第二章　聚类分析

2.2 模式相似性测度

2.2.2 相似测度

测度基础：以两矢量的方向是否相近作为考虑的基础，矢量长度并不重要。设

$$\vec{x}=(x_1,x_2,\ldots x_n)',\ \vec{y}=(y_1,y_2,\ldots y_n)'$$

1. 角度相似系数（夹角余弦）

$$\cos(\vec{x},\vec{y})=\frac{\vec{x}'\vec{y}}{\|\vec{x}\|\|\vec{y}\|}=\frac{\vec{x}'\vec{y}}{\left[(\vec{x}'\vec{x})(\vec{x}'\vec{x})\right]^{1/2}}$$

注意：坐标系的旋转和尺度的缩放是不变的，但对一般的线形变换和坐标系的平移不具有不变性。

2.2 模式相似性测度

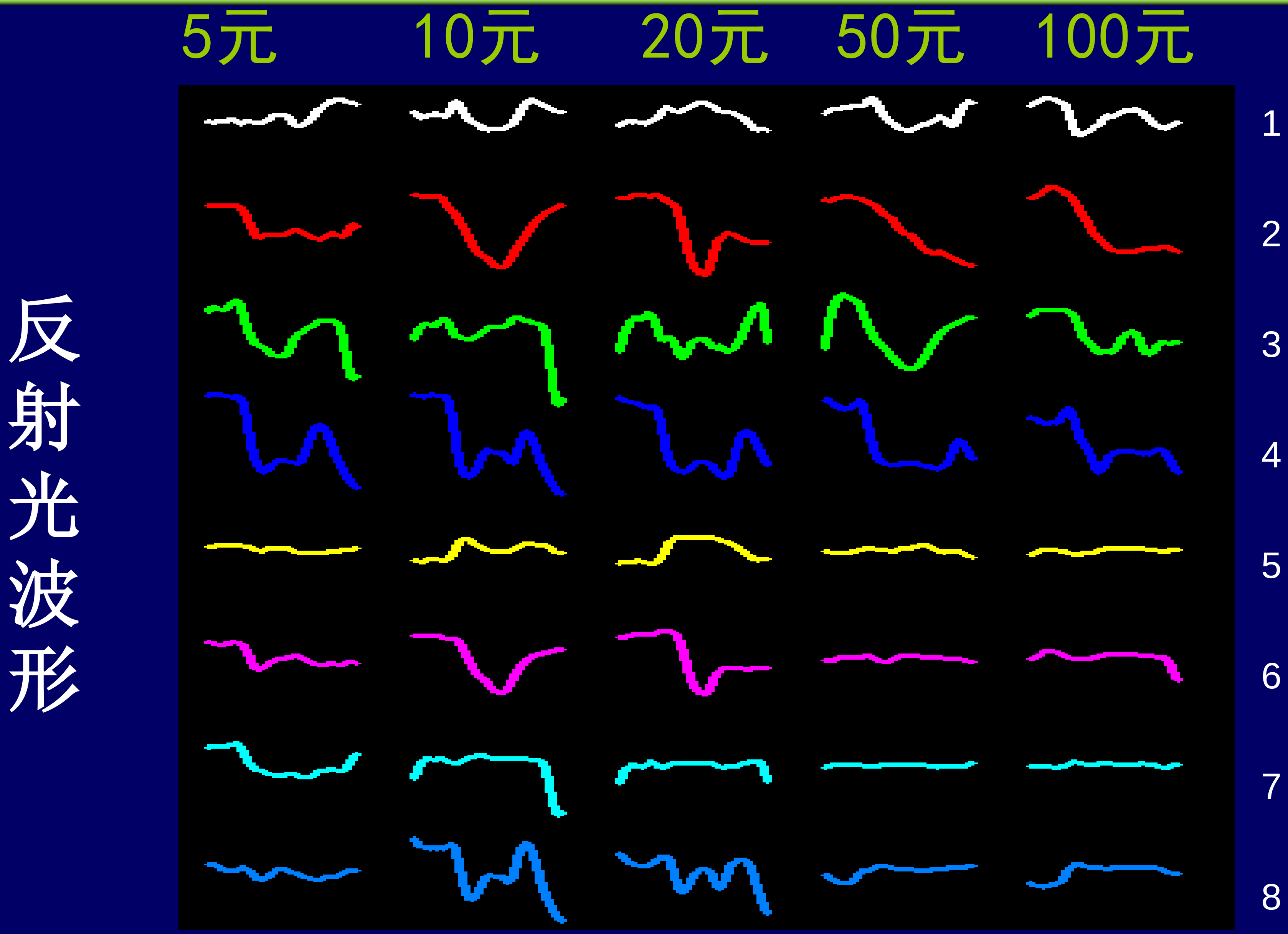
2.2.2 相似测度

2. 相关系数

它实际上是数据中心化后的矢量夹角余弦。

$$r(\vec{x},\vec{y})=\frac{(\vec{x}-\bar{\vec{x}})'(\vec{y}-\bar{\vec{y}})}{\left[(\vec{x}-\bar{\vec{x}})'(\vec{x}-\bar{\vec{x}})(\vec{y}-\bar{\vec{y}})'(\vec{y}-\bar{\vec{y}})\right]^{1/2}}$$

2.2 模式相似性测度



2.2 模式相似性测度

现金识别例子——100圆A面传感器1与其它各面的相关系数

	a	b	c	d	e	f	g	h
100圆	0.87,	0.68,	0.73,	-0.21,	0.47,	0.68,	0.72,	-0.16
50圆	0.75,	0.45,	0.72,	-0.71,	-0.03,	0.65,	0.71,	-0.68
20圆	0.67,	-0.42,	0.68,	-0.64,	0.59,	-0.30,	0.54,	-0.72
10圆	0.76,	0.00,	0.68,	-0.79,	0.76,	0.15,	0.56,	-0.77

2.2 模式相似性测度

2.2.2 相似测度

3. 指数相似系数

$$e(\vec{x},\vec{y})=\frac{1}{n}\sum_{i=1}^{n}\exp\left[-\frac{3}{4}\frac{(x_i-y_i)^2}{\sigma_i^2}\right]$$

式中 σ_i^2 为相应分量的协方差，n 为矢量维数。它不受量纲变化的影响。

2.2 模式相似性测度

现金识别例子——100圆A面传感器1与其它各面的相关系数

	a	b	c	d	e	f	g	h
100圆	3.52,	0.00,	1.46,	0.55,	3.58,	0.00,	1.31,	0.21
50圆	0.00,	0.00,	0.00,	0.00,	0.50,	0.00,	0.14,	0.37
20圆	0.02,	0.00,	0.00,	0.33,	2.63,	0.10,	0.76,	0.52
10圆	0.00,	0.00,	0.14,	0.00,	0.00,	0.00,	0.07,	0.01

2.2 模式相似性测度

2.2.3 匹配测度

当特征只有两个状态（0，1）时，常用匹配测度。

0表示无此特征，1表示有此特征。故称之为二值特征。

对于给定的x和y中的某两个相应分量x_i与y_j

若$x_i=1, y_j=1$，则称 x_i与y_j是 (1-1)匹配；

若$x_i=1, y_j=0$，则称 x_i与y_j是 (1-0)匹配；

若$x_i=0, y_j=1$，则称 x_i与y_j是 (0-1)匹配；

若$x_i=0, y_j=0$，则称 x_i与y_j是 (0-0)匹配。

2.2 模式相似性测度

2.2.3 匹配测度

对于二值n维特征矢量可定义如下相似性测度

令 $a=\sum_i x_i y_i$ 为 $\vec{x}$ 与 $\vec{y}$ 的 (1–1) 匹配的特征数目

$b=\sum_i y_i(1-x_i)$ (0–1) 匹配的特征数目

$c=\sum_i x_i(1-y_i)$ (1–0) 匹配的特征数目

$e=\sum_i (1-x_i)(1-y_i)$ (0–0) 匹配的特征数目

2.2.3 匹配测度

(1)Tanimoto测度

$$s(\vec{x},\vec{y})=\frac{a}{a+b+c}=\frac{\vec{x}'\vec{y}}{\vec{x}'\vec{x}+\vec{y}'\vec{y}-\vec{x}'\vec{y}}$$

2.2 模式相似性测度

2.2.3 匹配测度

例：

设 $\vec{x}=(0,1,0,1,1,0)'$ $\vec{y}=(0,0,1,1,0,1)'$

则 $\vec{x}'\vec{x}=3$， $\vec{y}'\vec{y}=3$ ， $\vec{x}'\vec{y}=1$

$$s(\vec{x},\vec{y})=\frac{1}{3+3-1}=\frac{1}{5}$$

可以看出，它等于共同具有的特征数目与分别具有的特征种类总数之比。这里只考虑(1-1)匹配而不考虑(0-0)匹配。

2.2 模式相似性测度

2.2.3 匹配测度

(2) Rao测度

$$s(\vec{x},\vec{y})=\frac{a}{a+b+c+e}=\frac{\vec{x}'\vec{y}}{n}$$

注：(1-1)匹配特征数目和所选用的特征数目之比。

2.2.3 匹配测度

例：

设 $\vec{x}=(0,1,0,1,1,0)'$ $\vec{y}=(0,0,1,1,0,1)'$

则 $$n=6\ ,\quad \vec{x}'\vec{y}=1$$

$$s(\vec{x},\vec{y})=\frac{1}{6}$$

2.2.3 匹配测度

(3) 简单匹配系数

$$m(\vec{x},\vec{y})=\frac{a+e}{n}$$

注：上式分子为(1-1)匹配特征数目与(0-0)匹配特征数目之和，分母为所考虑的特征数目。

2.2 模式相似性测度

2.2.3 匹配测度

例：

设 $\vec{x}=(0,1,0,1,1,0)' \quad \vec{y}=(0,0,1,1,0,1)'$

则 $n=6\ , \quad a=1 \quad ,e=1$

$$s(\vec{x},\vec{y})=\frac{2}{6}=\frac{1}{3}$$

2.2.3 匹配测度

(4) Dice系数

$$m(\vec{x},\vec{y})=\frac{a}{2a+b+c}=\frac{\vec{x}'\vec{y}}{\vec{x}'\vec{x}+\vec{y}'\vec{y}}=\frac{(1\text{-}1)\text{匹配个数}}{\text{俩矢量中}1\text{的总数}}$$

设 $\vec{x}=(0,1,0,1,1,0)'$ $\vec{y}=(0,0,1,1,0,1)'$

则

$$m(\vec{x},\vec{y})=\frac{\vec{x}'\vec{y}}{\vec{x}'\vec{x}+\vec{y}'\vec{y}}=\frac{1}{3+3}=\frac{1}{6}$$

2.2 模式相似性测度

2.2.3 匹配测度

(5) Kulzinsky系数

$$m(\vec{x},\vec{y})=\frac{a}{b+c}=\frac{\vec{x}'\vec{y}}{\vec{x}'\vec{x}+\vec{y}'\vec{y}-2\vec{x}'\vec{y}}=\frac{(1-1)\text{匹配个数}}{(0-1)+(1-0)\text{匹配个数}}$$

设 $\vec{x}=(0,1,0,1,1,0)'\quad \vec{y}=(0,0,1,1,0,1)'$

则

$$m(\vec{x},\vec{y})=\frac{\vec{x}'\vec{y}}{\vec{x}'\vec{x}+\vec{y}'\vec{y}-2\vec{x}'\vec{y}}=\frac{1}{3+3-2}=\frac{1}{4}$$

2.2 模式相似性测度

现金识别例子——100圆A面与其它各面的匹配系数Kulzinsky

	a	b	c	d	e	f	g	h
100圆	9.44,	2.95,	4.33,	3.50,	5.61,	3.15,	3.04,	3.00
50圆	4.53,	2.92,	4.25,	3.37,	4.55,	2.77,	3.95,	2.97
20圆	3.32,	1.43,	2.67,	1.81,	1.60,	1.01,	1.32,	1.31
10圆	2.68,	1.21,	1.98,	1.29,	2.01,	1.26,	2.33,	1.36

第二章　聚类分析

2.3.1 类的定义

- 研究分类算法之前应了解类的定义
- 类的不同定义适用于不同的模式分布情况
- 类的定义很多情况下融于准则函数中

类的约束

对于一个待分类的集合S，要求分类后的各类$S_1, S_2, \cdots, S_c$满足：

(1)　$S_i \neq \varnothing$　　(2)　$\bigcup_{i=1}^{c} S_i = S$

(3)　$S_i \cap S_j = \varnothing, \quad i, j = 1, 2, \cdots, c; \; i \neq j$

2.3.1 类的定义

定义1: 若集合 S 中任两个元素 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$ 的距离 d_{ij} 有

$$d_{ij} \le h$$

则称 S 相对于阈值 h 组成一类。

定义2: 若集合 S 中任一元素 $\boldsymbol{x}_i$ 与其他各元素 $\boldsymbol{x}_j$ 间的距离 d_{ij} 均满足

$$\frac{1}{k-1}\sum_{\boldsymbol{x}_j \in S} d_{ij} \le h$$

则称 S 相对于阈值 h 组成一类（$\boldsymbol{k}$为集合元素个数）。

2.3.1 类的定义

定义3: 若集合 S 中任两个元素 $\boldsymbol{x}_i$ 、$\boldsymbol{x}_j$ 的距离 d_{ij} 满足

$$\frac{1}{k(k-1)}\sum_{x_i\in S}\sum_{x_j\in S} d_{ij}\le h \quad 且 \quad d_{ij}\le r$$

则称 S 相对于阈值 h 、$\boldsymbol{r}$ 组成一类。

定义4: 若集合 S 中元素满足对于任一 $x_i\in S$，都存在某 $x_j\in S$ 使它们的距离

$$d_{ij}\le h$$

则称 S 相对于阈值 h 组成一类。

2.3.1 类的定义

定义5: 若集合S 任意分成两类S_i、S_j，这两类的距离

$$D(S_1, S_2) \le h$$

则称 S相对于阈值h组成一类。

作业

P63:   2.1, 2.2

谢谢！