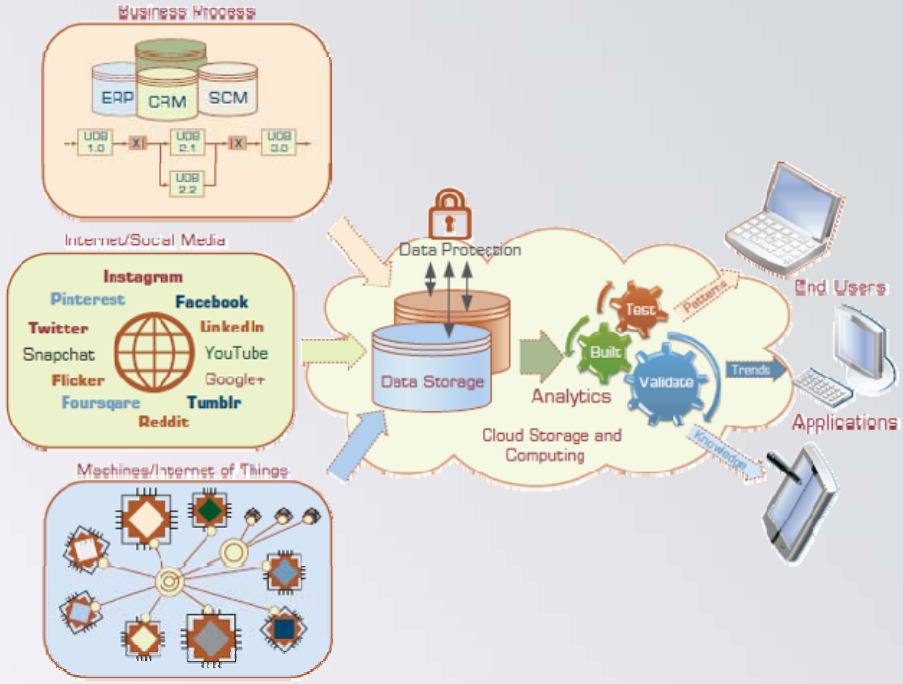
商务智能
Business Intelligence

认识数据（I）

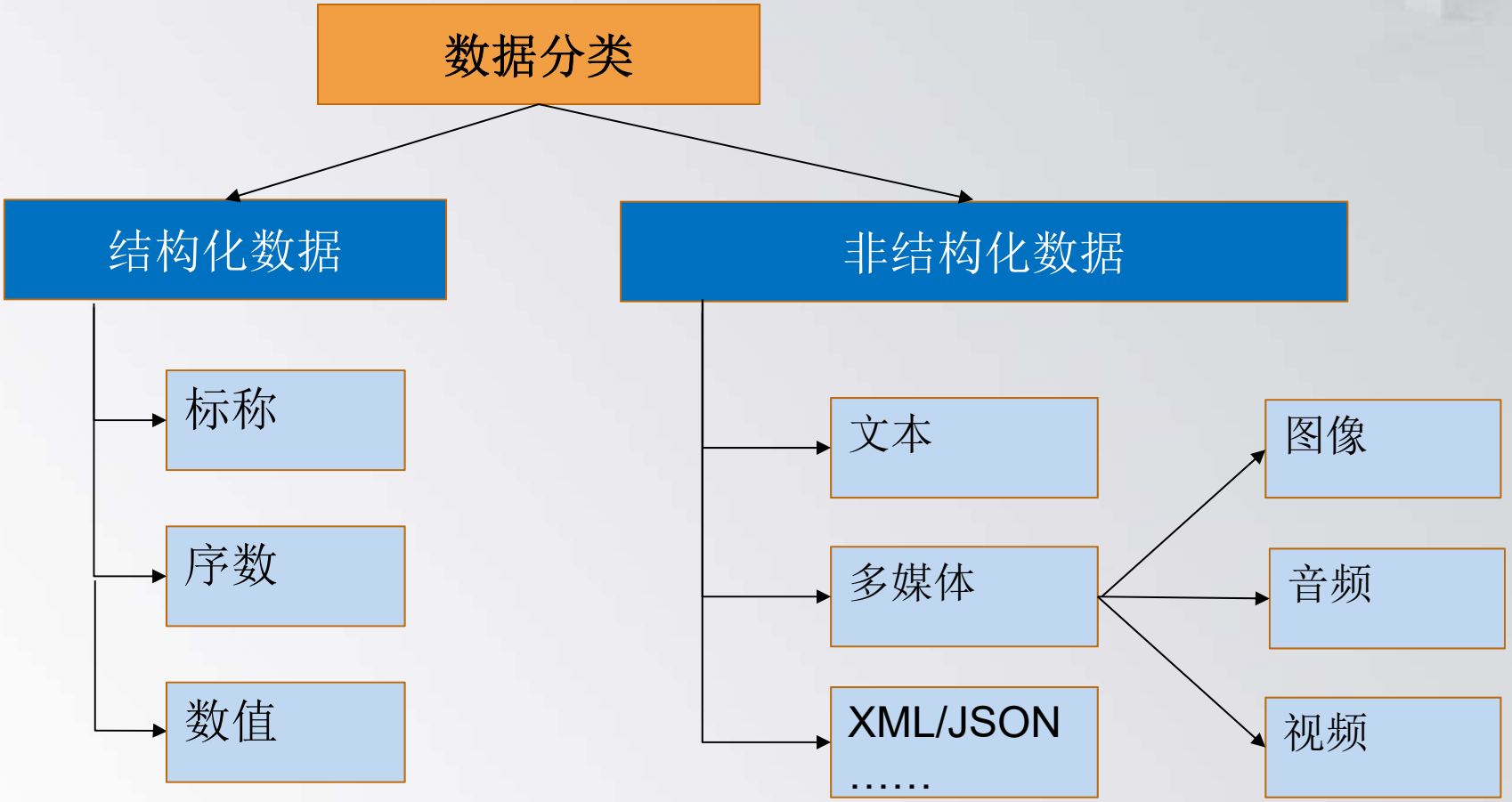
朱卫平 博士
计算机学院
武汉大学

数据的性质

- 数据: 一系列事实的集合
  - 通常是实验、观察、事务或经验获取的结果
- 数据可能包括数字，文字，图像....
- 数据是最低层次的抽象(从中产生信息和知识)



数据的分类



如何判断数据的相似性

学号	姓名	班级号	网络课成绩	商务智能成绩	是否感冒	.....
1	Jack	A	优秀	90	N	
2	Jim	B	一般	95	N	.....
3	Mary	C	好	85	Y	
4	Tom	A	优秀	96	N	

如何判断数据之间是否相似，以及如何表征相似程度?

衡量数据相异性的方法

- 相异性矩阵
  - 给定n对象，相异性矩阵D存放n个对象两两之间的邻近度
  - d(i,j) 值越趋近于0，相异性越小
  - 相似性可以由此计算：$\mathrm{sim}(i,j) = 1 - \mathrm{d}(i,j)$

对象1和2之间的相异性

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

数据对象和属性类型

- 数据集是由数据对象构成的，一个数据对象表示一个实体
  - 在大学数据库中，数据对象是学生
- 数据对象用属性来描述
  - 数据对象具有多个属性，如班级号、成绩、是否感冒等
  - 每个属性具有不同的数据类型，如标称型、序数型和数值型等
- 相似性计算和属性类型相关

标称属性

- 标称属性(nomimal attribute) 是事物的标号或者名称
  - 每一个值表示类别、编码或者状态，因此也被称为是分类
  - 值没有次序信息
  - 在计算机领域，也被称为枚举型
  - 举例如：班级号和发色
    - 班级可以是A班、 B班、 C班等
    - 发色可以是黑色、棕色、灰色等



标称属性

- 尽管名词属性是标号或者名称，但也可以是数值的表示形式
  - 比如，对于班级属性，可以用0表示班级A，1表示班级B等
  - 但是，在这种情况下，该数据并不被当成数值来使用

标称属性的相异性度量

■标称属性的相异性度量

- 标称属性可以取两个或多个状态
- 给定两个对象i和j，它们之间的相异性可以根据不匹配率来计算

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

其中p是对象的属性总数, m是匹配属性数目（取值相同的属性数），p-m表示不匹配属性数目

标称属性的相异性度量

例 假设我们有以下数据，求其相异性矩阵为

学号	班级号
1	A
2	B
3	C
4	A

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

由于只有一个标称属性，我们令属性总数p=1
当对象i和j匹配时，不匹配属性数目为0，d(i,j)=0
当对象不同时不匹配属性数目为1，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之外，所有对象都互不相似

二元属性

- 二元属性是一种特殊的标称属性，只有两个状态：0和1
  - 0一般表示属性缺失，1表示存在
  - 又称为布尔属性，两个状态表示真和假
  - 如果二进制属性的两个状态是同等重要的，称为对称二元属性
    - 比如性别属性的两个值男和女
  - 如果两个状态不是同等重要的，则为非对称二元属性
    - 比如感冒检查的结果呈阴性和阳性
    - 通常，用1表示更重要的（更稀少）的结果

二元属性的相异性度量

■对于对称二元属性，对象i和j的相异性可以利用以下列表

		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

则i和j的相异性为： $d(i, j) = \frac{r + s}{q + r + s + t}$

二元属性的相异性度量

■非对称二元属性的相异性度量

- 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
- 很多时候，负匹配的数值将远大于正匹配，使得相异性绝对值变的很小
- 因此，不重要的负匹配数在计算时被忽略，如下所示:

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

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

		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

二元属性的相异性度量

例 如下图学生数据表计算其相异性，除姓名外其余属性都是非对称二元属性

姓名	是否发烧	是否感冒	检测1	检测2	检测3	检测4
Jack	Y	N	P	N	N	N
Jim	Y	Y	N	N	N	N
Mary	Y	N	P	N	P	N
...	...	...	...	...	...	...

对于这些属性，值Y（yes）和P(positive)被设置为1，值N(no或negative)被设置为0。
假设对象之间的距离只基于上述非对称二元属性来计算，则三个对象之间的距离如下：

二元属性的相异性度量

姓名	是否发烧	是否感冒	检测1	检测2	检测3	检测4
Jack	Y	N	P	N	N	N
Jim	Y	Y	N	N	N	N
Mary	Y	N	P	N	P	N
...	...	...	...	...	...	...

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

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

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

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

这些度量显示如何患病的话， Jim 和 Mary 不大可能患类似的疾病，因为他们有最高相异性；在这三个人中，Jack 和 Mary 最可能患类似的疾病。

数值属性的相异性

- 数值型属性是可测量的数值
  - 用于计算数值属性相异性的距离度量包括欧氏距离、曼哈顿距离和闵可夫斯基距离
  - 欧氏距离

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

  - 曼哈顿距离

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

数值属性的相异性

- 闵可夫斯基距离

  闵可夫斯基距离是欧氏距离和曼哈顿距离的推广：

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

其中h是大于等于1的实数

  - 当h=1时，它表示曼哈顿距离
  - 当h=2时，它表示欧氏距离

数值属性的相异性

- 可以对每个变量根据其重要性赋予一个权重，从而形成加权距离
  - 加权的欧几里得距离可以用下式计算：

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

数值属性的相异性

例 请基于下表计算JacK和Jim间的欧几里得距离和曼哈段距离?

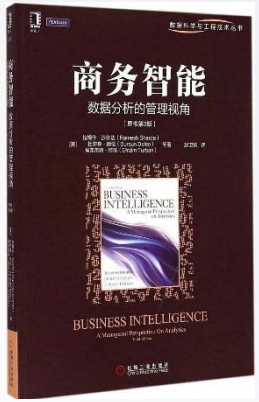
姓名	C语言成绩	商务智能成绩
Jack	90	90
Jim	80	95

欧式距离为 $\sqrt{(90-80)^2+(90-95)^2}=11.18$

曼哈顿距离为 |90-80|+|90-95|=15

序数属性

- 序数属性具有次序或者级别的值。但是相邻值之间的差是未知的
  - 如课程成绩A-, A 和 A+ 之间有排序，但不能分辨A+比A多多少



A-	A	A+
85-89	90-94	95-100

序数属性

- 更多的例子：饮料尺寸可以是“小杯”，“中杯”，“大杯”。值有顺序的意义，但是不能分辨中杯比大杯大多少。
- 序数属性被用来衡量无法客观衡量的属性，用主观的评估定质量。在调查中常用来排序。比如，参与者作为顾客，他们的满意度可以是：0：非常不满意，1 有点不满意，2 中立 3 满意 4 很满意
- 把数值数据离散化，把它们按照值的范围分类，也可以得到序数属性的数据。

序数属性的相异性度量

■ 序数属性的相异性计算过程

- 假设第*i*个对象的属性*f*值为x_{if}, *f*有M_f个有序的状态1,...M_f
- 用对应的排位$r_{if} \in \{1,...M_f\}$ 取代x_{if}
- 进行数据归一化

$$Z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

由于每个序数属性可以有不同的状态数，需要将每个属性的值域映射到[0,1]上，以便每个属性都有相同的权重。

- 基于z_{if}，相异性可以用数值属性相异性来计算

序数属性的相异性度量

例 对下表中的数据求相异性矩阵

姓名	商务智能成绩	
Jack	优秀	3
Jim	一般	1
Mary	好	2
Tom	优秀	3

属性有三个状态，也就是说$M_f = 3$

第一步：将属性值替换为它的排位，对应3,1,2,3

第二步：归一化，排位1映射为0.0, 2为0.5, 3为1.0

第三步：使用欧式距离计算相异性矩阵

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

因此Jack和Jim成绩最不相似，Jim与Tom也不相似

混合类型属性的相异性

- 在许多实际的数据集中，对象是由多种不同类型的属性共同描述的
- 计算混合属性类型的对象之间的相异性
  - 将每种类型的属性分成一组，分别进行数据挖掘，当得到兼容的结果时这种方法是可行的
  - 更可取的方法是将所有属性类型一起处理，只做一次分析。把所有有意义的属性转换到共同的区间[0,1]上，然后计算单个相异矩阵
  - 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：

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

混合类型属性的相异性

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

- 两个项目相除，分子是带指示器的相异性求和，分母是指示器求和
- 在每个求和中，依次考虑每个属性，一共有p个属性
- 对属性f， 对象i, j之间 的计算如下

- 指示符 $\delta_{ij}^{(f)}$
  - 指示符 $\delta_{ij}^{(f)}$ =0
    1) x_{if} 或者x_{jf} 缺失
    2) x_{if}= x_{jf}=0，且f是非对称二元属性
  - 否则指示符$\delta_{ij}^{(f)}$ =1

- 相异性贡献$d_{ij}(f)$根据它的类型计算
  - 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是数值型： $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max x_f - \min x_f}$
  - f是序数的：计算排位 r_{if} 和 z_{if}，然后做为数值属性对待

混合类型属性的相异性

例 计算相异性矩阵

学号	班级号	商务智能成绩	运动次数
1	A	优秀	45
2	B	一般	22
3	C	好	64
4	A	优秀	28

在前面的例子中，我们对**班级号**（标称属性）和**商务智能成绩**（序数属性）计算了相异性矩阵。

我们需要计算**运动次数**（数值）的相异性矩阵，即我们必须计算$d_{ij}^{(3)}$。根据数值的规则，我们令$\max x_3=64, \min x_3=22$。两者之差用来归一化相异性矩阵的值

混合类型属性的相异性

三者的相异性矩阵,分别为

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

对于每个属性f，指示符$d_{ij}^{(f)}=1$。

例如，我们得到 $$d(3,1)=\frac{1(1)+1(0.5)+1(0.45)}{3}=0.65$$

得到的结果相异性矩阵如下

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

Thank You!

Q&A