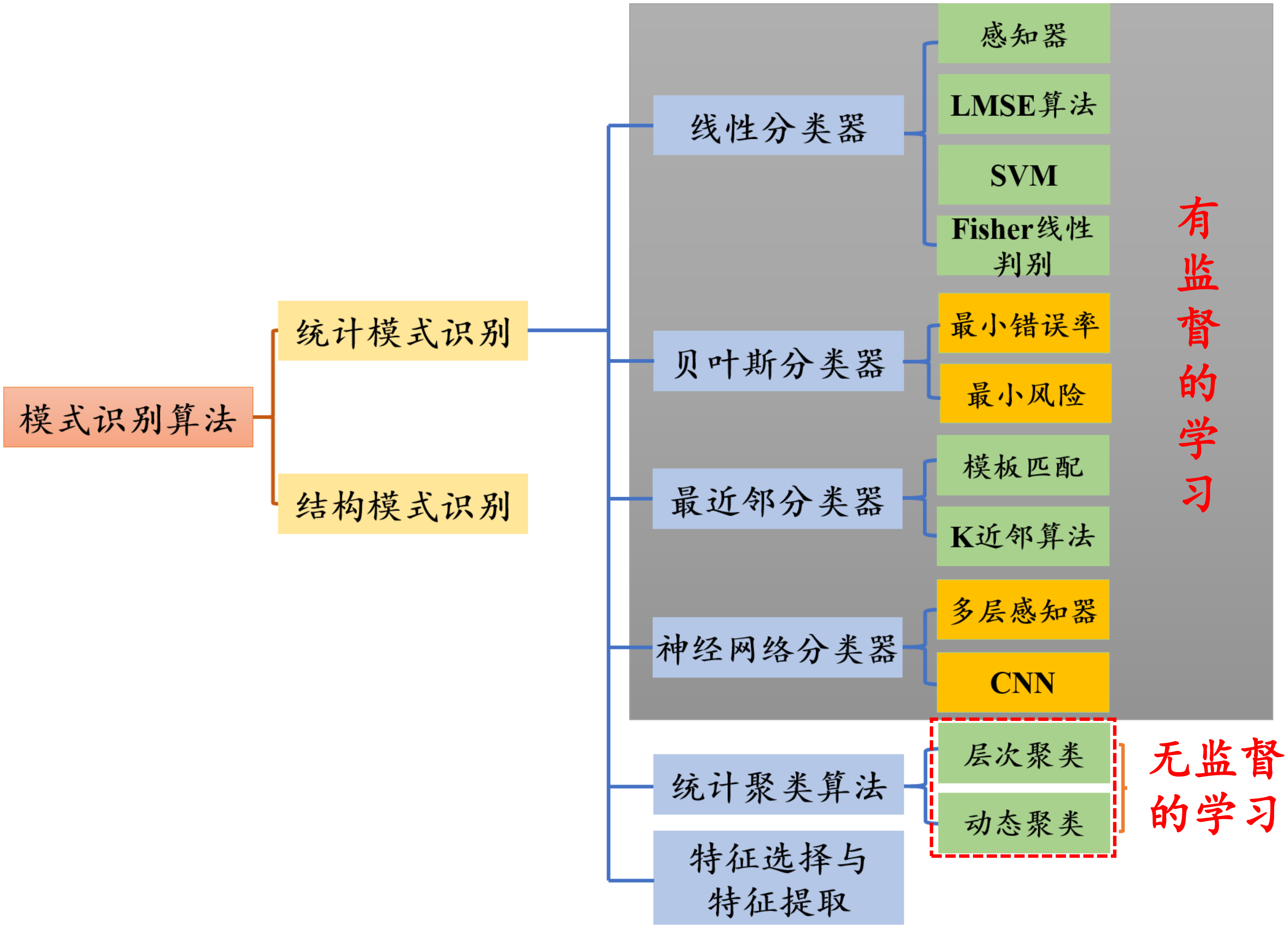
模式识别课程

统计聚类算法

张俊超

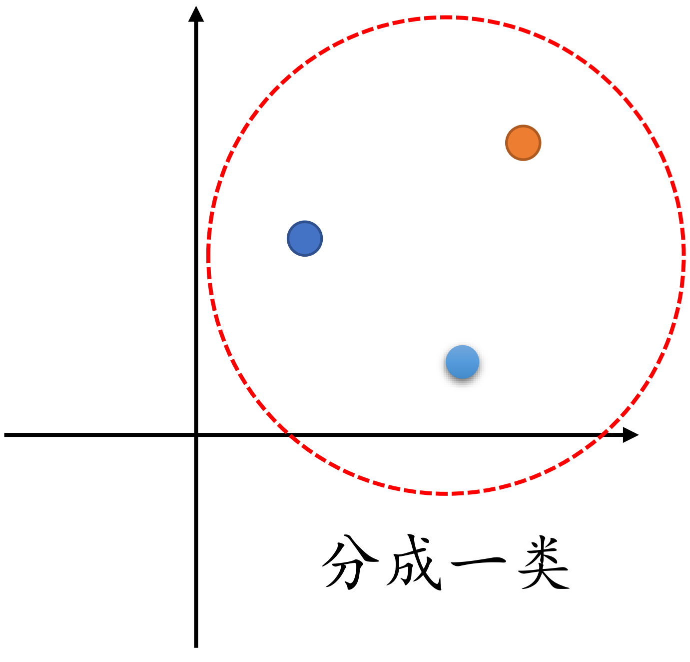
中南大学
航空航天学院

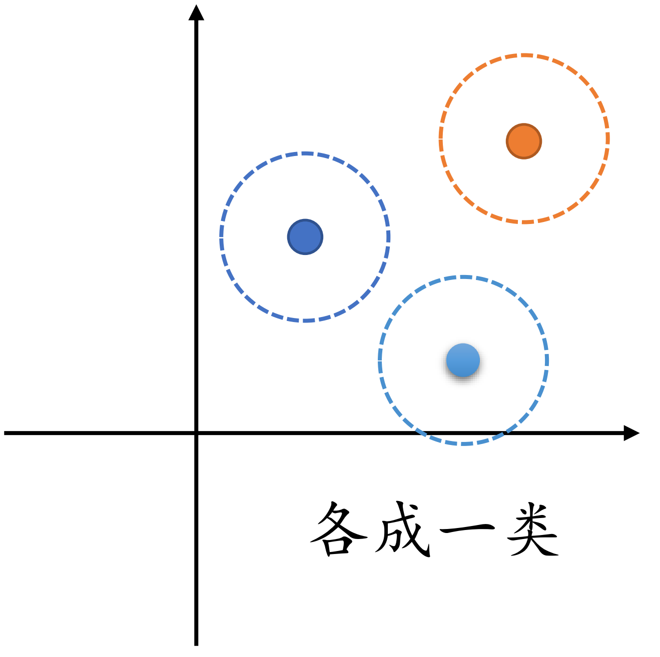
模式识别-统计聚类算法



模式识别算法
统计模式识别
结构模式识别
线性分类器
感知器
LMSE算法
SVM
Fisher线性判别
贝叶斯分类器
最小错误率
最小风险
最近邻分类器
模板匹配
K近邻算法
神经网络分类器
多层感知器
CNN
有监督的学习
统计聚类算法
层次聚类
动态聚类
无监督的学习
特征选择与特征提取

② 层次聚类



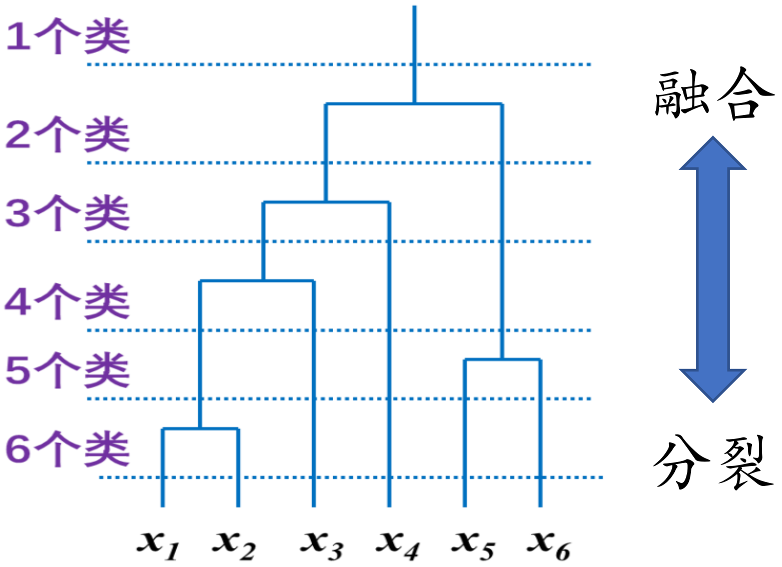


模式识别-统计聚类算法

层次聚类：

- **融合**：从N类→1类
- **分裂**：从1类→N类

本质：从多到少(或从少到多)进行类别划分，求得一系列类别数的划分方案，基于聚类准则，选择适当的划分方案作为聚类的结果。



层次聚类算法的融合算法流程为：

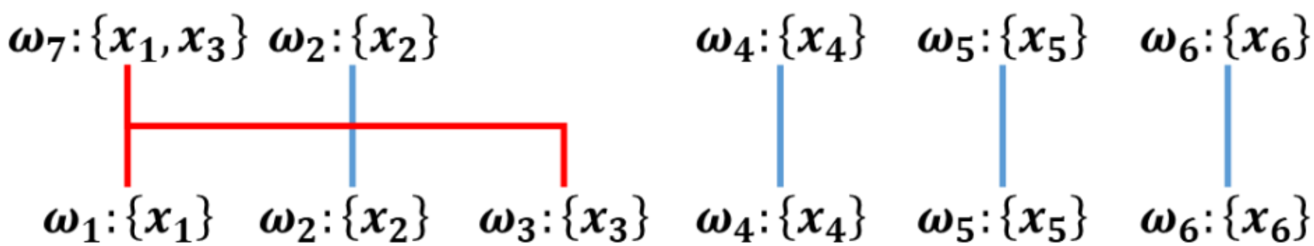
- 对于含 n 个样本的样本集，**先令每个样本自成一类**，总分类数 $c=n$
- 计算类间距离，**将距离最小(最相似)的两个类合并**，总分类数减少为 $c=n-1$
- 继续合并类，直至总分类数 c 或类间距离满足要求

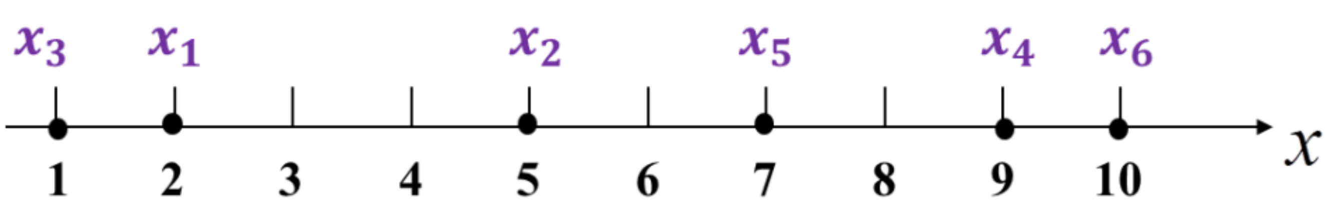
层次聚类算法的分裂算法流程为：

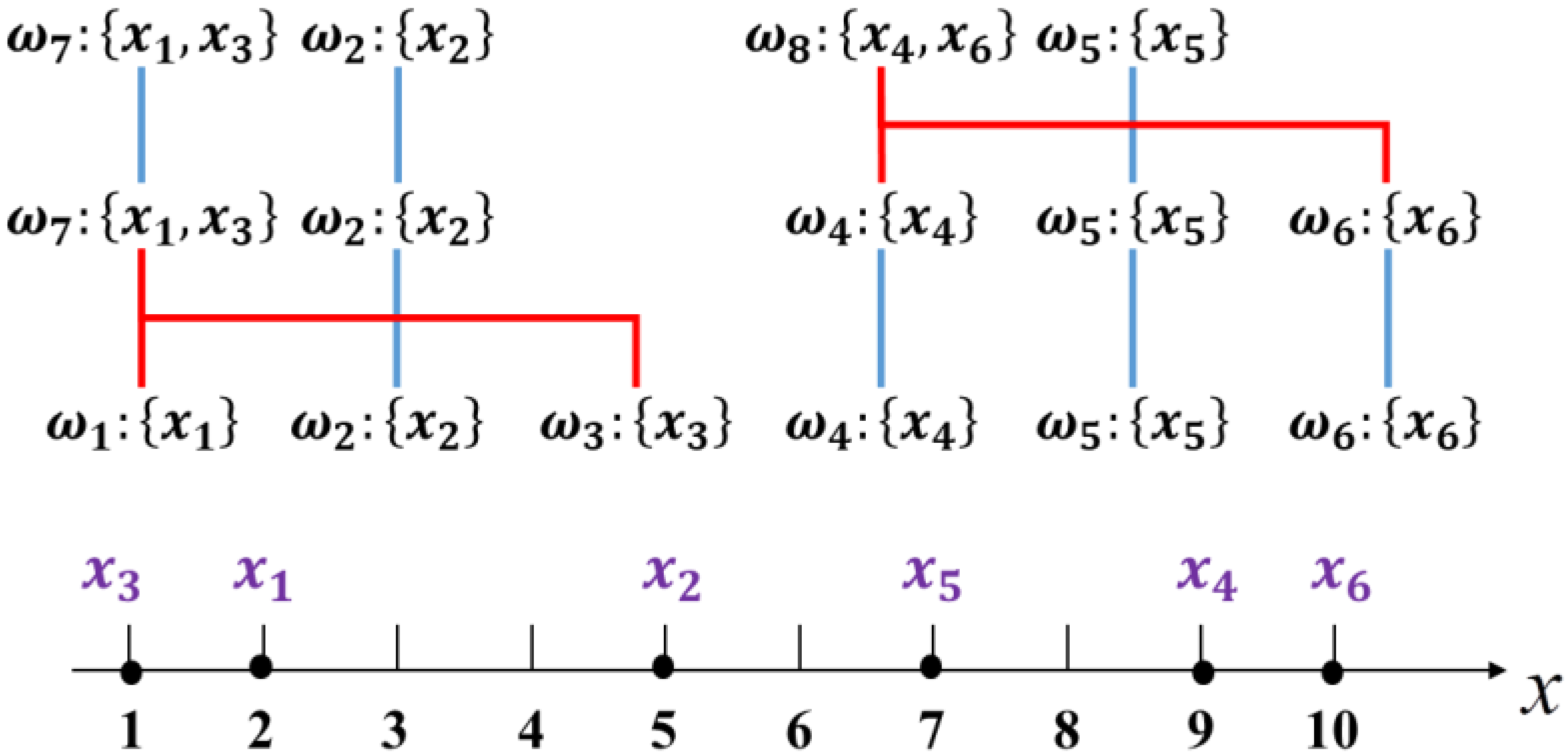
- 对于含 n 个样本的样本集，**先将所有样本作为一类**，总分类数 $c=1$
- 将已得到的类分成两类，计算类间距离，**将类间距离最大(最不相似)的分类方法作为本级分类结果**，总分类数增加为 $c=c+1$
- 对每一个得到的类再进行分类，直至总分类数 c 或类间距离满足要求

融合算法【举例】

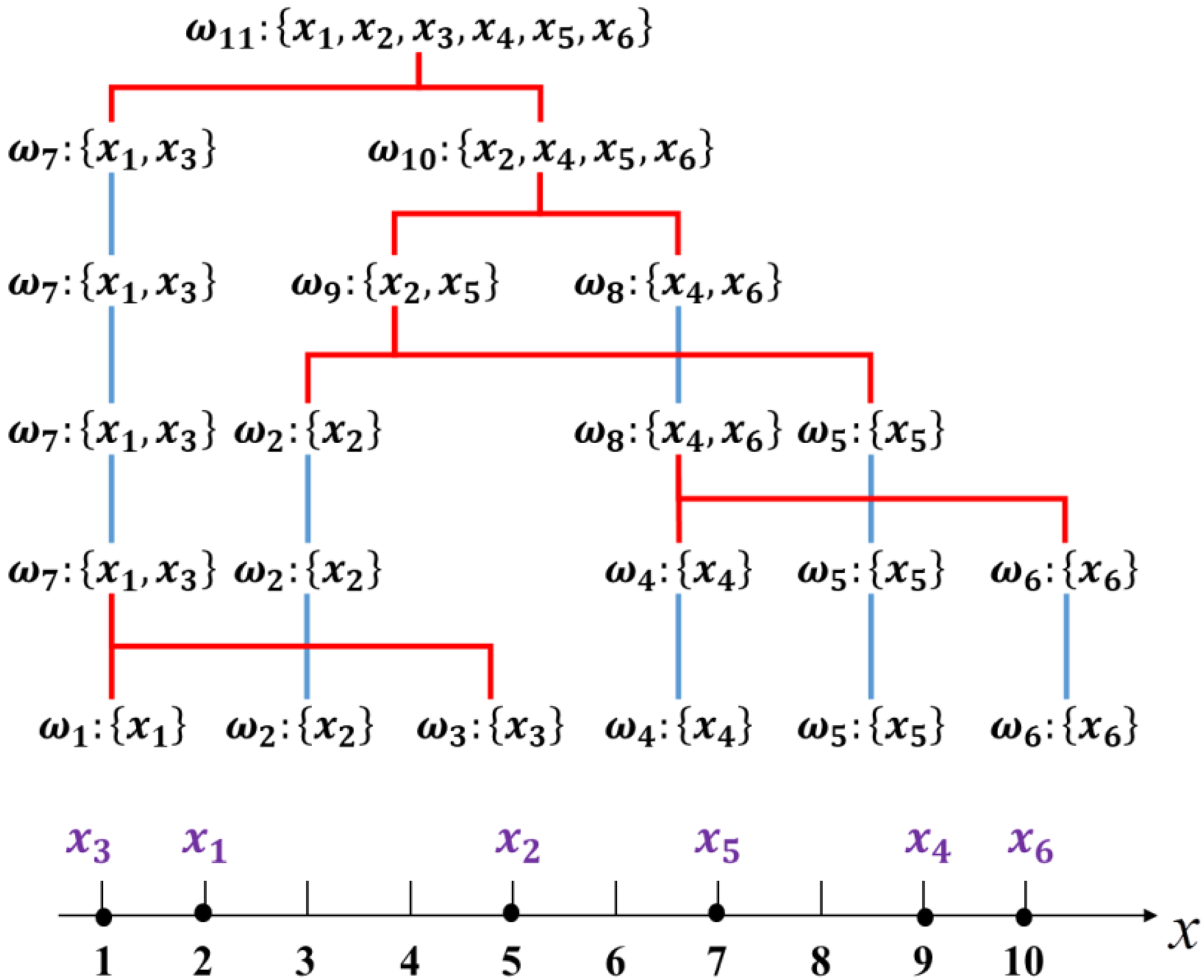
	ω_1	ω_2	ω_3	ω_4	ω_5
ω_2	3				
ω_3	1	4			
ω_4	7	4	8		
ω_5	5	2	6	2	
ω_6	8	5	9	1	3






$\omega_7:\{x_1,x_3\}$
$\omega_2:\{x_2\}$
$\omega_8:\{x_4,x_6\}$
$\omega_5:\{x_5\}$
$\omega_7:\{x_1,x_3\}$
$\omega_2:\{x_2\}$
$\omega_4:\{x_4\}$
$\omega_5:\{x_5\}$
$\omega_6:\{x_6\}$
$\omega_1:\{x_1\}$
$\omega_2:\{x_2\}$
$\omega_3:\{x_3\}$
$\omega_4:\{x_4\}$
$\omega_5:\{x_5\}$
$\omega_6:\{x_6\}$
x_3
x_1
x_2
x_5
x_4
x_6
1 2 3 4 5 6 7 8 9 10
x

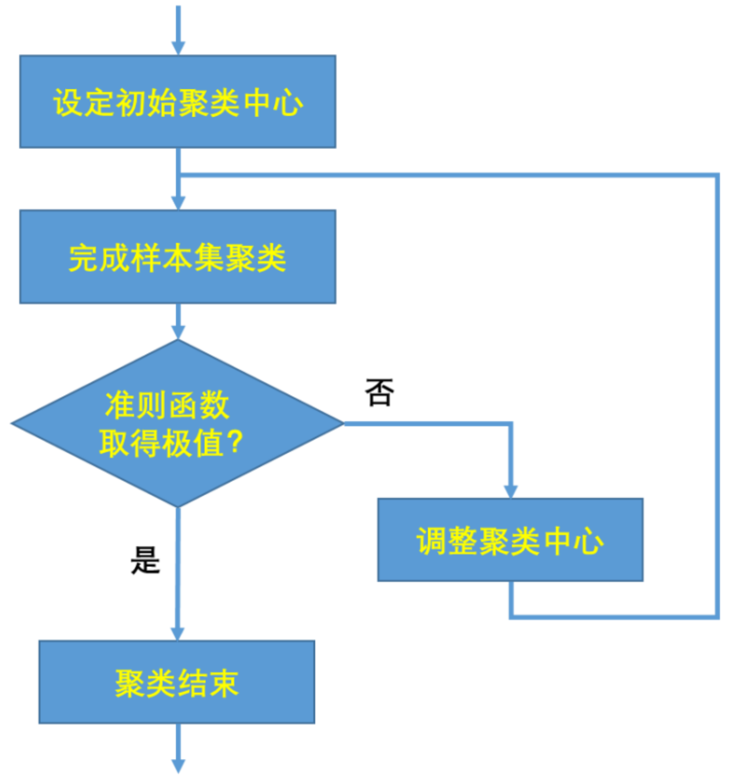
模式识别-统计聚类算法



$\omega_{11}:\{x_1,x_2,x_3,x_4,x_5,x_6\}$
$\omega_7:\{x_1,x_3\}$
$\omega_{10}:\{x_2,x_4,x_5,x_6\}$
$\omega_7:\{x_1,x_3\}$
$\omega_9:\{x_2,x_5\}$
$\omega_8:\{x_4,x_6\}$
$\omega_7:\{x_1,x_3\}$
$\omega_2:\{x_2\}$
$\omega_8:\{x_4,x_6\}$
$\omega_5:\{x_5\}$
$\omega_7:\{x_1,x_3\}$
$\omega_2:\{x_2\}$
$\omega_4:\{x_4\}$
$\omega_5:\{x_5\}$
$\omega_6:\{x_6\}$
$\omega_1:\{x_1\}$
$\omega_2:\{x_2\}$
$\omega_3:\{x_3\}$
$\omega_4:\{x_4\}$
$\omega_5:\{x_5\}$
$\omega_6:\{x_6\}$
x_3 x_1 x_2 x_5 x_4 x_6
1 2 3 4 5 6 7 8 9 10
x

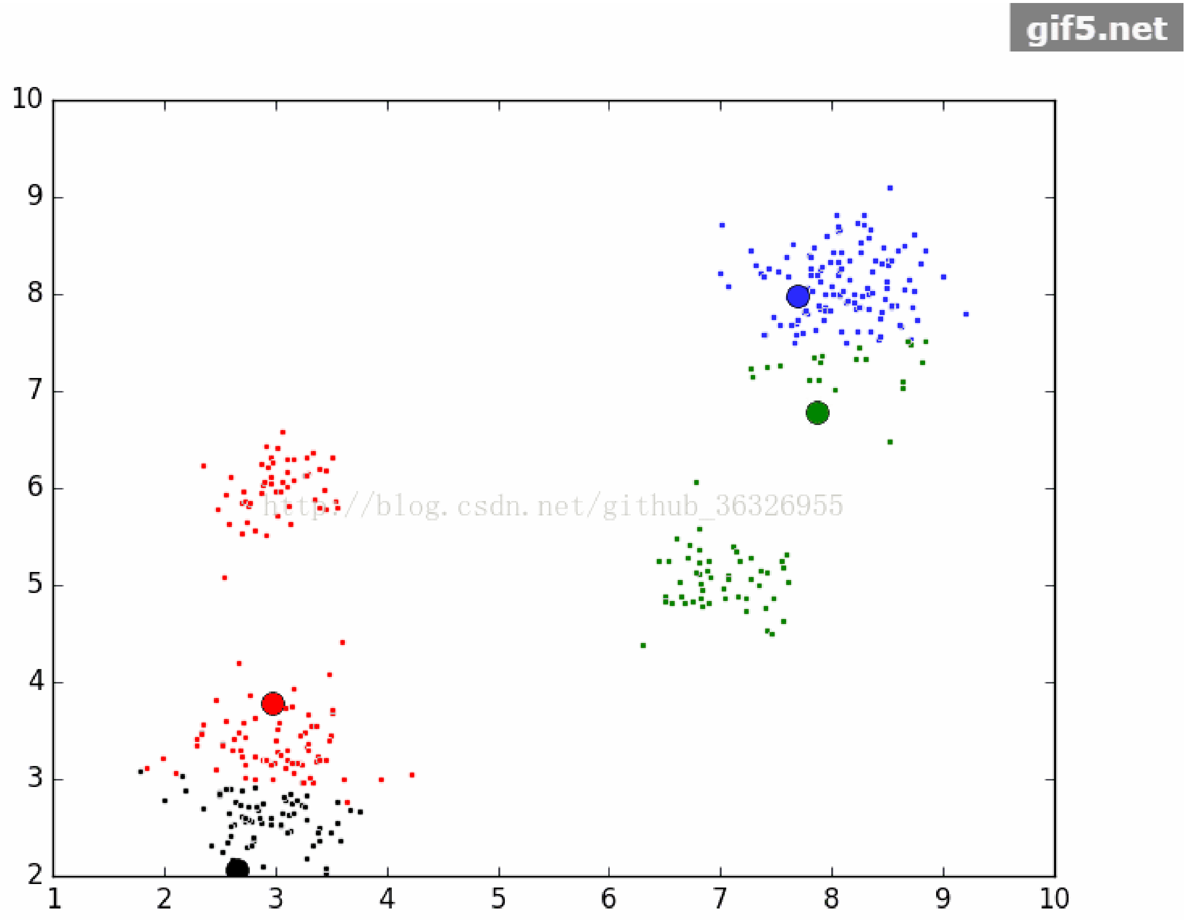
③ 动态聚类

动态聚类算法是一种迭代算法，通过反复修改聚类结果来进行优化，以达到最满意的聚类结果。



模式识别-统计聚类算法

- K-Means算法(C均值算法)



模式识别-统计聚类算法

- K-means算法的基础是：$\boxed{\textbf{最小误差平方和准则}}$

误差平方和的准则函数：$J_e=\sum_{i=1}^{c}\sum_{k=1}^{n_i}\left\|\mathbf{x}_k^{(i)}-\mathbf{m}_i\right\|^2$

$\boxed{n_i:\text{类别}\omega_i\text{的样本总数}}$

$\mathbf{x}_k^{(i)}$:类别ω_i的第k个样本

$\mathbf{m}_i$:类别ω_i的样本均值, $\mathbf{m}_i=\frac{1}{n_i}\sum_{k=1}^{n_i}\mathbf{x}_k^{(i)}$

①类别ω_i中的$\mathbf{x}_k^{(i)}$(记为$\mathbf{y}$)移到类别ω_j中, 则两类的均值变为:

$$\overline{\mathbf{m}_i} = \mathbf{m}_i + \frac{1}{n_i - 1}\left(\mathbf{m}_i - \mathbf{y}\right)$$

$$\overline{\mathbf{m}_j} = \mathbf{m}_j + \frac{1}{n_j + 1}\left(\mathbf{y} - \mathbf{m}_j\right)$$

②两类各自的误差平方和变为:

$$\overline{J_i} = J_i - \frac{n_i}{n_i - 1}\left\|\mathbf{y} - \mathbf{m}_i\right\|^2$$

$$\overline{J_j} = J_j + \frac{n_j}{n_j + 1}\left\|\mathbf{y} - \mathbf{m}_j\right\|^2$$

模式识别-统计聚类算法

$$\overline{J_i} = J_i - \frac{n_i}{n_i - 1}\|\mathbf{y} - \mathbf{m}_i\|^2$$

$$\overline{J_j} = J_j + \frac{n_j}{n_j + 1}\|\mathbf{y} - \mathbf{m}_j\|^2$$

如果减少量大于增加量，则有利于总体误差平方和的减少。即：

$$\frac{n_i}{n_i - 1}\|\mathbf{y} - \mathbf{m}_i\|^2 > \frac{n_j}{n_j + 1}\|\mathbf{y} - \mathbf{m}_j\|^2$$

针对多类情况，将样本**y**移到剩余类别中，其中误差平方和的**最小增加量<减少量**，则把样本**y**归属于最小增加量的类别中。

模式识别-统计聚类算法

- K-means算法的步骤：

(1)初始划分c个聚类中心

(2)任取一个样本$\mathbf{y}$，设$\mathbf{y}\in\omega_i$

(3)若$n_i=1$,则转(2);否则继续

(4)计算ρ_j

$$\begin{cases}\rho_j=\dfrac{n_j}{n_j+1}\left\|\mathbf{y}-\mathbf{m}_j\right\|^2, j\neq i\\ \rho_i=\dfrac{n_i}{n_i-1}\left\|\mathbf{y}-\mathbf{m}_i\right\|^2\end{cases}$$

(5)考查ρ_j中的最小值ρ_k,若$\rho_k<\rho_i$,则把$\mathbf{y}$归属于ω_k

(6)重新计算$\mathbf{m}_i, i=1,2,\ldots,c$和误差平方和$J_e$

(7)若连续N次迭代J_e不改变，则停止；否则转(2)

模式识别-统计聚类算法

- K-means算法II的步骤：

(1)初始划分c个聚类中心

(2)从样本集中依次选取一个样本$\mathbf{y}$，计算相似度$d\left(\mathbf{y},\mathbf{m}_i\right), i=1,2,\ldots,c$

若$d\left(\mathbf{y},\mathbf{m}_k\right)=\min\limits_{i=1,2,\ldots,c}\left\{d\left(\mathbf{y},\mathbf{m}_i\right)\right\}$,则$\mathbf{y}\in\omega_k$

(3)重新计算$\mathbf{m}_i, i=1,2,\ldots,c$

若聚类中心$\mathbf{m}_i, i=1,2,\ldots,c$不再改变，迭代终止；否则转(2)

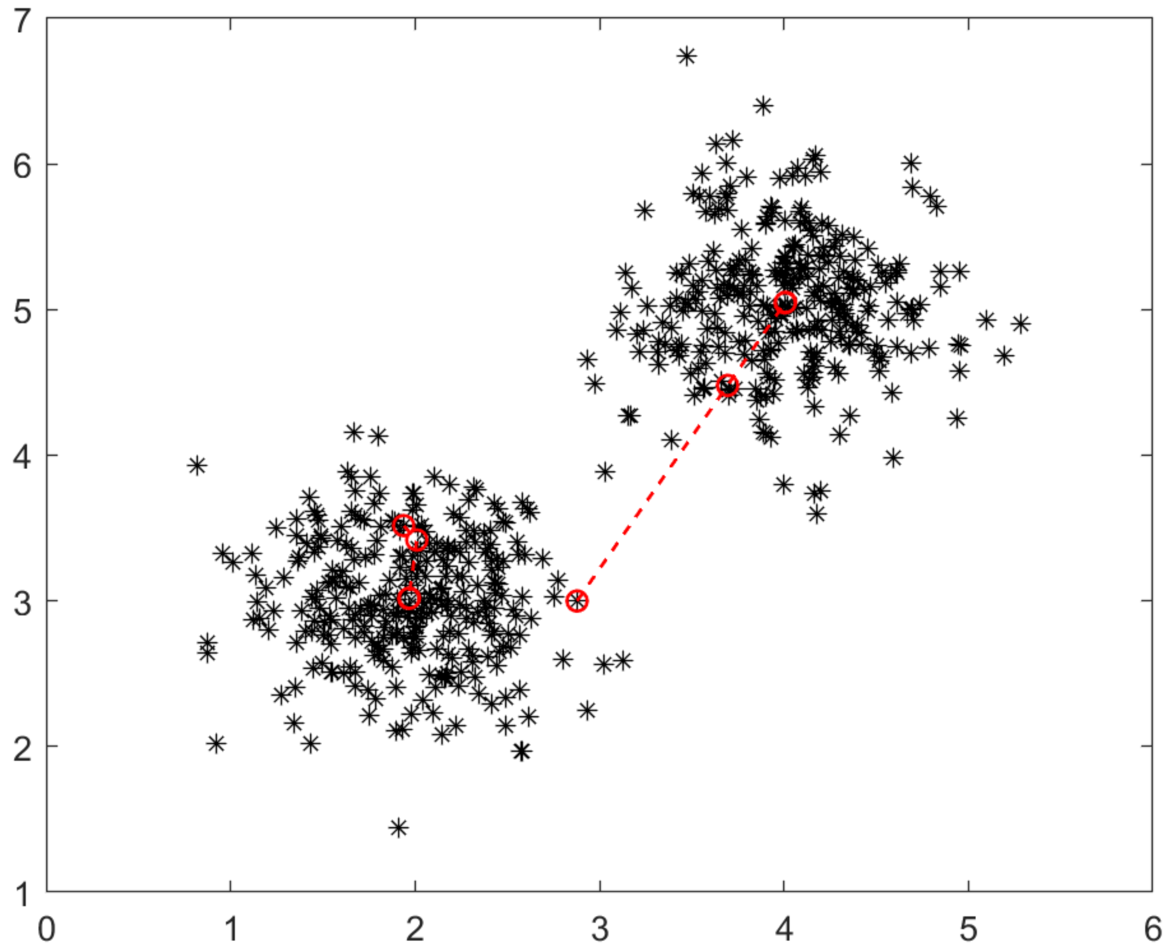
模式识别-统计聚类算法

```
%% K-means
C = 2;%类别数
start_L = randperm(N);
Z = Data(start_L(1:C),:);%初始聚类中心
Labels = zeros(N,1);
Cluster=cell(C,1);
Je=0;
Itera = 1;
Center_pt{Itera} = Z;
while(1)
    Je_old = Je;
    for j = 1:C
        Cluster{j}=[];
    end
    for i = 1:N
        dist = Z - repmat(Data(i,:),[C,1]);
        dist = sum(dist.^2,2);
        [minv,idx] = min(dist);
        Cluster{idx} = [Cluster{idx};Data(i,:)];
```

模式识别-统计聚类算法

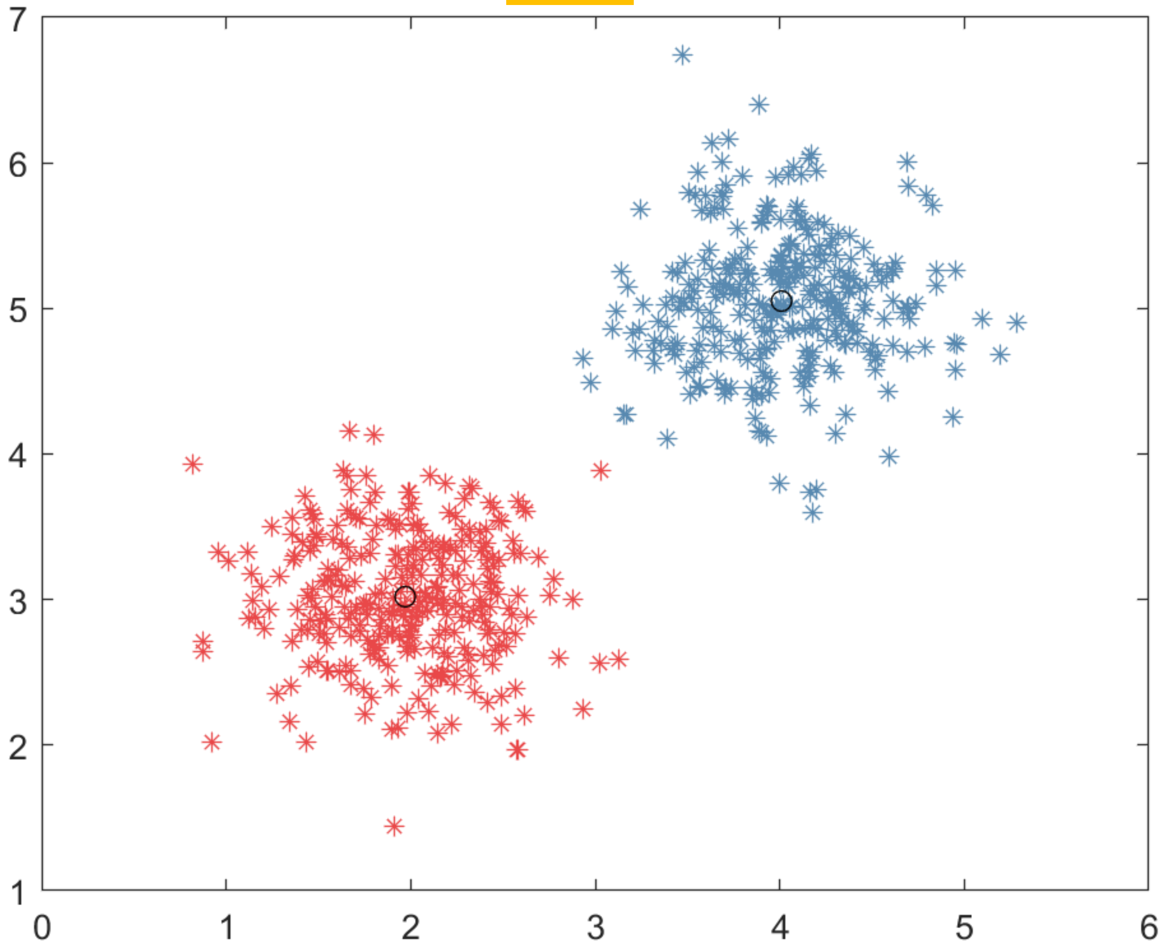
```
    for i = 1:N
        dist = Z - repmat(Data(i,:),[C,1]);
        dist = sum(dist.^2,2);
        [minv,idx] = min(dist);
        Cluster{idx} = [Cluster{idx};Data(i,:)];
        Labels(i) = idx;
    end
    S = 0;
    for k = 1:C
        Z(k,:) = mean(Cluster{k});
        tmp=bsxfun(@minus,Cluster{k},Z(k,:));
        tmp = sum(sum(tmp.^2));
        S = S+tmp;
    end
    Itera = Itera + 1;
    Center_pt{Itera} = Z;
    Je = S;
    if abs(Je-Je_old) < 1e-5
        break;
    end
end
```

模式识别-统计聚类算法



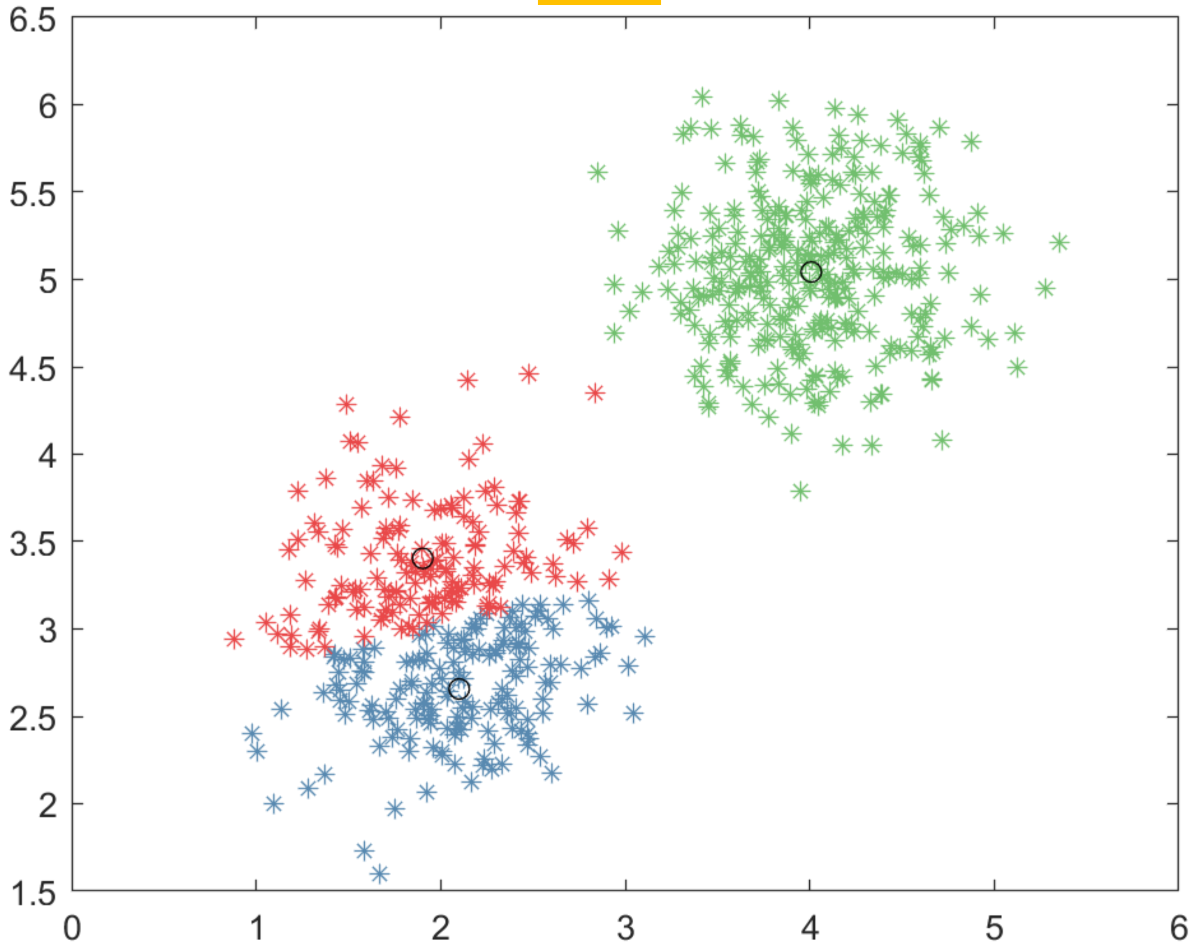
模式识别-统计聚类算法

$c = 2$



模式识别-统计聚类算法

$c=3$



模式识别-统计聚类算法

- K-means: K需要事先确定，且一旦确定，不再改变。
- ISODATA: 动态调整K的值


若干素材取自网络，特此致谢！

谢谢聆听！