第二章 认识数据

笔记本：	数据挖掘：概念与技术		
创建时间：	2017/12/20 15:39	更新时间：	2017/12/26 11:26
作者：	Passero		

2.1数据对象与属性类型

数据集由数据对象组成。一个数据对象代表一个实体。
e.g.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

属性：属性是一个数据字段，表示数据对象的一个特征。属性，维，特征和变量可以互换的使用。（通常的使用范围是：机器学习->特征 统计学家->变量 数据挖掘，数据库人士->属性）
e.g. 描述顾客的属性可能有：customer_ID name address
涉及的术语：
观测：给定属性的观测值
属性向量（或特征向量）：用来描述一个给定对象的一组属性
单变量分布：涉及一个属性的数据分布
双变量分布：涉及两个属性的数据分布

标称属性：标称属性的值是一些符号或事物的名称，每个值代表某种类别，编码或状态，因此标称属性被看做是分类的。
e.g. 假设hair_color和marital_status是两个描述人的属性，hair_color的取值可能有黑色，棕色，淡黄色，红色，赤褐色，marital_status的取值可以为单身，已婚，离异和丧偶。这两个都是标称属性。尽管标称属性的值是一些符号或事物的名称，但是可以用数来表示这些符号或名称。例如hair_color，可以指定代码0表示黑色，1表示棕色等。在标称属性之上，数学运算没有意义。与一个年龄值减去另外一个年龄值不同，一个顾客号减去另一个顾客号并没有意义，也就是说，尽管可以给标称属性取整数值，但是不能将其视为数值属性，因为不会定量的使用这些整数。

二元属性：二元属性是一种标称属性，只有两个类别或状态：0或1，其中0表示该属性不出现，而1表示出现。二元属性又称布尔属性，当两种状态对应于true和false的时候。若两种状态具有同等价值并且携带相同的权重，则二元属性是对称的，反之，则是非对称的。（比如描述性别时，便是对称的，当描述化验检查的阴性阳性时，比如艾滋病病毒化验结果通常会用1对最重要的结果[通常是稀有的]，比如阳性进行编码，而另外一个结果则用0）
e.g.假设属性smoker描述患者对象，1表示患者抽烟，0表示患者不抽烟。

序数属性：序数属性的可能的值之间具有有意义的序或者秩评定，但是相继值之间的差是未知的
e.g.职位可以按顺序枚举，如对教师有助教，讲师，副教授和教授，对于军阶有列兵，一等兵，专业军士，下士，中士等
对于记录不能客观度量的主观质量评估，序数属性是有用的，因此序数属性通常用于等级评定调查，如调查顾客满意程度（用0,1,2,3,4来表示满意度）
序数属性的中心趋势可以用它的众数和中位数（有序序列的中间值）表示，但不能定义均值。

总结：标称，二元和序数属性都是定性的，也就是说，它们描述对象的特征，而不给出实际大小或数量。这种定性属性的值通常是代表类别的词。如果使用整数，那么它们代表类别的计算机编

码，而不是可测量的量。（例如，0表示小杯饮料，1表示中号杯，2表示大杯）

数值属性：数值属性是定量的，即它是可度量的量，用整数或实数值表示。数值属性是可以区间标度或比率标度的。
---区间标度属性：用相等的单位尺度度量，区间属性值是有序的，可以为正，0或负，因此除了值的秩评定之外，这种属性允许我们比较和定量评估值之间的差。
e.g.温度，日历日期（我们不能用比率谈论这些值，比如我们不能说10度比5度温暖2倍）
---比率标度属性：比率标度是具有固定零点的数值属性，也就是说，如果度量是比率标度的，便可以说一个值是另一个的倍数，而且这些值是有序的，因此可以计算值之间的差，均值，中位数和众数。
e.g.重量，高度，速度，货币量

离散属性与连续属性：
---离散属性：具有有限或无限可数个值，可以用或不用整数表示。如hair_color等都有有限个值，因此是离散的。离散属性可以具有数值值，如二元属性取0和1。如果一个属性的可能的值集合是无限的，但是可以建立一个与自然数的一一对应，则这个属性是无限可数的，入邮政编码，如用户ID，顾客数量是无限增长的，但事实上实际的值集合是可数的（可以建立这些值与证书集合的一一对应）
---连续属性：若属性不是离散的，则它是连续的。连续属性一般用浮点变量表示。

2.2数据的基本统计描述

中心趋势度量：均值，中位数和众数
---均值mean：（算术）均值，加权算术均值（或加权平均，也就是每个取值与一个权重相关联的情形）
Notes: 均值对极端值（例如：离群点）很敏感，例如公司平均薪水可能被少数几个高收入的人拉高。因此可以使用截尾均值，即丢弃高低极端值后的均值，但应避免截取太多，因为这样可能会丢失有价值的信息。
---中位数median：有序数据值的中间值

当观测的数量很大时，中位数的计算开销很大。然而，对于数值属性，我们可以很容易计算中位数的近似值。假定数据根据它们的 x_i 值划分成区间，并且已知每个区间的频率（即数据值的个数）。例如，可以根据年薪将人划分到诸如 10 000 ~ 20 000 美元、20 000 ~ 30 000 美元等区间。令包含中位数频率的区间为中位数区间。我们可以使用如下公式，用插值计算整个数据集的中位数的近似值（例如，薪水的中位数）：

$$median = L_1 + \left(\frac{N/2 + (\sum freq)_l}{freq_{median}}\right) width \qquad (2.3)$$

其中，L_1 是中位数区间的下界，N 是整个数据集中值的个数，$(\sum freq)_l$ 是低于中位数区间的所有区间的频率和，$freq_{median}$ 是中位数区间的频率，而 $width$ 是中位数区间的宽度。

---众数mode：数据集的众数是集合中出现最频繁的值。可能最高频率对应对个不同值，导致多个众数。具有一个，两个，三个众数的数据集合分别称为单峰的，双峰的和三峰的。一般而言，具有两个及以上众数的数据集是多峰的，如果每个数据值仅出现一次，则没有众数。

对于适度倾斜（非对称）的单峰数值数据，我们有下面的经验关系

$$mean - mode \approx 3 \times (mean - median) \quad (2.4)$$

这意味：如果均值和中位数已知，则适度倾斜的单峰频率曲线的众数容易近似计算。

中列数（midrange）也可以用来评估数值数据的中心趋势。中列数是数据集的最大和最小值的平均值。中列数容易使用SQL的聚集函数 **max()** 和 **min()** 计算。

在具有完全**对称的**数据分布的单峰频率曲线中，均值、中位数和众数都是相同的中心值，如图2.1a所示。

在大部分实际应用中，数据都是不对称的。它们可能是**正倾斜的**，其中众数出现在小于中位数的值上（见图2.1b）；或者是**负倾斜的**，其中众数出现在大于中位数的值上（见图2.1c）。

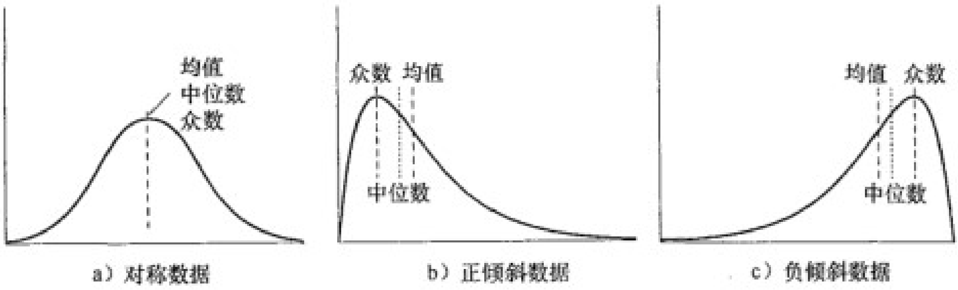


图2.1 对称、正倾斜和负倾斜数据的中位数、均值和众数

度量数据散布：极差，四分位数，方差，标准差和四分位数极差

---极差range：集合的极差是最大值和最小值之差

设 $x_1, x_2, \cdots, x_N$ 是某数值属性 X 上的观测的集合。该集合的**极差**（range）是最大值（**max()**）与最小值（**min()**）之差。

假设属性 X 的数据以数值递增序排列。想象我们可以挑选某些数据点，以便把数据分布划分成大小相等的连贯集，如图2.2所示。这些数据点称做分位数。**分位数**（quantile）是取自数据分布的每隔一定间隔上的点，把数据划分成基本上大小相等的连贯集合。（我们说“基本上”，因为可能不存在把数据划分成恰好大小相等的诸子集的 X 的数据值。为简单起见，我们将称它们相等。）给定数据分布的第 k 个 q-分位数是值 x，使得小于 x 的数据值最多为 k/q，而大于 x 的数据值最多为 $(q-k)/q$，其中 k 是整数，使得 $0<k<q$。我们有 $q-1$ 个 q-分位数。

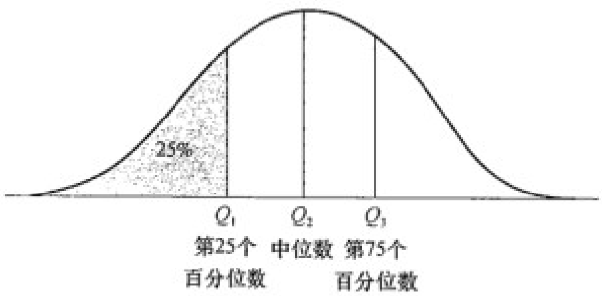


图2.2 某属性 X 的数据分布图。这里绘制的分位数是四分位数。3个四分位数把分布划分成4个相等的部分。第2个四分位数对应于中位数

2-分位数是一个数据点，它把数据分布划分成高低两半。2-分位数对应于中位数。4-分位数是3个数据点，它们把数据分布划分成4个相等的部分，使得每部分表示数据分布的四分之一。通常称它们为**四分位数**（quartile）。100-分位数通常称做**百分位数**（percentile），它们把数据分布划分成100个大小相等的连贯集。中位数、四分位数和百分位数是使用最广泛的分位数。

四分位数给出分布的中心、散布和形状的某种指示。**第1个四分位数**记作 Q_1，是第25个百分位数，它砍掉数据的最低的25%。**第3个四分位数**记作 Q_3，是第75个百分位数，它砍掉数据的最低的75%（或最高的25%）。第2个四分位数是第50个百分位数，作为中位数，它给出数据分布的中心。

---四分位数quartile：把数据划分成4个相等的部分，使得每部分表示数据分布的四分之一

---四分位数极差IQR：即第一个和第三个四分位数之间的距离，它给出被数据的中间一半所覆盖的范围，IQR = Q3 - Q1
---五数概括：分布的五数概括由中位数（Q2），四分位数Q1和Q3，最小和最大观测值组成，按次序minimum，Q1，median，Q3，maximum写出
---盒图：盒图体现了五数概括

- 盒的端点一般在四分位数上，使得盒的长度是四分位数极差 *IQR*。
- 中位数用盒内的线标记。
- 盒外的两条线（称做胡须）延伸到最小（*Minimum*）和最大（*Maximum*）观测值。

当处理数量适中的观测值时，值得个别地绘出可能的离群点。在盒图中这样做：仅当最高和最低观测值超过四分位数不到 $1.5\times IQR$ 时，胡须扩展到它们。否则，胡须在出现在四分位数的 $1.5\times IQR$ 之内的最极端的观测值处终止，剩下的情况个别地绘出。盒图可以用来比较若干个可比较的数据集。

---方差
---标准差

数值属性 X 的 N 个观测值 $x_1, x_2, \cdots, x_N$ 的**方差**（variance）是：

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{n}x_i^2\right)^2 - \bar{x}^2 \tag{2.6}$$

其中，$\bar{x}$ 是观测的均值，由（2.1）式定义。观测值的**标准差**（standard deviation）σ 是方差 σ^2 的平方根。

[方差和标准差都是数据散布度量，它们指出数据分布的散布程度，低标准差意味着数据观测值趋向于非常靠近均值，高标准差意味着数据散布在一个大的值域中]

数据的基本统计描述的图形显示
---分位数图

这里和以下几小节我们介绍常用的数据分布的图形显示。**分位数图**（quantile plot）是一种观察单变量数据分布的简单有效方法。首先，它显示给定属性的所有数据（允许用户评估总的情况和不寻常的出现）。其次，它绘出了分位数信息（见 2.2.2 节）。对于某序数或数值属性 X，设 $x_i(i=1, \cdots, N)$ 是按递增序排序的数据，使得 x_1 是最小的观测值，而 x_N 是最大的。每个观测值 x_i 与一个百分数 f_i 配对，指出大约 $f_i\times 100\%$ 的数据小于值 x_i。我们说"大约"，因为可能没有一个精确的小数值 f_i，使得数据的 $f_i\times 100\%$ 小于值 x_i。注意，百分比 0.25 对应于四分位数 Q_1，百分比 0.50 对应于中位数，而百分比 0.75 对应于 Q_3。

令

$$f_i = \frac{i-0.5}{N} \tag{2.7}$$

这些数从 $\frac{1}{2N}$（稍大于0）到 $1-\frac{1}{2N}$（稍小于1），以相同的步长 $1/N$ 递增。在分位数图中，x_i 对应 f_i 画出。这使得我们可以基于分位数比较不同的分布。例如，给定两个不同时间段的销售数据的分位数图，我们一眼就可以比较它们的 Q_1、中位数、Q_3 以及其他 f_i 值。

例 2.13　分位数图。图 2.4 显示了表 2.1 的单价数据的分位数图。■

表 2.1　AllElectronics 的一个部门销售的商品单价数据集

单价（美元）	商品销售量
40	275
43	300
47	250
…	…
74	360
75	515
78	540
…	…
115	320
117	270
120	350

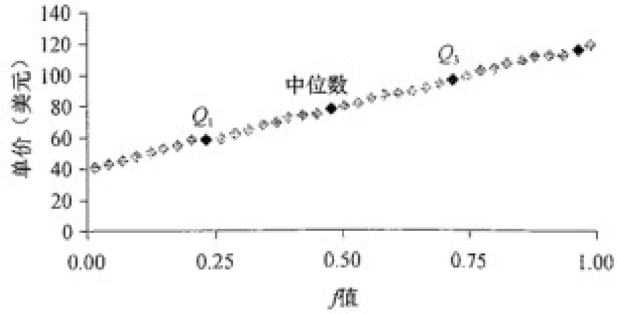


图 2.4　表 2.1 的单价数据的分位数图

---分位数-分位数图

分位数－分位数图（quantile-quantile plot）或 **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。它是一种强有力的可视化工具，使得用户可以观察从一个分布到另一个分布是否有漂移。

假定对于属性或变量 *unit price*（单价），我们有两个观测集，取自两个不同的部门。设 x_1，…，x_N 是取自第一个部门的数据，y_1，…，y_M 是取自第二个部门的数据，其中每组数据都已按递增序排序。如果 $M=N$（即每个集合中的点数相等），则我们简单地对着 x_i 画 y_i，其中 y_i 和 x_i 都是它们的对应数据集的第 $(i-0.5)/N$ 个分位数。如果 $M<N$（即第二个部门的观测值比第一个少），则可能只有 M 个点在 q-q 图中。这里，y_i 是 y 数据的第 $(i-0.5)/M$ 个分位数，对着 x 数据的第 $(i-0.5)/M$ 个分位数画。在典型情况下，该计算涉及插值。

例 2.14　分位数－分位数图。图 2.5 显示在给定的时间段 AllElectronics 的两个不同部门销售的商品的单价数据的分位数－分位数图。每个点对应于每个数据集的相同的分位数，并对该分位数显示部门 1 与部门 2 的销售商品单价。（为帮助比较，我们也画了一条直线，它代表对于给定的分位数，两个部门的单价相同的情况。此外，加黑的点分别对应于 Q_1、中位数和 Q_3。）

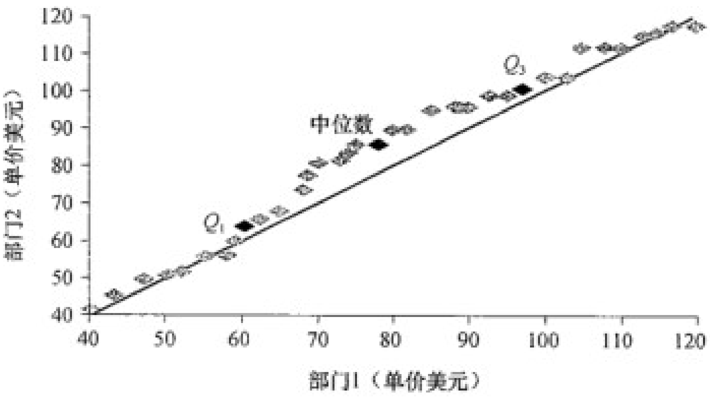


图 2.5　两个不同部门的单价数据的分位数－分位数图

例如，我们看到，在 Q_1，部门 1 销售的商品单价比部门 2 稍低。换言之，部门 1 销售的商品 25% 低于或等于 60 美元，而在部门 2 销售的商品 25% 低于或等于 64 美元。在第 50 个分位数（标记为中位数，即 Q_2），我们看到部门 1 销售的商品 50% 低于或等于 78 美元，而在部门 2 销售的商品 50% 低于或等于 85 美元。一般地，我们注意到部门 1 的分布相对于部门 2 有一个漂移，因为部门 1 销售的商品单价趋向于比部门 2 低。■

---直方图

直方图（histogram）**或频率直方图**（frequency histogram）至少已经出现一个世纪，并且被广泛使用。“histo”意指柱或杆，而“gram”表示图，因此 histogram 是柱图。直方图是一种概括给定属性 X 的分布的图形方法。如果 X 是标称的，如汽车型号或商品类型，则对于 X 的每个已知值，画一个柱或竖直条。条的高度标示该 X 值出现的频率（即计数）。结果图更多地称做条形图（bar chart）。

例 2.15 直方图。图 2.6 显示了表 2.1 的数据集的直方图，其中桶（或箱）定义成等宽的，代表增量 20 美元，而频率是商品的销售数量。■

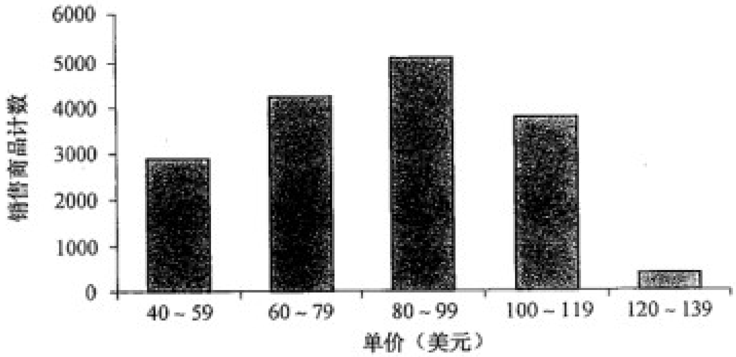


图 2.6 表 2.1 中数据集的直方图

尽管直方图被广泛使用，但是对于比较单变量观测组，它可能不如分位数图、q-q 图和盒图方法有效。

---散点图

散点图（scatter plot）是确定两个数值变量之间看上去是否存在联系、模式或趋势的最有效的图形方法之一。为构造散点图，每个值对视为一个代数坐标对，并作为一个点画在平面上。图 2.7 显示表 2.1 中数据的散点图。

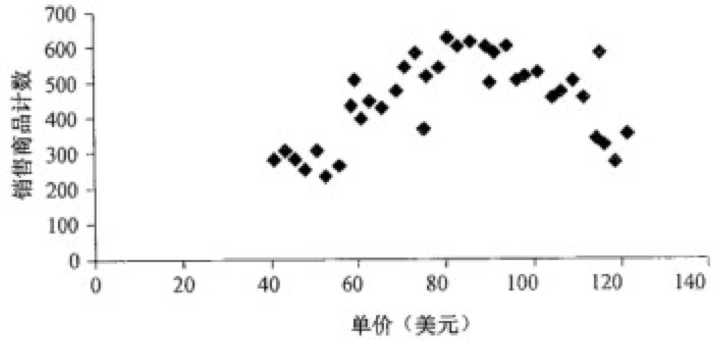


图 2.7 表 2.1 中数据的散点图

散点图是一种观察双变量数据的有用的方法，用于观察点簇和离群点，或考察相关联系的可能性。两个属性 X 和 Y，如果一个属性蕴含另一个，则它们是**相关的**。相关可能是正的、负的或零（null）相关（不相关的）。图 2.8 显示了两个属性之间正相关和负相关的例子。如果标绘点的模式从左下到右上倾斜，则意味 X 的值随 Y 的值增加而增加，暗示正相关（见图 2.8a）。如果标绘点的模式从左上到右下倾斜，则意味 X 的值随 Y 的值减小而增加，暗示负相关（见图 2.8b）。可以画一条最佳拟合的线，研究变量之间的相关性。相关性统计检验在第 3 章介绍数据集成时给出（见（3.3）式）。图 2.9 显示了三种情况，每个给定的数据集的两个属性之间都不存在相关关系。2.3.2 节说明如何把散点图扩展到 n 个属性，得出散点图矩阵。

总结：基本数据描述和图形统计显示提供了数据总体情况的概貌，这有助于识别噪声和离群点，

因此对数据清理很有用。

2.3数据可视化

数据可视化旨在通过图形表示清晰有力的表达数据

基于像素的可视化技术

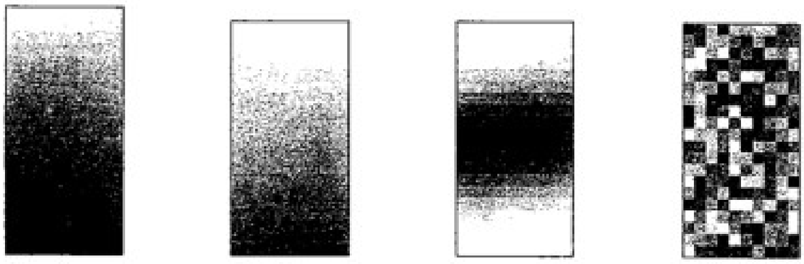
一种可视化一维值的简单方法是使用像素，其中像素的颜色反映该维的值。对于一个 m 维数据集，**基于像素的技术**（pixel-oriented technique）在屏幕上创建 m 个窗口，每维一个。记录的 m 个维值映射到这些窗口中对应位置上的 m 个像素。像素的颜色反映对应的值。

在窗口内，数据值按所有窗口共用的某种全局序安排。全局序可以用一种对手头任务有一定意义方法，通过对所有记录排序得到。

e.g.

例 2.16 基于像素的可视化。AllElectronics 维护了一个顾客信息表，包含 4 个维：*income*（收入），*credit_limit*（信贷额度），*transaction_volume*（成交量）和 *age*（年龄）。我们能够通过可视化技术分析 *income* 与其他属性之间的相关性吗？

我们可以对所有顾客按收入的递增序排序，并使用这个序，在 4 个可视化窗口安排顾客数据，如图 2.10 所示。像素颜色这样选择：值越小，颜色越淡。使用基于像素的可视化，我们可以很容易地得到如下观察：*credit_limit* 随 *income* 增加而增加；收入处于中部区间的顾客更可能从 AllElectronics 购物；*income* 与 *age* 之间没有明显的相关性。 ■



a）*income*　b）*credit_limit*　c）*transaction_volume*　d）*age*

图 2.10 通过按 *income* 的递增序对所有的顾客排序，4 个属性的基于像素的可视化

几何投影可视化技术

e.g.散点图

基于图符的可视化技术

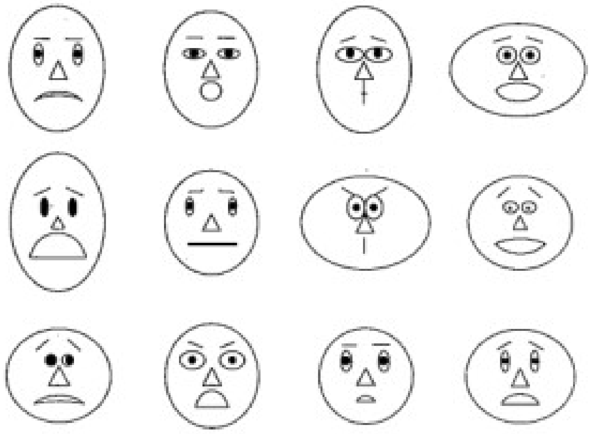


图 2.17　切尔诺夫脸。每张脸表示一个 n 维数据点（$n \leqslant 18$）

e.g.切尔诺夫脸

层次可视化技术

e.g.世界中的世界（n-Vision），数图

可视化复杂对象和关系

e.g.标签云

图 2.20　新闻图：使用树图对 Google 新闻报道标题可视化。资料来源：*www.cs.umd.edu/class/spring2005/cmsc838s/viz4all/ss/newsmap.png*

通常，标签云的用法有两种。首先，对于单个术语的标签云，我们可以使用标签的大小表示该标签被不同的用户用于该术语的次数。其次，在多个术语上可视化标签统计量时，我们可以用标签

的大小表示该标签用于的术语数，即标签的人气。

2.4度量数据的相似形和相异性

相似性和相异性都称邻近性。相似性和相异性是有关联的。如果两个对象i和j不相似，则它们的相似性度量将返回0.相似性值越高，对象之间的相似性越大。相异性度量正好相反，如果对象相同，它将返回值0，相异性值越高，两个对象越相异。

数据矩阵与相异性矩阵

---数据矩阵

数据矩阵（data matrix）或称对象－属性结构：这种数据结构用关系表的形式或 $n \times p$（n 个对象 $\times p$ 个属性）矩阵存放 n 个数据对象：

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix} \tag{2.8}$$

每行对应于一个对象。在记号中，我们可能使用 f 作为遍取 p 个属性的下标。

---相异性矩阵

相异性矩阵（dissimilarity matrix）或称对象－对象结构：存放 n 个对象两两之间的邻近度（proximity），通常用一个 $n \times n$ 矩阵表示：

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix} \tag{2.9}$$

其中d(i,j)是对象i和j之间的相异性或差别的度量。一般而言，d(i,j)是一个非负的数值，对象i和j彼此高度相似或接近时，其值接近于0；而越不同，该值越大。

Notes: d(i,i) = 0，即一个对象与自己的差别为0。此外，d(i,j) = d(j,i)

相似性度量可以表示成相异性度量的函数。

e.g. 对于标称数据，sim(i,j) = 1 - d(i,j)，其中sim(i,j)是对象i和j之间的相似性。

数据矩阵由两种实体或事物组成，即行（代表对象），列（代表属性）。因而，数据矩阵经常被称为二模矩阵。相异性矩阵只包含一类实体，因此被称为单模矩阵。

标称属性的邻近性度量

设一个标称属性的状态数目是 M。这些状态可以用字母、符号或者一组整数（如1，2，…，M）表示。注意这些整数只是用于数据处理，并不代表任何特定的顺序。

“如何计算标称属性所刻画的对象之间的相异性？”两个对象 i 和 j 之间的相异性可以根据不匹配率来计算：

$$d(i,j) = \frac{p-m}{p} \tag{2.11}$$

其中，m 是匹配的数目（即 i 和 j 取值相同状态的属性数），而 p 是刻画对象的属性总数。我们可以通过赋予 m 较大的权重，或者赋给有较多状态的属性的匹配更大的权重来增加 m 的影响。

e.g.

例 2.17　标称属性之间的相异性。假设我们有表 2.2 中的样本数据，不过只有对象标识符和属性 *test*-1 是可用的，其中 *test*-1 是标称的。（在后面的例子中，我们将会用到 *test*-2 和 *test*-3。）让我们来计算相异性矩阵，即（2.9）式

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

由于我们只有一个标称属性 *test*-1，在（2.11）式中，我们令 $p=1$，使得当对象 i 和 j 匹配时，$d(i,j)=0$；当对象不同时，$d(i,j)=1$。于是，我们得到

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

由此，我们看到除了对象 1 和 4（即 $d(4,1)=0$）之外，所有对象都互不相似。■

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

或者，相似性可以用下式计算：

$$sim(i,j) = 1 - d(i,j) = \frac{m}{p} \tag{2.12}$$

标称属性刻画的对象之间的邻近性也可以使用编码方案计算。标称属性可以按以下方法用非对称的二元属性编码：对 M 种状态的每个状态创建一个新的二元属性。对于一个具有给定状态值的对象，对应于该状态值的二元属性设置为 1，而其余的二元属性都设置为 0。例如，为了对标称属性 *map_color* 进行编码，可以对上面所列的五种颜色分别创建一个二元变量。如果一个对象是黄色（*yellow*），则 *yellow* 属性设置为 1，而其余的 4 个属性都设置为 0。对于这种形式的编码，可以用下面讨论的方法来计算邻近度。

用
给
例
变
0。

二元属性的邻近性度量

"那么，如何计算两个二元属性之间的相异性?"一种方法涉及由给定的二元数据计算相异性矩阵。如果所有的二元都被看做具有相同的权重，则我们得到一个两行两列的列联表——表2.3，其中 q 是对象 i 和 j 都取1的属性数，r 是在对象 i 中取1、在对象 j 中取0的属性数，s 是在对象 i 中取0、在对象 j 中取1的属性数，而 t 是对象 i 和 j 都取0的属性数。属性的总数是 p，其中 $p=q+r+s+t$。

表2.3 二元属性的列联表

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

回忆一下，对于对称的二元属性，每个状态都同样重要。基于对称二元属性的相异性称做**对称的二元相异性**。如果对象 i 和 j 都用对称的二元属性刻画，则 i 和 j 的相异性为

$$d(i,j)=\frac{r+s}{q+r+s+t} \tag{2.13}$$

对于非对称的二元属性，两个状态不是同等重要的；如病理化验的阳性（1）和阴性（0）结果。给定两个非对称的二元属性，两个都取值1的情况（正匹配）被认为比两个都取值0的情况（负匹配）更有意义。因此，这样的二元属性经常被认为是"一元的"（只有一种状态）。基于这种属性的相异性被称为**非对称的二元相异性**，其中负匹配数 t 被认为是不重要的，因此在计算时被忽略，如下所示：

$$d(i,j)=\frac{r+s}{q+r+s} \tag{2.14}$$

互补地，我们可以基于相似性而不是基于相异性来度量两个二元属性的差别。例如，对象 i 和 j 之间的**非对称的二元相似性**可以用下式计算：

$$sim(i,j)=\frac{q}{q+r+s}=1-d(i,j) \tag{2.15}$$

（2.15）式的系数 $sim(i,j)$ 被称做 **Jaccard** 系数，它在文献中被广泛使用。

e.g.

例2.18 二元属性之间的相异性。假设一个患者记录表（见表2.4）包含属性 *name*（姓名）、*gender*（性别）、*fever*（发烧）、*cough*（咳嗽）、*test*-1、*test*-2、*test*-3 和 *test*-4，其中 *name* 是对象标识符，*gender* 是对称属性，其余的属性都是非对称二元的。

表2.4 用二元属性描述的患者记录的关系表

name	*gender*	*fever*	*cough*	*test*-1	*test*-2	*test*-3	*test*-4
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N
…	…	…	…	…	…	…	…

对于非对称属性，值 Y(*yes*) 和 P(*positive*) 被设置为1，值 N(*no* 或 *negative*) 被设置为0。假设对象（患者）之间的距离只基于非对称属性来计算。根据（2.14）式，三个患者

Jack、Mary 和 Jim 两两之间的距离如下：

$$d(\text{Jack},\text{Jim}) = \frac{1+1}{1+1+1} = 0.67$$

$$d(\text{Jack},\text{Mary}) = \frac{0+1}{2+0+1} = 0.33$$

$$d(\text{Jim},\text{Mary}) = \frac{1+2}{1+1+2} = 0.75$$

这些度量显示 Jim 和 Mary 不大可能患类似的疾病，因为他们具有最高的相异性。在这三个患者中，Jack 和 Mary 最可能患类似的疾病。■

数值属性的相异性：闵可夫斯基距离

最流行的距离度量是**欧几里得距离**（即，直线或“乌鸦飞行”距离）。令 $i=(x_{i1}, x_{i2}, \cdots, x_{ip})$ 和 $j=(x_{j1}, x_{j2}, \cdots, x_{jp})$ 是两个被 p 个数值属性描述的对象。对象 i 和 j 之间的欧几里得距离定义为：

$$d(i,j) = \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \cdots + (x_{ip}-x_{jp})^2} \qquad (2.16)$$

另一个著名的度量方法是**曼哈顿**（或**城市块**）**距离**，之所以如此命名，是因为它是城市两点之间的街区距离（如，向南 2 个街区，横过 3 个街区，共计 5 个街区）。其定义如下：

$$d(i,j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \cdots + |x_{ip}-x_{jp}| \qquad (2.17)$$

欧几里得距离和曼哈顿距离都满足如下数学性质：

非负性：$d(i, j) \geqslant 0$：距离是一个非负的数值。

同一性：$d(i, i)=0$：对象到自身的距离为 0。

对称性：$d(i, j)=d(j, i)$：距离是一个对称函数。

三角不等式：$d(i, j) \leqslant d(i, k)+d(k, j)$：从对象 i 到对象 j 的直接距离不会大于途经任何其他对象 k 的距离。

满足这些条件的测度称做**度量**（metric）⊖。注意非负性被其他三个性质所蕴含。

例 2.19 欧几里得距离和曼哈顿距离。令 $\boldsymbol{x}_1=(1, 2)$ 和 $\boldsymbol{x}_2=(3, 5)$ 表示如图 2.23 所示的两个对象。两点间的欧几里得距离是 $\sqrt{2^2+3^2}=3.61$。两者的曼哈顿距离是 $2+3=5$。■

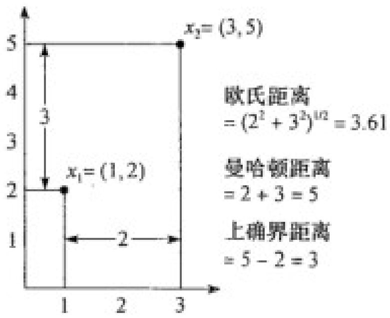


图 2.23 两个对象间的欧几里得距离和曼哈顿距离

闵可夫斯基距离（Minkowski distance）是欧几里得距离和曼哈顿距离的推广，定义如下：

$$d(i,j) = \sqrt[h]{|x_{i1}-x_{j1}|^h + |x_{i2}-x_{j2}|^h + \cdots + |x_{ip}-x_{jp}|^h} \quad (2.18)$$

其中，h 是实数，$h \geq 1$。（在某些文献中，这种距离又称 L_p **范数**（norm），其中 p 就是我们的 h。我们保留 p 作为属性数，以便于本章的其余部分一致。）当 $p=1$ 时，它表示曼哈顿距离（即，L_1 范数）；当 $p=2$ 表示欧几里得距离（即，L_2 范数）。

上确界距离（又称 L_{max}，L_∞ **范数和切比雪夫**（Chebyshev）**距离**）是 $h \to \infty$ 时闵可夫斯基距离的推广。为了计算它，我们找出属性 f，它产生两个对象的最大值差。这个差是上确界距离，更形式化地定义为：

$$d(i,j) = \lim_{h\to\infty}\left(\sum_{f=1}^{p}|x_{if}-x_{jf}|^h\right)^{\frac{1}{h}} = \max_{f}^{p}|x_{if}-x_{jf}| \quad (2.19)$$

L_∞ 范数又称**一致范数**（uniform norm）。

例 2.20 上确界距离。让我们使用相同的数据对象 $\boldsymbol{x}_1=(1,2)$ 和 $\boldsymbol{x}_2=(3,5)$，如图 2.23所示。第二个属性给出这两个对象的最大值差为 $5-2=3$。这是这两个对象间的上确界距离。■

如果对每个变量根据其重要性赋予一个权重，则**加权的欧几里得距离**可以用下式计算：

$$d(i,j) = \sqrt{w_1|x_{i1}-x_{j1}|^2 + w_2|x_{i2}-x_{j2}|^2 + \cdots + w_p|x_{ip}-x_{jp}|^2} \quad (2.20)$$

加权也可以用于其他距离度量。

序数属性的邻近性度量

序数属性的值之间具有有意义的序或排位，而相继值之间的量值未知（2.1.4 节）。例子包括 *size* 属性的值序列 *small*，*medium*，*large*。序数属性也可以通过把数值属性的值域划分成有限个类别，对数值属性离散化得到。这些类别组织成排位。即，数值属性的值域可以映射到具有 M_f 个状态的序数属性 f。例如，区间标度的属性 *temperature*（摄氏温度）可以组织成如下状态：$-30\sim-10$，$-10\sim10$，$10\sim30$，分别代表 *cold temperature*，*moderate temperature* 和 *warm temperature*。令序数属性可能的状态数为 M。这些有序的状态定义了一个排位 $1,\cdots,M_f$。

“*如何处理序数属性?*”在计算对象之间的相异性时，序数属性的处理与数值属性的非常类似。假设 f 是用于描述 n 个对象的一组序数属性之一。关于 f 的相异性计算涉及如下步骤：

1. 第 i 个对象的 f 值为 x_{if}，属性 f 有 M_f 个有序的状态，表示排位 $1,\cdots,M_f$。用对应的排位 $r_{if}\in\{1,\cdots,M_f\}$ 取代 x_{if}。

2. 由于每个序数属性都可以有不同的状态数，所以通常需要将每个属性的值域映射到 [0.0，1.0] 上，以便每个属性都有相同的权重。我们通过用 z_{if} 代替第 i 个对象的 r_{if} 来实现数据规格化，其中

$$z_{if} = \frac{r_{if}-1}{M_f-1} \quad (2.21)$$

3. 相异性可以用 2.4.4 节介绍的任意一种数值属性的距离度量计算，使用 z_{if} 作为第 i 个对象的 f 值。

e.g.

例 2.21 序数型属性间的相异性。 假定我们有前面表 2.2 中的样本数据，不过这次只有对象标识符和连续的序数属性 *test*-2 可用。*test*-2 有三个状态，分别是 *fair*、*good* 和 *excellent*，

也就是 $M_f=3$。第一步，如果我们把 *test*-2 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。第二步，通过将排位 1 映射为 0.0，排位 2 映射为 0.5，排位 3 映射为 1.0 来实现对排位的规格化。第三步，我们可以使用比如说欧几里得距离（(2.16) 式）得到如下的相异性矩阵：

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

因此，对象 1 与对象 2 最不相似，对象 2 与对象 4 也不相似（即，$d(2,1)=1.0$，$d(4,2)=1.0$）。这符合直观，因为对象 1 和对象 4 都是 *excellent*。对象 2 是 *fair*，在 *test*-2 的值域的另一端。 ■

序数属性的相似性值可以由相异性得到：$sim(i,j)=1-d(i,j)$。

混合类型属性的相异性

"那么，我们如何计算混合属性类型的对象之间的相异性？"一种方法是将每种类型的属性分成一组，对每种类型分别进行数据挖掘分析（例如，聚类分析）。如果这些分析得到兼容的结果，则这种方法是可行的。然而，在实际的应用中，每种属性类型分别分析不大可能产生兼容的结果。

一种更可取的方法是将所有属性类型一起处理，只做一次分析。一种这样的技术将不同的属性组合在单个相异性矩阵中，把所有有意义的属性转换到共同的区间［0.0，1.0］上。

假设数据集包含 p 个混合类型的属性，对象 i 和 j 之间的相异性 $d(i, j)$ 定义为：

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}} \tag{2.22}$$

其中，指示符 $\delta_{ij}^{(f)}=0$，如果 x_{if} 或 x_{jf} 缺失（即对象 i 或对象 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；否则，指示符 $\delta_{ij}^{(f)}=1$。属性 f 对 i 和 j 之间相异性的贡献 $d_{ij}^{(f)}$ 根据它的类型计算：

- 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。
- f 是标称或二元的：如果 $x_{if}=x_{jf}$，则 $d_{ij}^{(f)}=0$；否则 $d_{ij}^{(f)}=1$。
- f 是序数的：计算排位 r_{if} 和 $z_{if}=\frac{r_{if}-1}{M_f-1}$，并将 z_{if} 作为数值属性对待。

上面的步骤与我们所见到的各种单一属性类型的处理相同。唯一的不同是对于数值属性的处理，其中规格化使得变量值映射到了区间［0.0，1.0］。这样，即便描述对象的属性具有不同类型，对象之间的相异性也能够进行计算。

e.g.

例 2.22 混合类型属性间的相异性。我们来计算表 2.2 中对象的相异性矩阵。现在，

我们将考虑所有属性，它们具有不同类型。在例 2.17 到例 2.21 中，我们对每种属性计算了相异性矩阵。处理 *test*-1（它是标称的）和 *test*-2（它是序数的）的过程与上文所给出的处理混合类型属性的过程是相同的。因此，在下面计算（2.22）式时，我们可以使用由 *test*-1 和 *test*-2 所得到的相异性矩阵。然而，我们首先需要对第 3 个属性 *test*-3（它是数值的）计算相异性矩阵。即，我们必须计算 $d_{ij}^{(3)}$。根据数值属性的规则，我们令 $max_h x_h = 64$，$min_h x_h = 22$。二者之差用来规格化相异性矩阵的值。结果，*test*-3 的相异性矩阵为：

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

现在就可以在计算（2.22）式时利用这三个属性的相异性矩阵了。对于每个属性 f，指示符 $d_{ij}^{(f)} = 1$。例如，我们得到 $d(3, 1) = \frac{1(1) + 1(0.5) + 1(0.45)}{3} = 0.65$。由三个混合类型的属性所描述的数据得到的结果相异性矩阵如下：

$$\begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$

由表 2.2，基于对象 1 和对象 4 在属性 *test*-1 和 *test*-2 上的值，我们可以直观地猜测出它们两个最相似。这一猜测通过相异性矩阵得到了印证，因为 $d(4, 1)$ 是任何两个不同对象的最小值。类似地，相异性矩阵表明对象 2 和对象 4 最不相似。■

余弦相似性：余弦相似性是一种度量，它可以用来比较文档，或针对给定的查询词向量对文档排序，令x和y是两个待比较的向量，使用余弦度量作为相似性函数，有：

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$$

其中，$\|\boldsymbol{x}\|$ 是向量 $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)$ 的欧几里得范数，定义为 $\sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$。从概念上讲，它就是向量的长度。类似地，$\|\boldsymbol{y}\|$ 是向量 $\boldsymbol{y}$ 的欧几里得范数。该度量计算向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 之间夹角的余弦。余弦值 0 意味两个向量呈 90°夹角（正交），没有匹配。余弦值越接近于 1，夹角越小，向量之间的匹配越大。注意，由于余弦相似性度量不遵守 2.4.4 节定义的度量测度性质，因此它被称做非度量测度（nonmetric measure）。

e.g.

例 2.23　两个词频向量的余弦相似性。 假设 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 是表 2.5 的前两个词频向量。即 $\boldsymbol{x}=(5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$ 和 $\boldsymbol{y}=(3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$。$\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的相似性如何？使用（2.23）式计算这两个向量之间的余弦相似性，我们得到：

$$\begin{aligned}\boldsymbol{x}\cdot\boldsymbol{y} &= 5\times3+0\times0+3\times2+0\times0+2\times1+0\times1+0\times0+2\times1+0\times0+0\times1\\ &= 25\end{aligned}$$

$$\|\boldsymbol{x}\| = \sqrt{5^2+0^2+3^2+0^2+2^2+0^2+0^2+2^2+0^2+0^2} = 6.48$$

$$\|\boldsymbol{y}\| = \sqrt{3^2+0^2+2^2+0^2+1^2+1^2+0^2+1^2+0^2+1^2} = 4.12$$

$$sim(\boldsymbol{x},\boldsymbol{y}) = 0.94$$

因此，如果使用余弦相似性度量比较这两个文档，它们将被认为是高度相似的。■

当属性是二值属性时，余弦相似性函数可以用共享特征或属性解释。假设如果 $x_i=1$，则对象 $\boldsymbol{x}$ 具有第 i 个属性。于是，$\boldsymbol{x}\cdot\boldsymbol{y}$ 是 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 共同具有的属性数，而 $|\boldsymbol{x}|\,|\boldsymbol{y}|$ 是 $\boldsymbol{x}$ 具有的属性数与 $\boldsymbol{y}$ 具有的属性数的几何均值。于是，$sim(\boldsymbol{x}, \boldsymbol{y})$ 是公共属性相对拥有的一种度量。

对于这种情况，余弦度量的一个简单的变种如下：

$$sim(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\boldsymbol{x}\cdot\boldsymbol{x}+\boldsymbol{y}\cdot\boldsymbol{y}-\boldsymbol{x}\cdot\boldsymbol{y}} \tag{2.24}$$

这是 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 所共有的属性个数与 $\boldsymbol{x}$ 或 $\boldsymbol{y}$ 所具有的属性个数之间的比率。这个函数被称为 **Tanimoto** 系数或 **Tanimoto** 距离，它经常用在信息检索和生物学分类中。