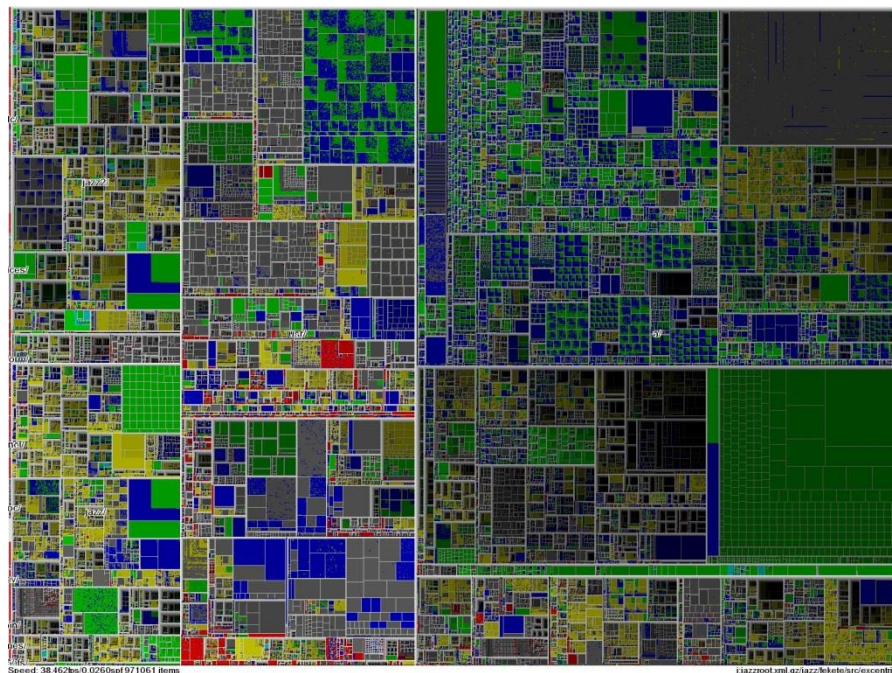


Tree-Map

- 把层次数据显示成嵌套矩形的集合



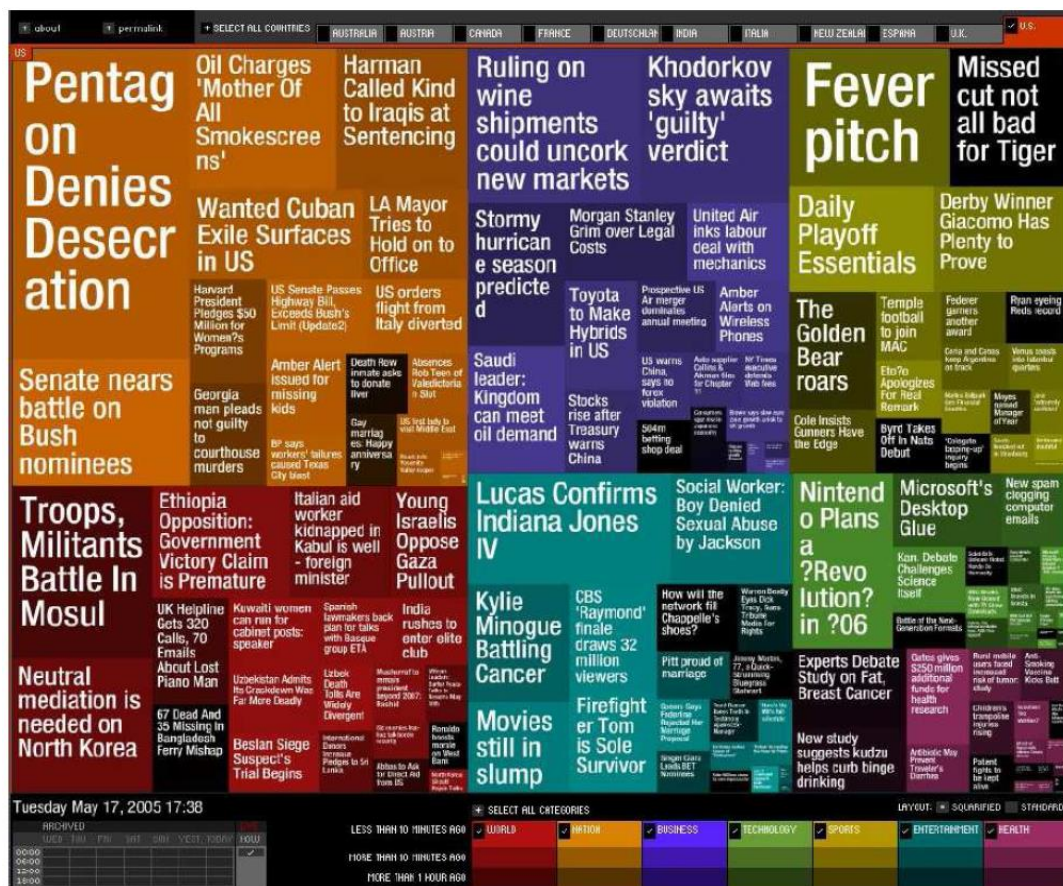
Schneiderman@UMD: Tree-Map of a File System



Schneiderman@UMD: Tree-Map to support large data sets of a million items

- 非数值数据的可视化: 文本与社交网络
- 标签云: 用户产生标签的统计量可视化

- 除了文本数据，还有用于可视化社交网络关系的技术



耐 劳 苦 尚 俭 朴
勤 学 业 爱 国 家



目 录

CONTENTS

01

数据对象与属性类型

Data Objects and Attribute Types

02

数据的基本统计描述

Basic Statistical Descriptions of Data

03

数据可视化

Data Visualization

04

度量数据的相似性和相异性

Measuring Similarity and Dissimilarity

概述

- 相似性(Similarity)
 - 两个对象相似程度的数量表示
 - 数值越高表明相似性越大
 - 通常取值范围为 $[0,1]$
- 相异性(Dissimilarity)(例如距离)
 - 两个对象不相似程度的数量表示
 - 数值越低表明相似性越大
 - 相异性的最小值通常为0
 - 相异性的最大值（上限）是不同的
- 邻近性(Proximity):相似性和相异性都称为邻近性

数据矩阵与相异性矩阵

■ 数据矩阵-对象-属性结构

- 行-对象: n个对象
- 列-属性: p个属性
- 二模矩阵(Two modes)

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

■ 相异性矩阵

(Dissimilarity matrix)

- n个对象两两之间的邻近度
- 对称矩阵
- 单模(Single mode)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

其中 $d(i,j)$ 表示对象i与对象j之间的相异性(距离)

标称属性的邻近性度量

- 标称属性(Nominal Attributes):可以取两个或多个状态
 - 例如: 颜色属性, 可以取值为: 红、黄、蓝、绿
- 两个对象*i*和*j*之间的相异性使用不匹配率来表示

$$d(i, j) = \frac{p - m}{p}$$

- *m*: 对象匹配数目, *p*: 对象的属性总数

表2.2 包含混合类型属性的样本数据表

对象标识符	Test-1 (标称的)	Test-2 (序数的)	Test-3 (数值的)
1	A	优秀	45
2	B	一般	22
3	C	好	64
4	A	优秀	28

只对标称属性test1计算相异性，
因此p=1，当对象i和j匹配时，
d(i,j)=0，当对象不同时d(i,j)=1

$$d(i, j) = \frac{p - m}{p}$$

0			
1	0		
1	1	0	
0	1	1	0

二元属性的邻近性度量

- 对象*i* 和对象*j* 的频数表

		对象 <i>j</i>		
		1	0	sum
对象 <i>i</i>	1	q	r	$q + r$
	0	s	t	$s + t$
sum		$q + s$	$r + t$	p

- 对称的二元相异性

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

- 非对称的二元相异性(*t*被认为不重要, 例如: 病理化验呈阴性)

$$d(i, j) = \frac{r + s}{q + r + s}$$

二元属性的邻近性度量

- Jaccard系数(非对称的二元相似性):

$$\text{sim}_{\text{Jaccard}}(i, j) = \frac{q}{q + r + s}$$

- Note: Jaccard系数与“一致性”计算相同:

$$\text{coherence}(i, j) = \frac{\text{sup}(i, j)}{\text{sup}(i) + \text{sup}(j) - \text{sup}(i, j)} = \frac{q}{(q + r) + (q + s) - q}$$

二元属性的相异性（例子）

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Name(姓名)是标称属性，Gender (性别)是对称二元属性
- 其他属性是非对称二元属性，假设只针对非对称二元属性进行相异性计算
- 值 Y 和 P 是 1, 值 N 是 0

数值属性的相异性:闵可夫斯基距离

- 闵可夫斯基距离(Minkowski Distance): 计算距离的通用的公式:

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

$i = (x_{i1}, x_{i2}, \dots, x_{ip})$ 和 $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ 是p维数据对象

- 距离需要满足的性质:
 - 非负性: $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$
 - 对称性: $d(i, j) = d(j, i)$
 - 三角不等式: $d(i, j) \leq d(i, k) + d(k, j)$
- 满足上述条件的测度称为度量(metric)

闵可夫斯基距离的特殊表现形式

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

- $h = 1$: 曼哈顿距离(或城市块距离 **Manhattan distance**)

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

- $h = 2$: 欧几里德距离(用的最多的)

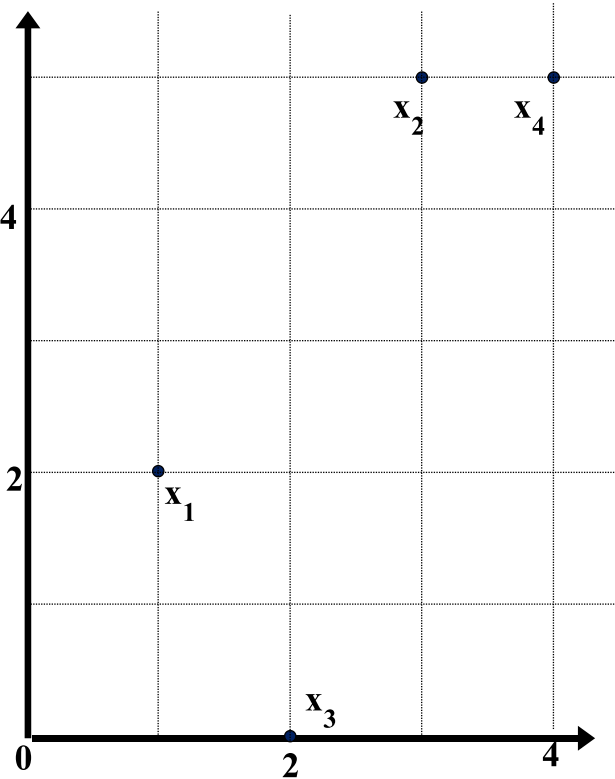
$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

- $h \rightarrow \infty$: 上确界距离 (又叫切比雪夫Chebyshev距离)
- 找出两个对象的属性中最大的距离

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{i,f} - x_{j,f}|^h \right)^{1/h} = \max_f |x_{i,f} - x_{j,f}|$$

例：闵可夫斯基距离

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



曼哈顿距离 (L_1)

相异性矩阵

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

欧氏距离 (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

上确界距离（切比雪夫距离）

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

序数属性的邻近性度量

- 序数属性：值之间是有意义的序或者排位
- 假设f为n个对象的一组序数属性之一，第i个对象的f值为 x_{if} ，属性f有 M_f 个有序状态，表示排位 $r_{if} \in \{1, \dots, M_f\}$

- 用下面公式实现数据规格化

$$z_{if} \in \frac{r_{if} - 1}{M_f - 1}$$

- 相异性计算可以用数值属性的距离度量来计算

序数属性的邻近性度量

对象标识符	Test-1 (标称的)	Test-2 (序数的)	Test-3 (数值的)
1	A	1.0	45
2	B	0.0	22
3	C	0.5	64
4	A	1.0	28

- $M=3$ ，把test2的每个值替换为它的排位，则4个对象将分别被赋值为3、1、2、3
- 实现规格化：将1映射为0.0，2映射为0.5，3映射为1.0
- 使用欧几里德距离求相异性矩阵

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

混合类型属性的相异性

- 数据库中可能包含各种属性类型
 - 标称的、对称二元的、非对称二元的、数值的或序数的
- 分别对每类数据进行数据挖掘分析，可能产生的结果不兼容
- 所有类型一起处理，公式为：

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

- 如果 x_{if} 或者 x_{jf} 缺失，即对象i或者对象j没有属性f的度量值，或者 $x_{if} = x_{jf} = 0$ ，并且f是非对称的二元属性，则 $\delta_{ij}^{(f)} = 0$
- 其他情况指示符 $\delta_{ij}^{(f)} = 1$

混合类型属性的相异性

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

- 若 f 是标称或二元的:
 - 如果 $x_{if} = x_{jf}$, 则 $d_{ij}^{(f)} = 0$, 否则 $d_{ij}^{(f)} = 1$

- 若 f 是数值的:

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$$

- 其中 h 遍取属性 f 的所有非缺失对象
- 若 f 是序数的:
 - 计算 r_{if} 和 z_{if} , 并将 z_{if} 作为数值属性对待。

$$z_{if} \in \frac{r_{if} - 1}{M_f - 1}$$


» 4 度量数据的相似性和相异性



混合类型属性的相异性

对象标识符	Test-1 (标称的)	Test-2 (序数的)	Test-3 (数值的)
1	A	优秀	45
2	B	一般	22
3	C	好	64
4	A	优秀	28

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$$



0			
0.55	0		
0.45	1.00	0	
0.40	0.14	0.86	0

Test1	Test2	Test3
$\begin{bmatrix} 0 \\ 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1.0 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0.55 & 0 \\ 0.45 & 1.00 & 0 \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$

对象标识符	Test-1 (标称的)	Test-2 (序数的)	Test-3 (数值的)
1	A	优秀	45
2	B	一般	22
3	C	好	64
4	A	优秀	28

对象1和对象4的最相似，对象1和对象2最不相似。

余弦相似性

- 对文档中的关键词或短语的频度表：

Document	teamcoach		hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- 词频向量通常很长，稀疏的，使用余弦相似性作为度量：

- $$\text{sim}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \bullet \mathbf{y}) / \|\mathbf{x}\| \|\mathbf{y}\| ,$$

其中：•表示向量积， $\|\mathbf{x}\|$:向量d的长度

例: 余弦相似性

- $\text{sim}(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\|$,

- 例: 求文档1与文档2的相似性

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

$$d_1 \bullet d_2 = 5*3 + 0*0 + 3*2 + 0*0 + 2*1 + 0*1 + 0*1 + 2*1 + 0*0 + 0*1 = 25$$

$$\|d_1\| = (5*5 + 0*0 + 3*3 + 0*0 + 2*2 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d_2\| = (3*3 + 0*0 + 2*2 + 0*0 + 1*1 + 1*1 + 0*0 + 1*1 + 0*0 + 1*1)^{0.5} = 4.12$$

- $\text{sim}(d_1, d_2) = 0.94$

练习

- 给定两个被元组(22, 1, 42, 10)和(20, 0, 36, 8)表示的对象

- (a)计算这两个对象之间的欧几里得距离。

$$\begin{aligned}d(i, j) &= \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \\ &= \sqrt{45} = 6.708\end{aligned}$$

- (b)计算这两个对象之间的曼哈顿距离。

$$\begin{aligned}d(i, j) &= |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}| \\ &= |22 - 20| + |1 - 0| + |42 - 36| + |10 - 8| = 11\end{aligned}$$

练习

- 给定两个被元组(22, 1, 42, 10)和(20, 0, 36, 8)表示的对象。
- (c)使用 $q=3$ ，计算这两个对象之间的闵可夫斯基距离。

$$\begin{aligned} d(i, j) &= \sqrt[3]{|x_{i1} - x_{j1}|^3 + |x_{i2} - x_{j2}|^3 + \cdots + |x_{ip} - x_{jp}|^3} \\ &= \sqrt[3]{8 + 1 + 216 + 8} = 6.15 \end{aligned}$$

- (d)计算这两个对象之间的上确界距离

$$d(i, j) = \max_f |x_{if} - x_{jf}| = 6$$



谢谢!

