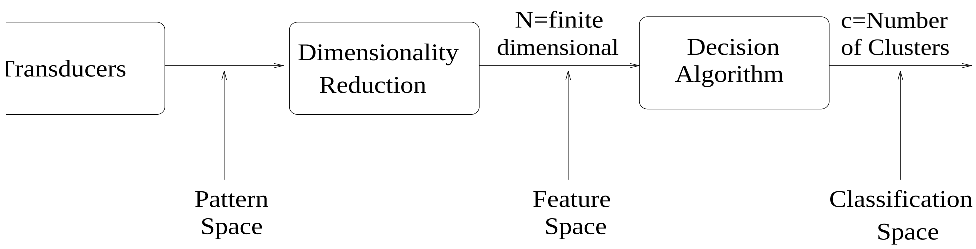
XI. Fuzzy Clustering for Pattern Recognition

Reference : 1. Zimmermann Ch. 11

2. J.C. Bezdek , " Pattern Recognition with Fuzzy Objective Function Algorithms", (1981).

More Reference



Pattern recognition.

Clustering

Once feature extraction is done , the task of clustering is to divide n objects $\{x^1, \cdots, x^n\}$ by p indicators (i.e. $x^i \in \mathbf{R}^p$) into c $(2 \le c < n)$ categorically homogeneous subsets.

Remark 1

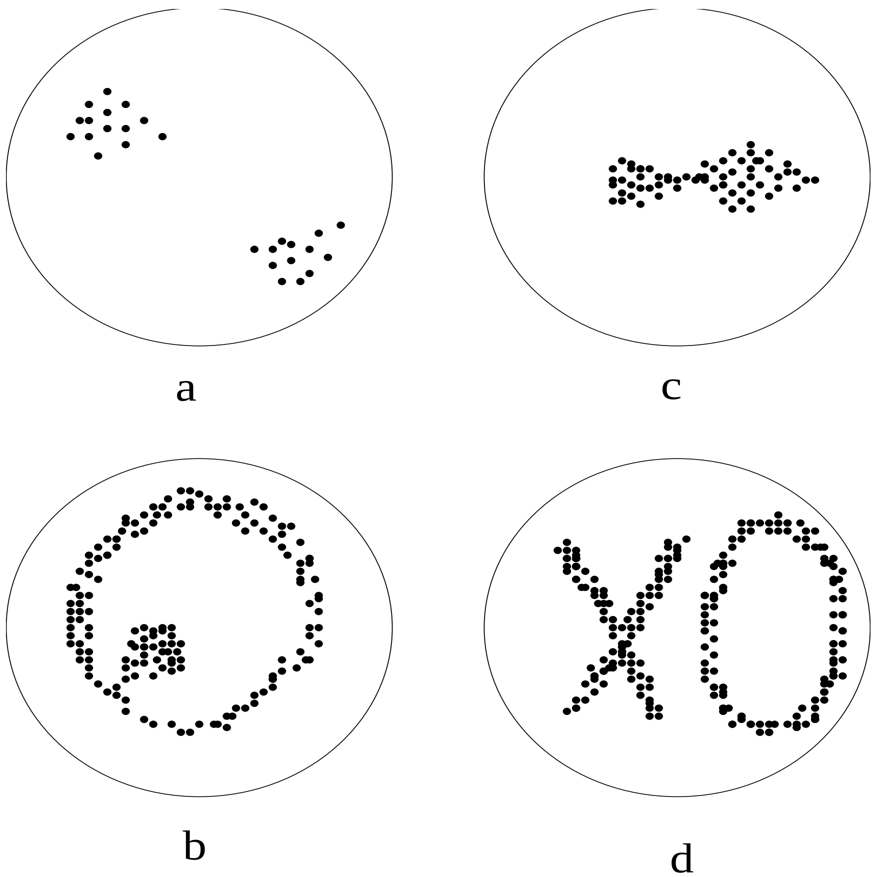
Each subset is called a cluster.

The objects in the same cluster should be <u>similar</u> and the objects of different clusters should be as <u>dissimilar</u> as possible.

Remark 2

The number of clusters , c , is normally unknown in advance.

Question : How ?



Some Possible Shapes of Clusters

Which criterion will lead you to the

“ right clustering ” ?

distance ? connectivity ? intensity ?

centering & variance ? $\cdots$

Common Clustering Methods

(1) Hierarchical

(2) Graph - theoretic

(3) Objective - function methods

(1) Hierarchical Method

Generate a hierarchy of partitions by successive merging and/or splitting of clusters

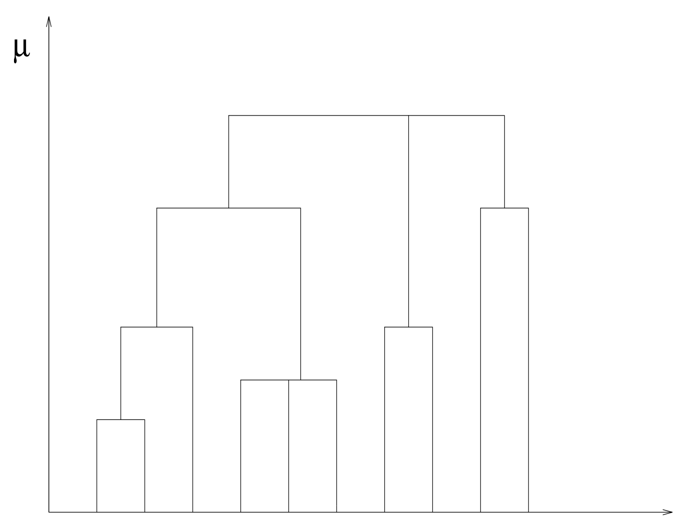


Figure 11-4. Dendogram for hierarchical clusters

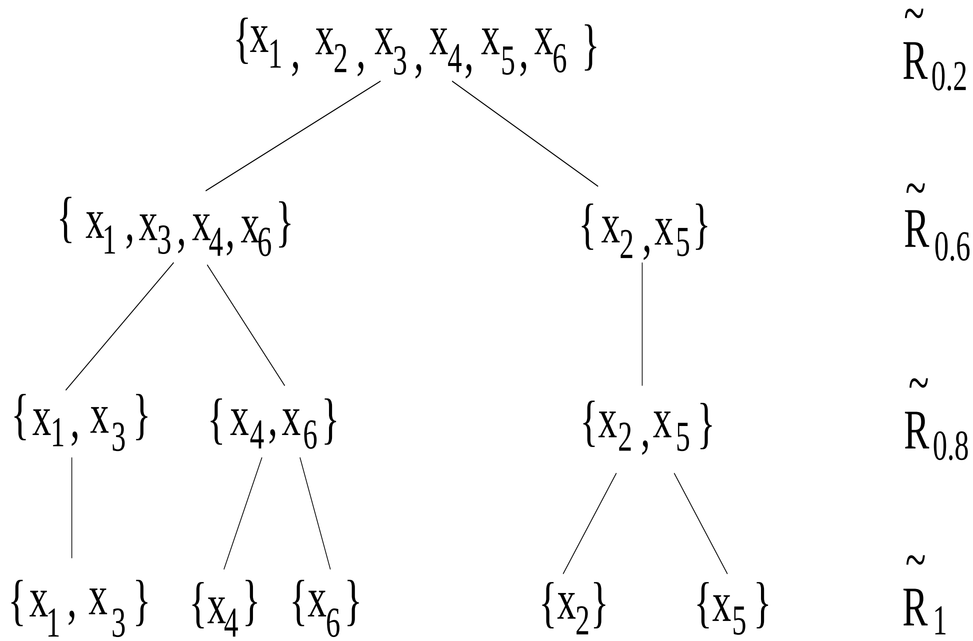
Advantages : Conceptual and computational simplicity .

disadvantages : Not iterative - difficult to change preceding levels

Example : Simmilarity Relation

$$\tilde{R} \triangleq \begin{array}{c|cccccc} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ \hline x_1 & 1 & 0.2 & 1 & 0.6 & 0.2 & 0.6 \\ x_2 & 0.2 & 1 & 0.2 & 0.2 & 0.8 & 0.2 \\ x_3 & 1 & 0.2 & 1 & 0.6 & 0.2 & 0.6 \\ x_4 & 0.6 & 0.2 & 0.6 & 1 & 0.2 & 0.8 \\ x_5 & 0.2 & 0.8 & 0.2 & 0.2 & 1 & 0.2 \\ x_6 & 0.6 & 0.2 & 0.6 & 0.8 & 0.2 & 1 \end{array}$$

$\tilde{R}$: Reflexive , symmetric , max - min tansitive



(2) <u>Graph - theoretic Method</u>

Check connectivity and break edges in a minimal spanning tree to form subgraphs

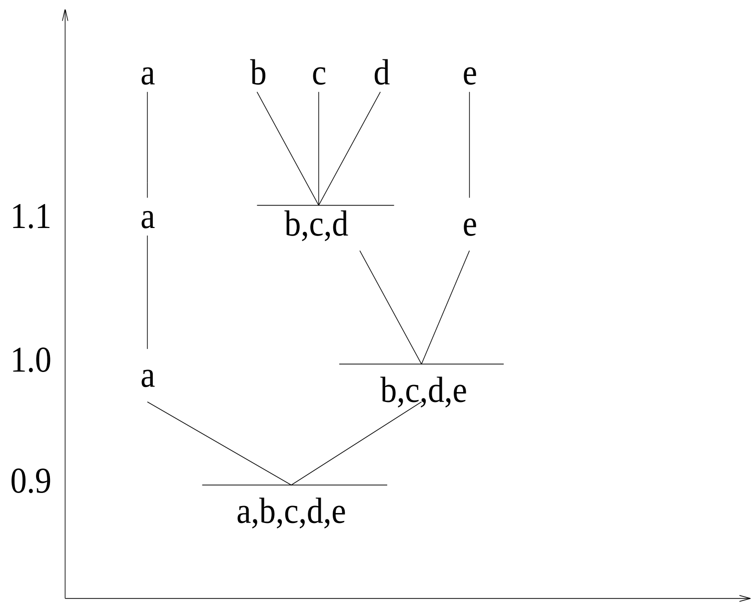


Figure 11-6 Dendogram for graph-theoretic clusters.

(3) Objective - function Methods

The “ desirability ” of clustering candidates is measured for each c by an objective function .

One frequently used method is the so-called $c-$mean algorithm , which defines “center of clusters ” and minimizing the total “spread ” around those centers .

c - mean Method

$$\mathbf{X} \triangleq \{ x^1, \cdots, x^n \}$$

$\tilde{S}_i$: clusters $\qquad i = 1, 2, \cdots, c$

$\mu_{\tilde{S}_i} : \mathbf{X} \longrightarrow [0, 1]$

$$x^k \longrightarrow \mu_{ik} \triangleq \mu_{\tilde{S}_i}(x^k)$$

Definition 1 : For a given integer $2 \leq c < n$, let

$V_{cn} \triangleq$ {all real matrix with dimensionality $c \times n$}.

The matrix $U = [\mu_{ik}] \in V_{cn}$ is a "crisp $c-$ partitioning"

if (a) $\mu_{ik} \in \{0, 1\}$, for $1 \leq i \leq c$, $1 \leq k \leq n$.

(b) $\Sigma_{i=1}^{c} \mu_{ik} = 1$, for $1 \leq k \leq n$.

(c) $0 < \Sigma_{k=1}^{n} \mu_{ik} < n$, for $1 \leq i \leq c$.

also $M_c \triangleq$ {all crisp $c-$ partitioning of $\mathbf{X}$}

Example :

$$\mathbf{X} = \{x_1, x_2, x_3\}$$

$$c = 2$$

$$U_1 = \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{l} \\ \leftarrow \text{ cluster 1 } S_1 = \{x_1, x_2\} \\ \leftarrow \text{ cluster 2 } S_2 = \{x_3\} \end{array}$$

$$U_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

$$U_3 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

How about

$$U_4 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \quad ?$$

$$U_5 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad ?$$

and $\quad M_c = ?$

Definition 2 : Same as in Definition 1 , the matrix

$\tilde{U} = [\mu_{ik}] \in V_{cn}$ is a "fuzzy $c-$ partitioning ",

if

(a) $\mu_{ik} \in [0, 1], \qquad \forall\ i, k$

(b) $\Sigma_{i=1}^{c} \mu_{ik} = 1, \qquad \forall\ k$

(c) $0 < \Sigma_{k=1}^{n} \mu_{ik} < n, \quad \forall\ i$

Also $M_{fc} \stackrel{\triangle}{=} \{\text{all fuzzy } c - \text{partitioning of } \mathbf{X}\}$

Example :

$$\mathbf{X} = \{x_1, x_2, x_3\}$$

$$\begin{array}{cl} & \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \tilde{U}_1 = & \begin{bmatrix} 1 & 0.5 & 0 \\ 0 & 0.5 & 1 \end{bmatrix} \begin{array}{l} \leftarrow \text{ cluster 1} \quad \tilde{S}_1 = \{(x_1, 1), (x_2, 0.5)\} \\ \leftarrow \text{ cluster 2} \quad \tilde{S}_2 = \{(x_2, 0.5), (x_3, 1)\} \end{array} \end{array}$$

$$\tilde{U}_2 = \begin{bmatrix} 0.7 & 0.4 & 0.8 \\ 0.3 & 0.6 & 0.2 \end{bmatrix}$$

$$\tilde{U}_3 = \begin{bmatrix} 0 & 0.99 & 0.8 \\ 1 & 0.01 & 0.2 \end{bmatrix}$$

For the " butterfly "

$$\begin{array}{cc} & \begin{array}{ccccccccccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 & x_{10} & x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \end{array} \\ \tilde{U} = & \begin{bmatrix} .86 & .97 & .86 & .94 & .99 & .94 & .86 & .5 & .14 & .06 & .01 & .06 & .14 & .03 & .14 \\ .14 & .03 & .14 & .06 & .01 & .06 & .14 & .5 & .86 & .94 & .99 & .94 & .86 & .97 & .86 \end{bmatrix} \end{array}$$

$$M_{fc} = ?$$

Definition 3 Given that $1 < c < n$ is known ,

$U \in M_c$ and $S_1, S_2, \cdots, S_c$ are clusters

defined by U, then

$$\mathrm{v}^i \triangleq \frac{1}{\mid S_i \mid} \sum_{x^k \in S_i} x^k, \qquad i = 1, \cdots, c$$

are called "<u>cluster centers</u>".

Remark 1 :

$$\mathrm{v}^i = \frac{1}{\sum_{k=1}^n \mu_{ik}} \sum_{k=1}^n \mu_{ik} x^k \qquad \forall\ i$$

Remark 2 :

$$d_{ik} \triangleq d(x^k, \mathrm{v}^i) = [\Sigma_{j=1}^p (x_{kj} - v_{ij})^2]^{1/2}$$

and $\Sigma_{x^k \in S_i} d_{ik}^2$ is the <u>variance of cluster</u> i

$$\|$$

$$\Sigma_{k=1}^n \mu_{ik} d_{ik}^2$$

The crisp c-mean method takes the minimum variance as objective function and consider the following problem

$$\text{Min} \quad z(U) = \Sigma_{i=1}^{c} \Sigma_{k=1}^{n} \mu_{ik} \|x^k - \mathrm{v}^i\|^2$$

$$\text{s.t.} \quad \mathrm{v}^i = \frac{1}{\mid S_i \mid} \sum_{x^k \in S_i} x^k, \qquad i = 1, \cdots, c$$

$$U \in M_c$$

Definition 4 Given that $1 < c < n$ is known ,

$\tilde{U} \in M_{fc}$, then

$$\mathrm{v}^i = \frac{1}{\sum_{k=1}^{n} \mu_{ik}} \sum_{k=1}^{n} \mu_{ik} x^k \qquad \forall\ i$$

are " cluster centers "

The fuzzy $c-$mean method considers the following

problem :

$$\text{Min} \quad z(\tilde{U}) = \textstyle\sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_{ik})^m \|x^k - \mathrm{v}^i\|^2$$

$$\text{s.t.} \quad \mathrm{v}^i = \frac{1}{\sum_{i=1}^{n} (\mu_{ik})^m} \sum_{k=1}^{n} (\mu_{ik})^m x^k \qquad \forall\ i$$

$$\tilde{U} \in M_{fc}$$

where $m \geq 1$ is a given number .

Remark 3 : For the above " m - weighted " model , the x^k with higher degree of membership has higher influence on v^i than those with lower degree of membership. The tendence is amplified for $m > 1$.

Remark 4 : Let G be a symmetric and positive - definite $p \times p$ matrix , then

$$\|x^k - \mathrm{v}^i\|_G^2 \triangleq (x^k - \mathrm{v}^i)^T G (x^k - \mathrm{v}^i)$$

defines a G - norm.

Remark 5 : When $G = I$, $\|x^k - \mathrm{v}^i\|_G^2 = \|x^k - \mathrm{v}^i\|^2$

therefore G-norm is more general .

The general fuzzy c-mean method considers the following problem, given $m \geq 1$ and G are known ,

$$\text{Min} \quad z_m(\tilde{U}; V) = \Sigma_{k=1}^{n} \Sigma_{i=1}^{c} (\mu_{ik})^m \|x^k - \mathrm{v}^i\|_G^2$$

$$(P_m) \quad \text{s.t.} \quad \tilde{U} \in M_{fc}$$

$$V \in \mathbf{R}^{cp}$$

Question : How to solve (P_m) ?

Necessary for a local optimum

$$\mathrm{v}^i = \frac{1}{\sum_{k=1}^{n}(\mu_{ik})^m}\sum_{k=1}^{n}(\mu_{ik})^m x_k \qquad i = 1, \cdots, c \qquad (11.1)$$

$$\mu_{ik} = \frac{\left(\frac{1}{\|x_k - \mathrm{v}^i\|_G^2}\right)^{1/(m-1)}}{\sum_{j=1}^{c}\left(\frac{1}{\|x_k - \mathrm{v}^j\|_G^2}\right)^{1/(m-1)}}, \qquad i = 1, \cdots, c;\ k = 1, \cdots, n \qquad (11.2)$$

Remark : v^i is determined by μ_{ik} while

μ_{ik} is determined by v^i

Fuzzy c-mean algorithm

Input data:

the number of clusters c, $2 \le c \le n$;

the exponential weight m, $1 < m < \infty$;

the $(p \times p)$ matrix G(G symmetric and positive-definite) which induces a norm;

the method to initialize the membership matrix $\tilde{U}^{(0)}$

the termination criteria $\triangle = \|\tilde{U}^{(l+1)} - \tilde{U}^{(l)}\|_G \le \epsilon$.

Procedure

Step 1. Choose $c(2 \le c \le n), m(1 < m < \infty)$ and the $(p \times p)$-matrix G with G symmetric and positive-definite.

Initialize $\tilde{U}^{(0)} \in M_{fc}$, set $l = 0$.

Step 2. Calculate the c fuzzy cluster centers $\{\nu^{i(l)}\}$ by using $\tilde{U}^{(l)}$ from condition (11.1).

Step 3. Calculate the new membership matrix $\tilde{U}^{(l+1)}$ by using $\{\nu^{i(l)}\}$ from condition (11.2) if $x_k \ne \nu^{i(l)}$. Else set

$$\mu_{jk} = \begin{cases} 1 & \text{for} j = i \\ 0 & \text{for} j \ne i \end{cases}$$

Step 4. Choose a suitable matrix norm and caculate $\triangle = \|\tilde{U}^{(l+1)} - \tilde{U}^{(l)}\|_G$. If $\triangle > \epsilon$ set $l = l + 1$ and go to step 2. If $\triangle \leq \epsilon \rightarrow$ stop.

For the fuzzy $c-$means algorithm a number of parameters have to be chosen :

Input data

(1) What is an optimal c ?

(2) What is an optimal m ?

In particular , $m \to \infty, \; \tilde{U} = [\frac{1}{c}]$

(3) G determins the shape of cluster , for example

$$G = [\mathrm{diag}(\sigma_j^2)]^{-1}$$

$\downarrow$

variance of feature j

rescales the data spread .

(4) How to find a good starting $\tilde{U}_0$?

Output Analysis :

" Cluster Validity " .

- an indicator of the quality of a clustering solution

Best known measures are

(partition coefficient) $F(\tilde{U}, c) \triangleq \Sigma_{k=1}^{n} \Sigma_{i=1}^{c} \frac{(\mu_{ik})^2}{n}$

(partition entropy) $H(\tilde{U}, c) \triangleq -\frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik} log_e(\mu_{ik})$

(proportion exponent) $P(\tilde{U}, c) \triangleq -log_e\{\Pi_{k=1}^{n}[\Sigma_{j=1}^{[\mu_k^{-1}]}(-1)^{j+1} \begin{pmatrix} c \\ j \end{pmatrix}$

$(1 - j\mu_k)^{(c-1)}]\}$

where $\mu_k = \max_{1 \le i \le c}\{\mu_{ik}\}$

and $[\mu_k^{-1}] =$ greatest integer $\le (\frac{1}{\mu_k})$

Remark 1

$$\frac{1}{c} \leq F(\tilde{U}, c) \leq 1$$

$$0 \leq H(\tilde{U}, c) \leq log_e c$$

$$0 \leq P(\tilde{U}, c) < \infty$$

Remark 2

Extrema for crisp partitions $U \in M_c$

$$F(\tilde{U}, c) = 1 \;\Leftrightarrow\; H(\tilde{U}, c) = 0 \;\Leftrightarrow\; \tilde{U} \in M_c$$

$$F(\tilde{U}, c) = \frac{1}{c} \;\Leftrightarrow\; H(\tilde{U}, c) = log_e(c) \;\Leftrightarrow\; \tilde{U} \in [\frac{1}{c}]$$

Remark 3

The (heuristic) rules for selecting the " correct " or best partitions are :

$$\max_c\{\max_{\tilde{U} \in \Omega_c}\{F(\tilde{U}, c)\}\} \quad c = 2, \cdots, n-1$$

$$\min_c\{\min_{\tilde{U} \in \Omega_c}\{H(\tilde{U}, c)\}\} \quad c = 2, \cdots, n-1$$

where Ω_c is the set of all " optimal " solutions for given c.

The heuristic for choosing a good partition is

$$\max_c\{\max_{\tilde{U} \in \Omega_c}\{P_i(\tilde{U}, c)\}\} \quad c = 2, \cdots, n-1$$