第二章　聚类分析

2.1　聚类分析的概念

2.2　模式相似性测度

2.3　类的定义与类间距离

2.4　准则函数

2.5　聚类的算法

2.1 聚类分析的概念

2.1.1 聚类分析的基本思想

- 假设

对象集客观存在着若干个自然类，每个自然类中个体的某些属性具有较强的相似性。

- 原理

将给定模式分成若干组，每组内的模式是相似的，而组间各模式差别较大。

2.1 聚类分析的概念

2.1.1 聚类分析的基本思想

一、基本思想

★相似的归为一类。

★模式相似性的度量和聚类算法。

★无监督分类（Unsupervised）。

二、特征量的类型

★物理量----（重量、长度、速度）

★次序量----（等级、技能、学识）

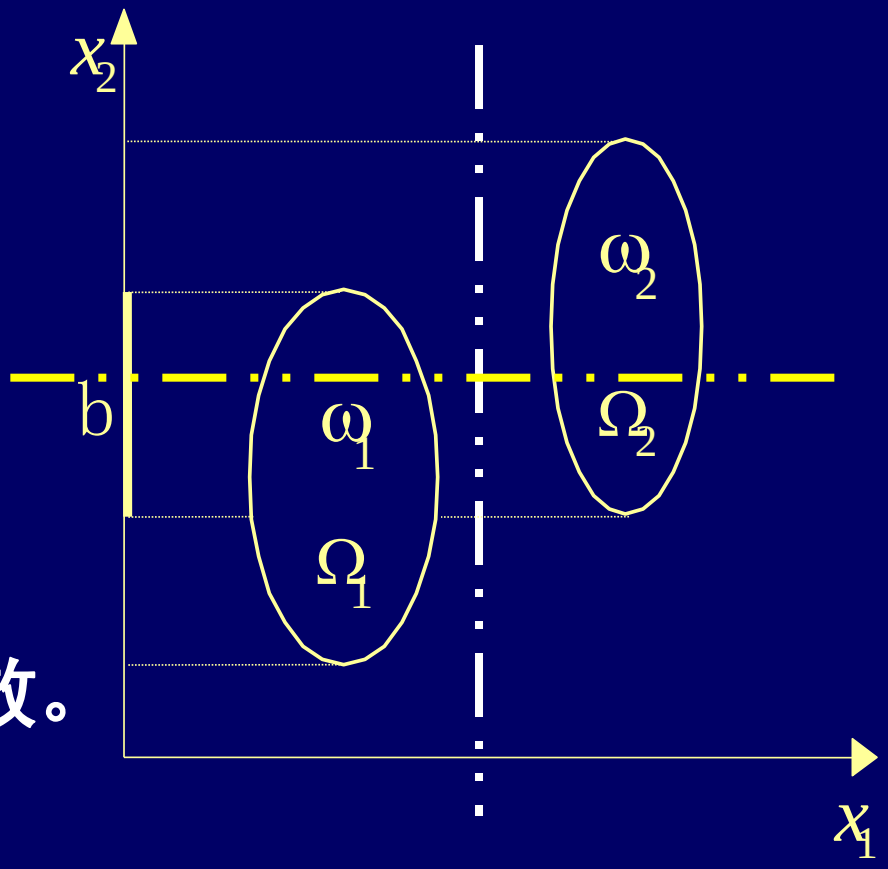
★名义量----（性别、状态、种类）

2.1 聚类分析的概念

2.1.1 聚类分析的基本思想

三、方法的有效性

取决于分类算法和特征点分布情况的匹配。



分类无效时的情况

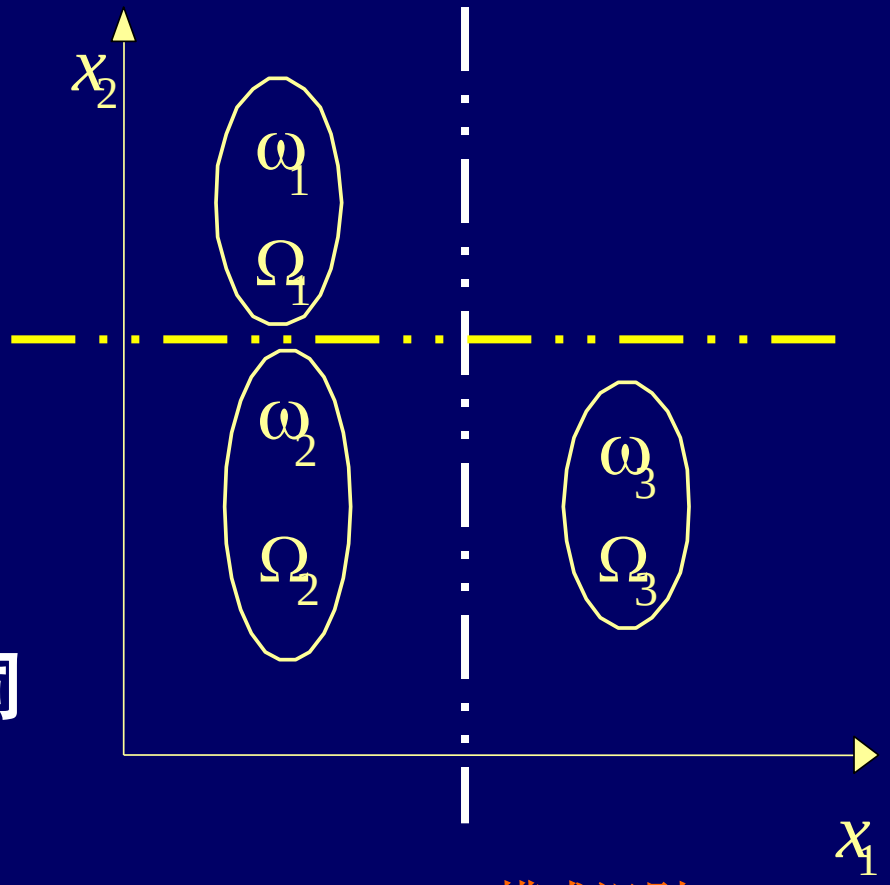
1. 特征选取不当使分类无效。

2.1 聚类分析的概念

2.1.1 聚类分析的基本思想

三、方法的有效性

取决于分类算法和特征点分布情况的匹配。



分类无效时的情况

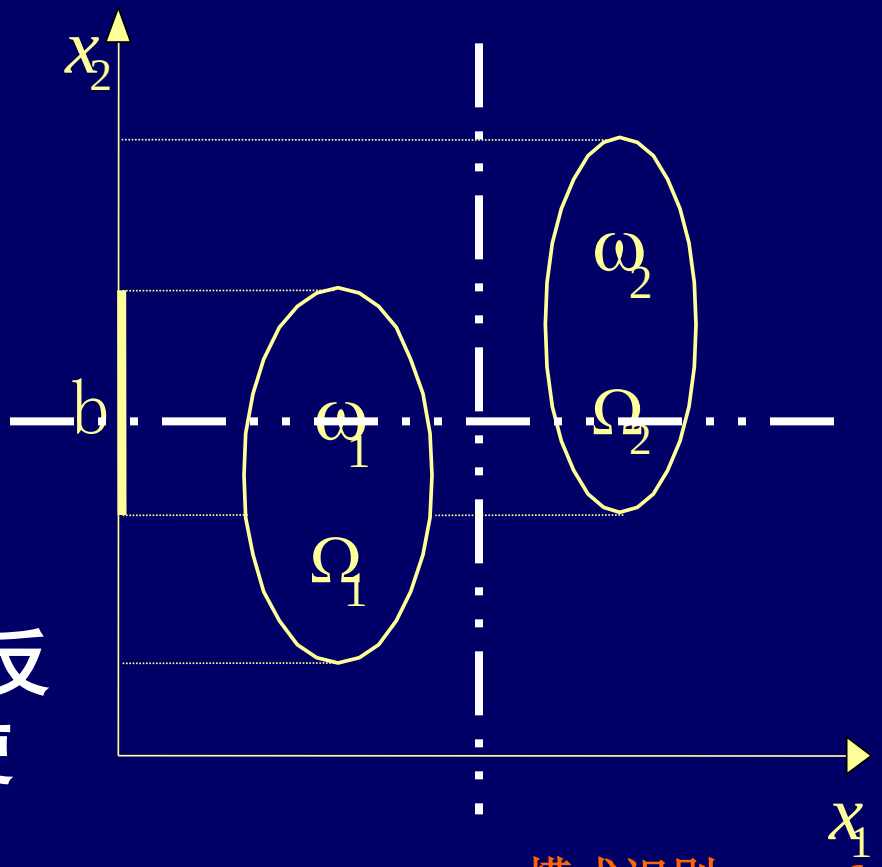
2. 特征选取不足可能使不同类别的模式判为一类。

2.1 聚类分析的概念

2.1.1 聚类分析的基本思想

三、方法的有效性

取决于分类算法和特征点分布情况的匹配。



分类无效时的情况

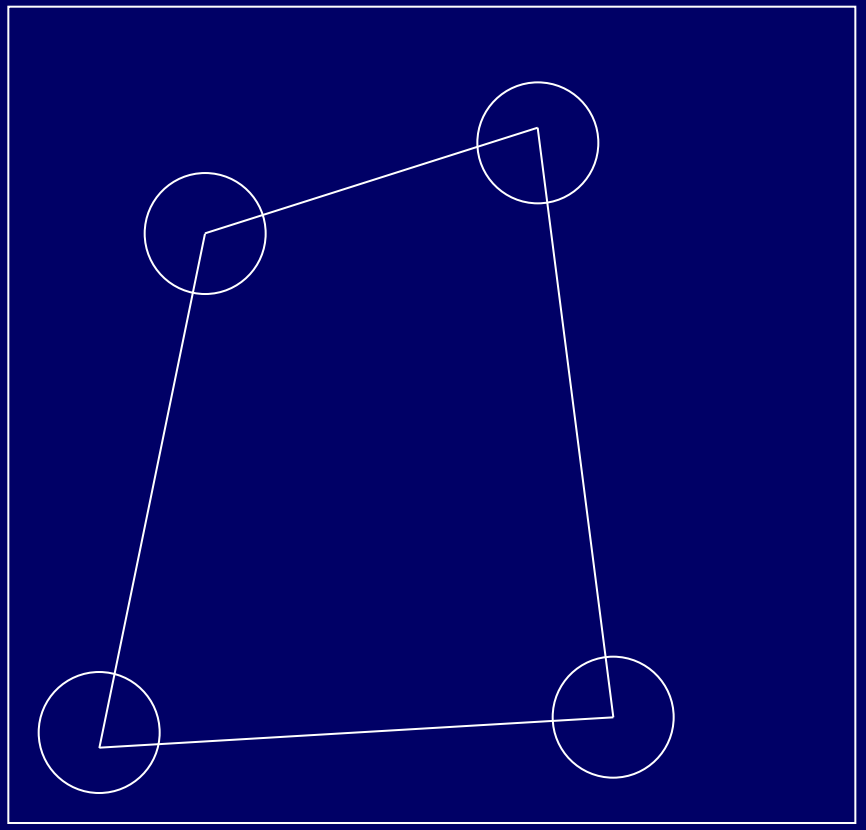
3. 特征选取过多可能无益反而有害，增加分析负担并使分析效果变差。

2.1 聚类分析的概念

2.1.1 聚类分析的基本思想

三、方法的有效性

取决于分类算法和特征点分布情况的匹配。



分类无效时的情况

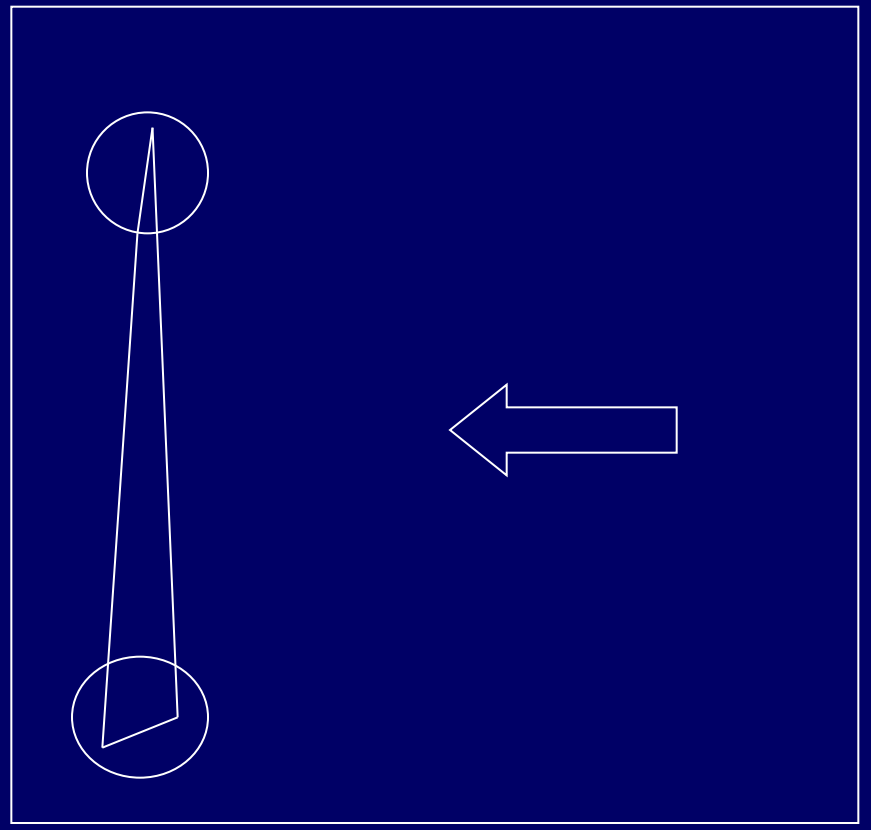
4. 量纲选取不当。

2.1 聚类分析的概念

2.1.1 聚类分析的基本思想

三、方法的有效性

取决于分类算法和特征点分布情况的匹配。



分类无效时的情况

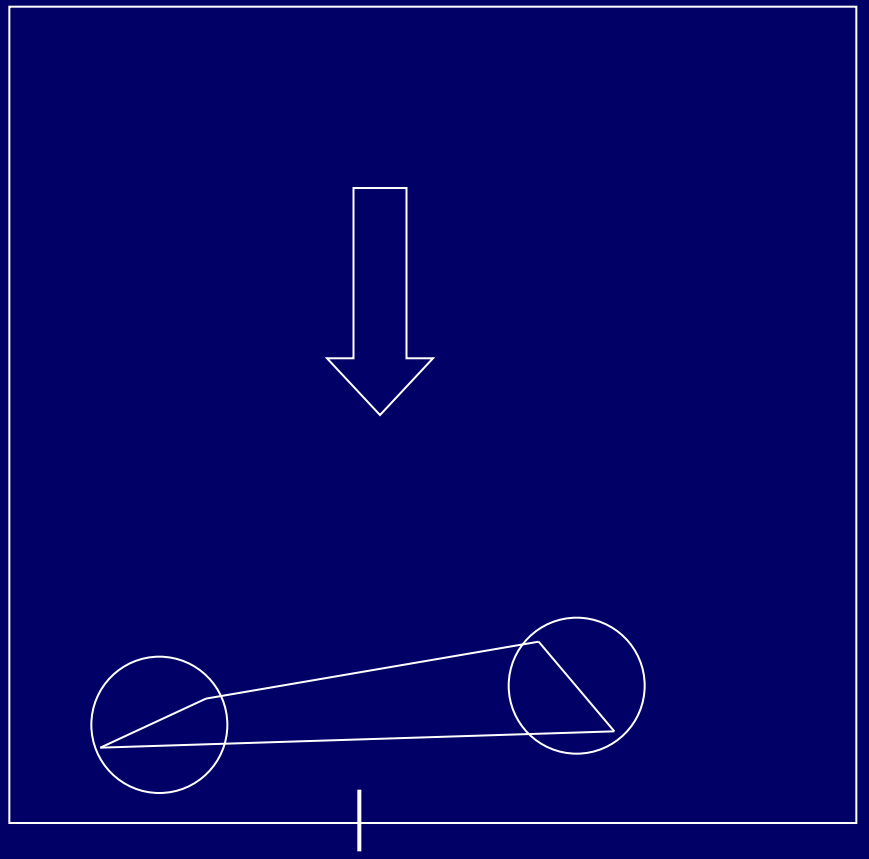
4. 量纲选取不当。

2.1 聚类分析的概念

2.1.1 聚类分析的基本思想

三、方法的有效性

取决于分类算法和特征点分布情况的匹配。



分类无效时的情况

4. 量纲选取不当。

2.1 聚类分析的概念

特征选取不同对聚类结果的影响

下列是一些动物的名称：

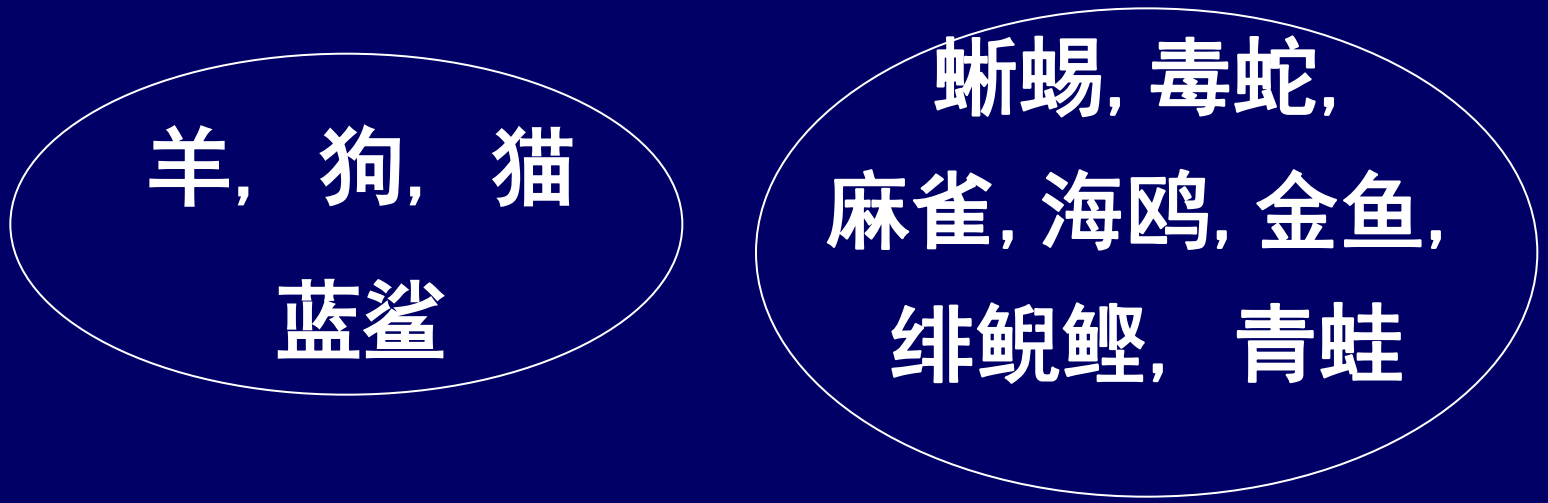
羊	（sheep）	狗	（dog）
蓝鲨	（blue shark）	蜥蜴	（lizard）
毒蛇	（viper）	猫	（cat）
麻雀	（sparrow）	海鸥	（seagull）
金鱼	（gold fish）	绯鲣鲣	（red-mullet）
蛙	（frog）		

要对这些动物进行分类，则不同的特征有不同的分法：

2.1 聚类分析的概念

特征选取不同对聚类结果的影响

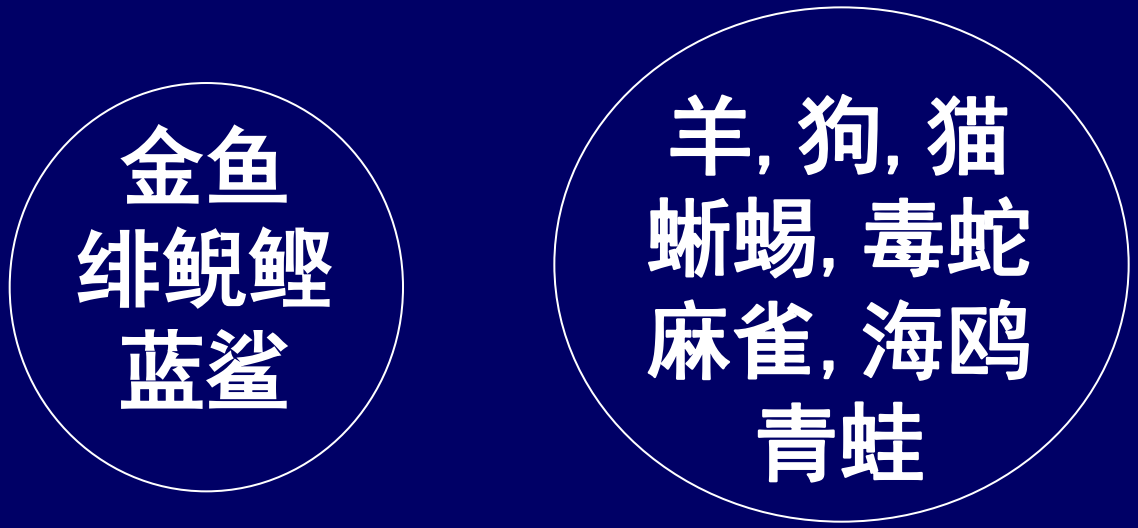
(a) 按繁衍后代的方式分



2.1 聚类分析的概念

特征选取不同对聚类结果的影响

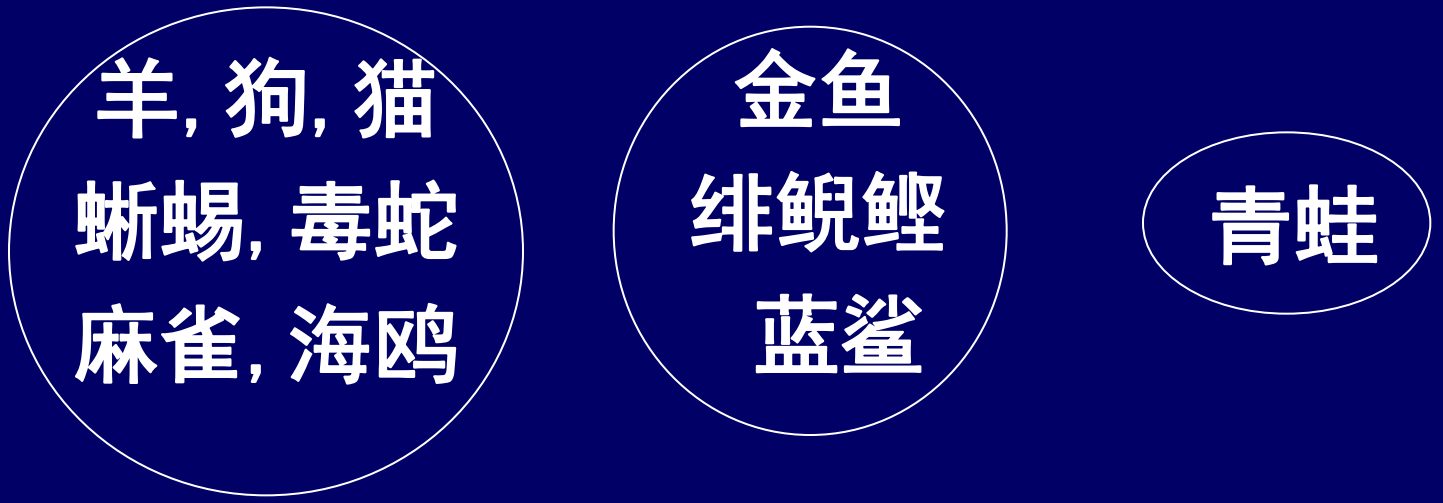
(b) 按肺是否存在分



2.1 聚类分析的概念

特征选取不同对聚类结果的影响

(c) 按生活环境分



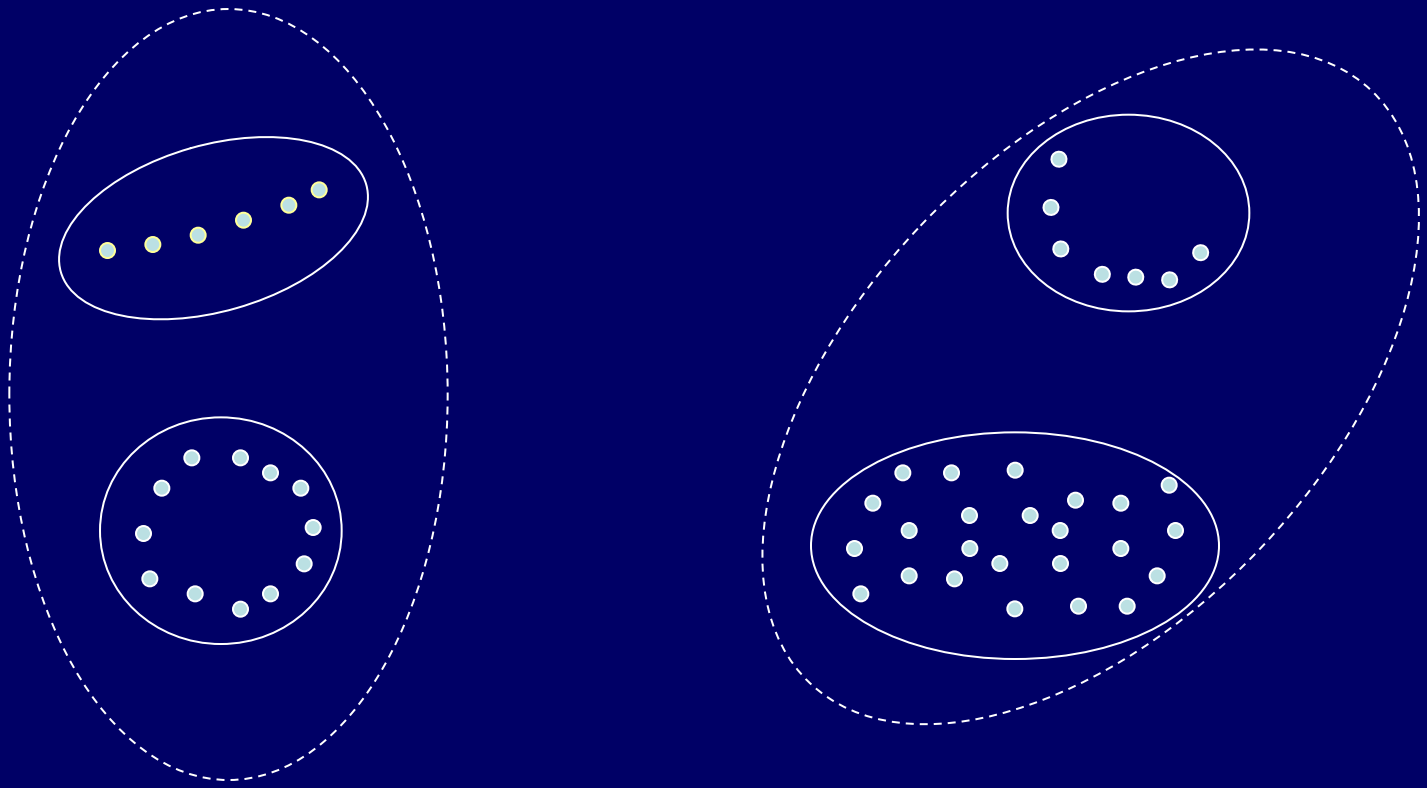
2.1 聚类分析的概念

特征选取不同对聚类结果的影响

(d) 按繁衍后代方式和肺是否存在分

2.1 聚类分析的概念

距离测度不同，聚类结果也不同



数据的粗聚类是两类,细聚类为4类

2.1 聚类分析的概念

综上可见:

选择什么特征?

选择多少个特征?

选择什么样的量纲?

选择什么样的距离测度?

这些对分类结果都会产生极大影响。

2.1 聚类分析的概念

聚类分析的基本思想

特征提取

模式相似性度量

点与类间的距离

类与类间的距离

聚类准则及聚类算法

有效性分析

2.1 聚类分析的概念

聚类算法的主要应用场合

- 在一些情况下，无法获得训练样本
- 可以获得样本，但耗费较多人、财力和时间
- 作为后续较复杂分类算法的预处理
- 用于数据压缩
- 用于数据挖掘，知识发现

第二章　聚类分析

2.1　聚类分析的概念

2.2　模式相似性测度

2.3　类的定义与类间距离

2.4　准则函数

2.5　聚类的算法

2.2　模式相似性测度

用于描述各模式之间特征的相似程度：

- 距 离 测 度
- 相 似 测 度
- 匹 配 测 度

2.2 模式相似性测度

一、距离测度(差值测度)

测度基础：两个矢量矢端的距离

测度数值：两矢量各相应分量之差的函数。

两矢量的距离定义应满足下面的公理：

设矢量 $\vec{x}$ 和 $\vec{y}$ 的距离记为 $d(\vec{x},\vec{y})$，

(1) $d(\vec{x},\vec{y})\geq 0$，当且仅当 $\vec{y}=\vec{x}$ 时，等号成立；

(2) $d(\vec{x},\vec{y})=d(\vec{y},\vec{x})$

(3) $d(\vec{x},\vec{y})\leq d(\vec{x},\vec{z})+d(\vec{z},\vec{y})$

2.2 模式相似性测度

2.2.1距离测度

常用的距离测度有：

1. 欧氏(Euclidean)距离

$$d(\vec{x},\vec{y})=\left\|\vec{x}-\vec{y}\right\|=[\sum_{i=1}^{n}(x_i-y_i)^2]^{1/2}$$

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

2.2 模式相似性测度

2.2.1距离测度

2. 绝对值距离(街坊距离或Manhattan距离)

$$d(\vec{x},\vec{y})=\sum_{i=1}^{n}|x_i-y_i|$$

3. 切氏(Chebyshev)距离

$$d(\vec{x},\vec{y})=\max_i|x_i-y_i|$$

4. 明氏(Minkowski)距离

$$d(\vec{x},\vec{y})=[\sum_{i=1}^{n}|x_i-y_i|^m]^{1/m}$$

2.2 模式相似性测度

2.2.1距离测度

5. 马氏(Mahalanobis)距离

设n维矢量$\vec{x}_i$和$\vec{x}_j$是矢量集$\{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_m\}$中的两个矢量，马氏距离$d$定义为

$$d^2(\vec{x}_i, \vec{x}_j) = (\vec{x}_i - \vec{x}_j)' V^{-1} (\vec{x}_i - \vec{x}_j)$$

其中

$$V = \frac{1}{m-1} \sum_{i=1}^{m} (\vec{x}_i - \vec{\bar{x}})(\vec{x}_i - \vec{\bar{x}})'$$

$$\vec{\bar{x}} = \frac{1}{m} \sum_{i=1}^{m} \vec{x}_i$$

2.2 模式相似性测度

2.2.1距离测度

马氏距离的性质：对一切非奇异线性变换都是不变的。即，具有坐标系比例、旋转、平移不变性，并且从统计意义上尽量去掉了分量间的相关性。

2.2 模式相似性测度

已知一个二维正态母体G的分布为 $N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.9\\0.9 & 1\end{pmatrix}\right)$

求点 $A:\begin{pmatrix}1\\1\end{pmatrix}$ 和 $B:\begin{pmatrix}1\\-1\end{pmatrix}$ 至均值点 $M:\ \vec{\mu}=\begin{pmatrix}0\\0\end{pmatrix}$ 的距离。

解：由题设，可得 $\Sigma=\begin{pmatrix}1 & 0.9\\0.9 & 1\end{pmatrix}$ $\Sigma^{-1}=\dfrac{1}{0.19}\begin{pmatrix}1 & -0.9\\-0.9 & 1\end{pmatrix}$

从而马氏距离：

$$d_M^{\ 2}(A,M)=\begin{pmatrix}1 & 1\end{pmatrix}\Sigma^{-1}\begin{pmatrix}1\\1\end{pmatrix}=0.2/0.19 \qquad d_M^{\ 2}(B,M)=\begin{pmatrix}1 & -1\end{pmatrix}\Sigma^{-1}\begin{pmatrix}1\\-1\end{pmatrix}=3.8/0.19$$

它们之比达 $\sqrt{19}$ 倍，若用欧氏距离，算得的距离值相同。

$$d_E^2(A,M)=2 \qquad d_E^2(B,M)=2$$

由分布函数知，A、B两点的概率密度分别为：

$$p(1,1)=0.2157 \qquad p(1,-1)=0.00001658$$

2.2 模式相似性测度

现金识别例子

数据样本介绍：10个文本文件

文件名：rmb00.txt ……rmb09.txt

每个文件有4个币种的数据，分别是：

100圆、50圆、20圆、10圆

每个币种有新旧两种版本，4个方向，故有8个数据块：

如100圆的8个数据块：

data100a, data100b, data100c, data100d——老版

data100e, data100f, data100g, data100h——新版

每个数据块有8个传感器数据：

传感器1，传感器2，……，传感器8

每个传感器有60个采样数据：

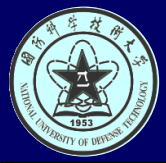
数据1，数据2，……，数据60

2.2 模式相似性测度

现金识别例子—马式平均距离

	100圆	50圆	20圆	10圆
a:	39.73	101.41	162.90	256.38
b:	91.89	230.25	288.69	659.47
c:	103.76	135.94	257.57	724.96
d:	78.58	171.10	330.97	675.90
e:	247.42	443.46	333.93	218.71
f:	108.10	328.11	305.19	607.51
g:	265.16	956.58	818.83	348.42
h:	107.56	339.64	387.10	628.88

其中马式矩阵为100圆A面的，上面是各面到100圆A面的均值点的平均马式距离。



谢谢！