Soft Computing

Clustering - II

Dr. Chun-Hao Chen

Outline

1、Clustering Validation Techniques

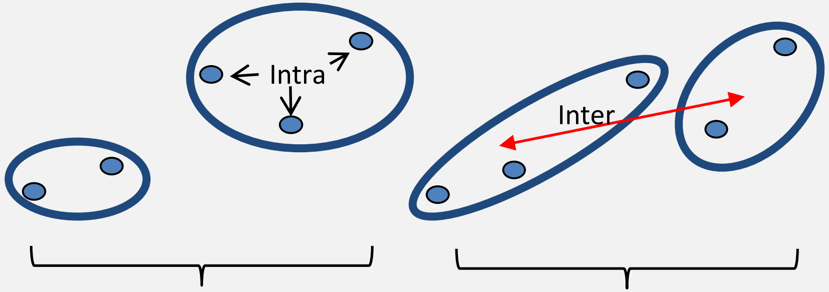
2、An Object belongs to Many Clusters

3、Clustering without Number of Clusters

4、Today’s Extra Task

How to Evaluate Clustering Results

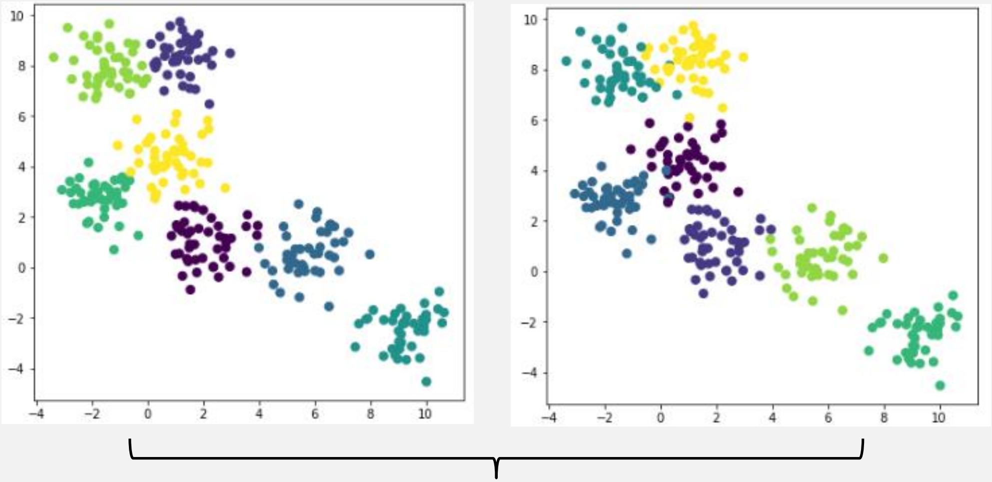
✓ Given the following clustering results



Good

Bad

https://towardsdatascience.com/spectral-graph-clustering-and-optimal-number-of-clusters-estimation-32704189afbe



Which clustering result is better?

➡ We need clustering validation techniques

Clustering Validation Techniques

- ✓ Types of validation techniques
  - ● External indices
    - ✓ Based on some "gold standards"
    - ✓ Validate a partition by comparing it with the correct partition (Arbelaitz et al.'13)
  - ● Internal indices
    - ✓ Based on some statistics of the results
    - ✓ Validate a partition by examining just the partitioned data (Arbelaitz et al.'13)

Validation Techniques – External Indices

✓ Given two binary matrices A and B of the same dimensions

		B	
		1	0
A	1	a	b
	0	c	d

- Matching coefficient: $(a+d) / (a+b+c+d)$
- Jaccard coefficient: $a / (a+b+c)$

An Example - Jaccard coefficient (↑)

✓ Jaccard coefficient → $0 \leq J(A, B) \leq 1$

- Proportion of dividing instances into correct groups
- Given two sets as follows:

		B	
		1	0
A	1	*a*	*b*
	0	*c*	*d*

Jaccard coefficient:

$a / (a+b+c)$

Set A: Clustering result of cluster A = {2, 4, 6}

Set B: Ground true of the cluster B = {0, 1, 2, 3, 4, 5, 6}

$$
\begin{aligned}
J(A, B) &= |A \cap B| \;/\; |A \cup B| \\
&= |2, 4, 6| \;/\; |0, 1, 2, 3, 4, 5, 6| \\
&= 0.5
\end{aligned}
$$

An Example - Hubert's Γ Statistics (↑)

✓ $X=[X(i,j)]$ and $Y=[Y(i,j)]$ are two $n \times n$ matrix

- $X(i, j)$: similarity of object i and object j
- $Y(i,j)=\begin{cases}1 & \text{if objects } i \text{ and } j \text{ are in same cluster,}\\ 0 & \text{otherwise.}\end{cases}$
- Hubert's Γ statistic represents the point serial correlation:

$$\Gamma=\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\frac{X(i,j)-\overline{X}}{\sigma_X}\right)\left(\frac{Y(i,j)-\overline{Y}}{\sigma_Y}\right), -1\leq\Gamma\leq 1$$

where $M = n\,(n - 1) / 2$ is the number of entries in the double sum

- A higher value of Γ represents the better clustering quality

An Example - Hubert's Γ Statistics (Cont.)

- ✓ $X=[X(i, j)]$ and $Y=[Y(i, j)]$ are two $n \times n$ matrix
- ✓ Let X and Y after standardization are shown as follows

$X(i, j)$: similarity of object *i* and object *j*

object *j*

object *i*

0	1.0	0.6	0.2
	0	0.3	0.4
		0	0.1
			0

Normalization

0	1.73	0.51	-0.71
	0	-0.4	-0.10
		0	-1.02
			0

$Y = [Y(i, j)]$: A clustering result

0	1	0	1
	0	1	0
		0	1
			0

Normalization

0	0.64	-1.29	0.64
	0	0.64	-1.29
		0	0.64
			0

$$\Gamma = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\frac{X(i,j)-\overline{X}}{\sigma_X}\right)\left(\frac{Y(i,j)-\overline{Y}}{\sigma_Y}\right), -1\le \Gamma \le 1$$

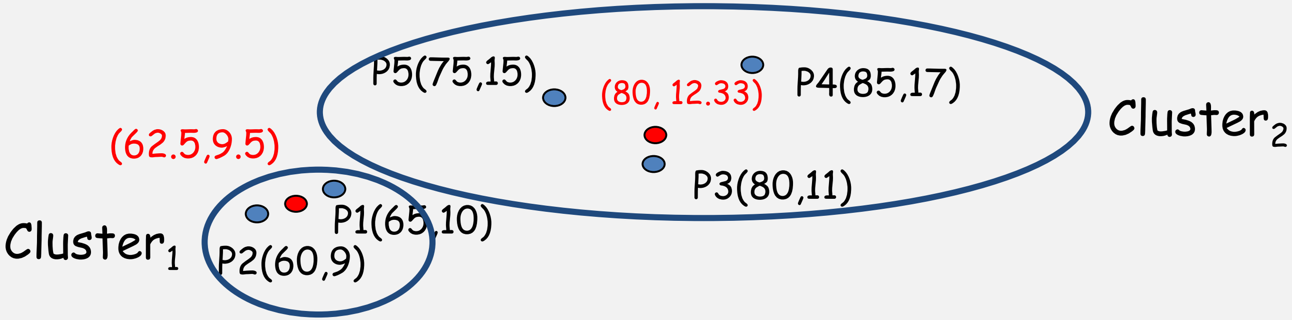
Hence, the value of Γ = (1.73*0.64 + 0.51*-1.29 + …+ -1.02*0.64) / 6
= -0.79 / 6 = -0.13

Other Validation Indices

- ✓ Cluster Validation Indices
  - C-index (Hubert and Schultz, 1976)
  - Davis-Bouldin index (Davies and Bouldin, 1979)
  - Dunn’s index (Dunn, 1974)
  - Goodman-Kruskal index (Goodman and Kruskal, 1954)
  - Silhouette index (Rousseeuw, 1987)

Discussion 1

k-means clustering



In *k*-means clustering algorithm, each data point should belong to a group. Do you think it reasonable?

An Example

✓ASUS

- ZenBook Duo UX481FL → Laptop
- ZenFone 6 (ZS630KL) → Smart Phone

“Computer Manufacturing”

or

“Smart Phone Manufacturing”

or

Both?

Two Solutions

- ✓Solution I - PoCluster algorithm
- ✓Solution II - Soft clustering algorithm (Fuzzy c-means algorithm)
  - By applying soft computing on the k-means clustering

PoCluster Algorithm - Concept

✓Continue previous example

- ASUS （華碩） belongs to “Computer Manufacturing” and “Smart Phone Manufacturing” with degree 0.75 and 0.25
- Acer （宏碁） belongs to “Computer Manufacturing” and “Smart Phone Manufacturing” with degree 0.8 and 0.2
- Compal （仁寶） belongs to “Computer Manufacturing” and “Smart Phone Manufacturing” with degree 0.9 and 0.1

Similarity(ASUS, Acer) ➔ Above 90%

Similarity(ASUS, Compal) ➔ Around 80%

Similarity(Acer, Compal) ➔ Around 85%

>0.9 →1

0.9~0.8 → 2

	ASUS	Acer	Compal
ASUS	-	**1**	**2**
Acer	1	-	**2**
Compal	2	2	-

PoCluster Algorithm

✓Input

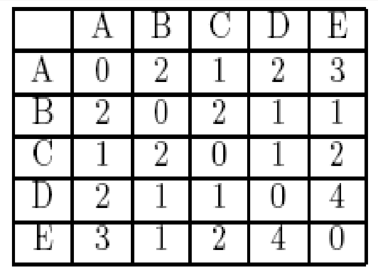
- *E*: An ordered list of edges (Can be generate from a similarity matrix)

✓Output

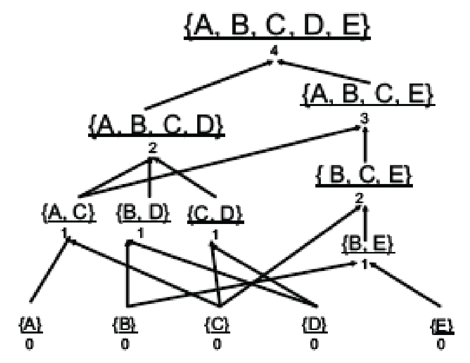
- A PoCluster

✓Three Steps

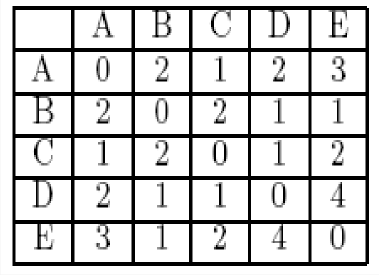
- Step 1: Select objects *O* with minimum value from E^t and $E^{t+1} = E^t - O$
- Step 2: Let $i = 0$ and *O* be a cluster C^{i+1} and $PoCluster = PoCluster \cup C^{i+1}$
- Step 3: If E^{t+1} is not empty, repeat Steps 2 & 3
- Step 4: Output *PoCluster*



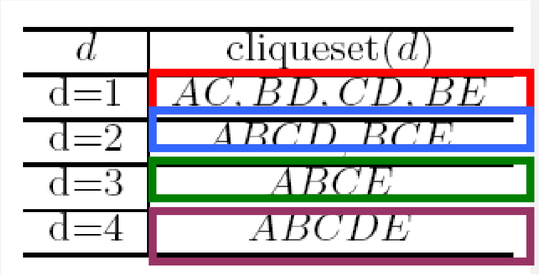
	A	B	C	D	E
A	0	2	1	2	3
B	2	0	2	1	1
C	1	2	0	1	2
D	2	1	1	0	4
E	3	1	2	4	0



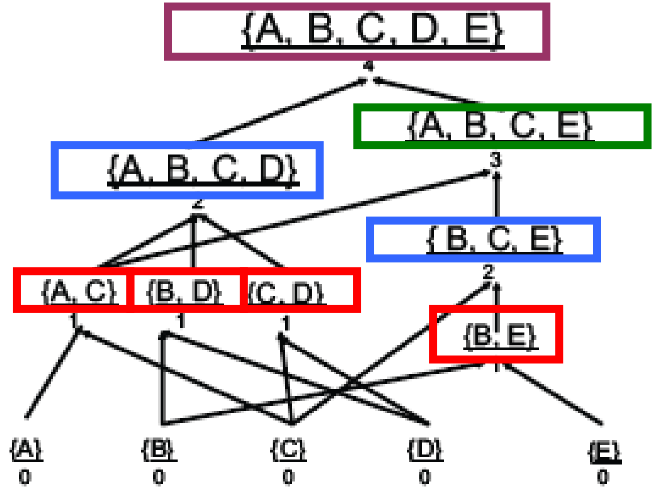
As a Result We Get



	A	B	C	D	E
A	0	2	1	2	3
B	2	0	2	1	1
C	1	2	0	1	2
D	2	1	1	0	4
E	3	1	2	4	0



d	cliqueset(d)
d=1	AC, BD, CD, BE
d=2	$ABCD, BCE$
d=3	$ABCE$
d=4	$ABCDE$

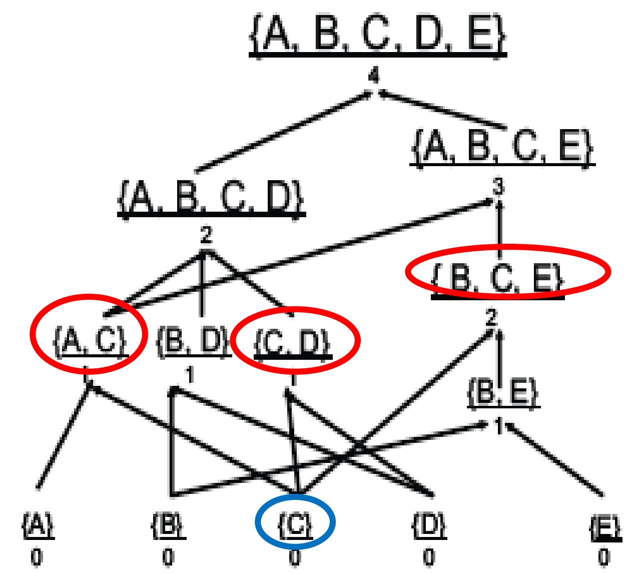


(A) An example PoCluster

A PoCluster

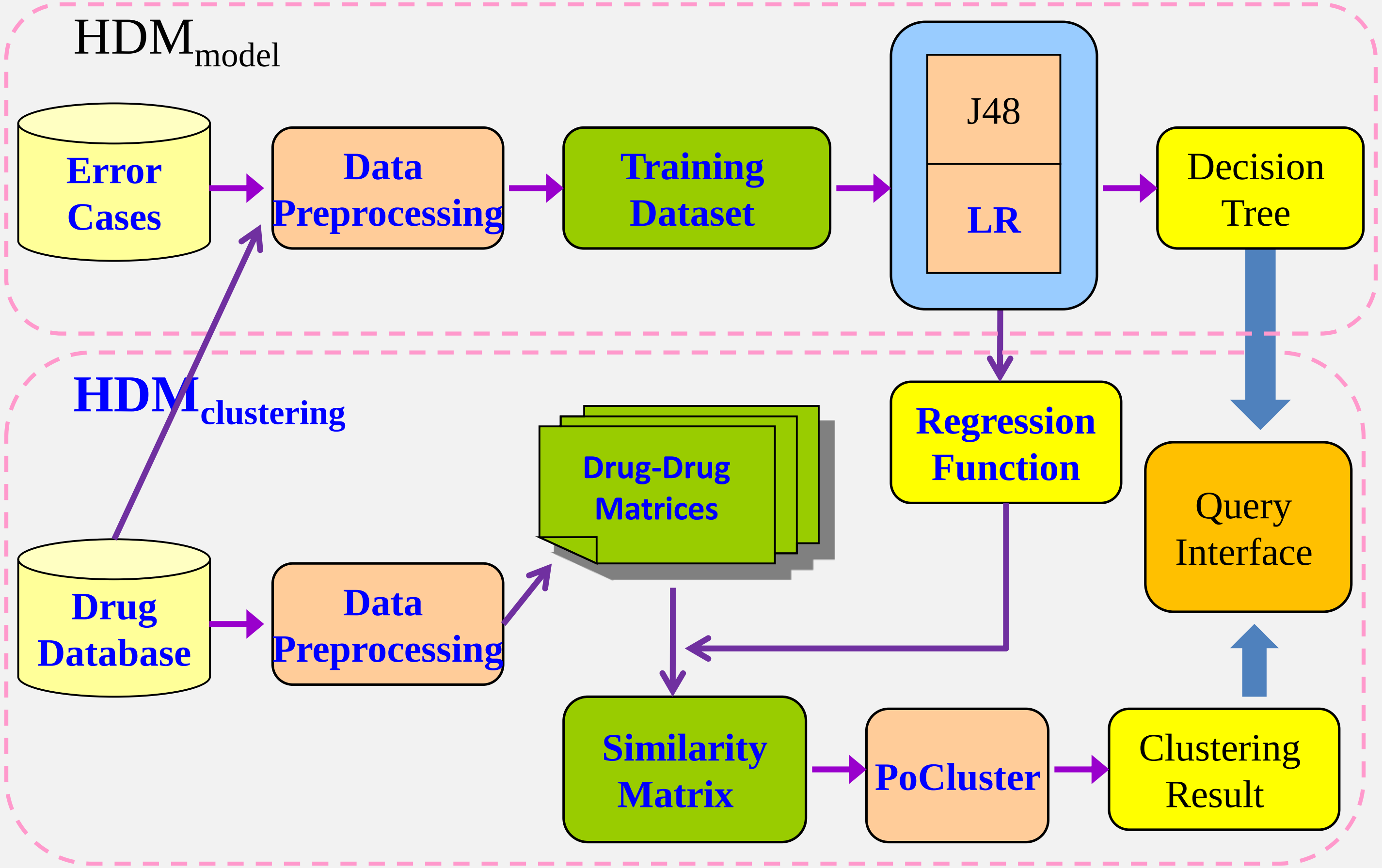
✓ One single element can belong to multiple clusters

✓ Reserve more information than other clustering approaches



Case Study – Drug Dispensing Error

- ✓ Find clusters for drug dispensing error prevention



Input Dataset - Drug Dataset

✓915 drugs with 10 attributes

	Drug_1	…	Drug_{141}	…	Drug_{915}
Generic Name	…	…	CARBAMAZEPINE C.R	…	…
Trade Name	…	…	TEGRETOL C.R. TAB	…	…
Pharmacology	…	…	6910	…	…
Location	…	…	架C	…	…
Dose Amount	…	…	200MG	…	…
Fifth Number	…	…	2	…	…
Dose Form Unit	…	…	TAB	…	…
Shape	…	…	凸長圓形	…	…
Color	…	…	橙色	…	…
Size	…	…	5.5*12	…	…

Input Dataset - Dispensing Error Cases

- ✓Drug dispensing error cases
- ✓e.g.

	Drug1 (Correct)	**Drug2 (Incorrect)**
ID	141	143
Generic Name	CARBAMAZEPINE C.R	CARBIDOPA/L-DOPA 25/100(SINEMET)
Trade Name	TEGRETOL C.R. TAB	SINEMET 25/100
Pharmacology	6910	6954
Location	架C	少4C
Dose Amount	200MG	125MG
Fifth Number	2	2
Dose Form Unit	TAB	TAB
Shape	凸長圓形	扁橢型
Color	橙色	黃色
Size	5.5*12	7*13

The First Problem Should be Handed

✓How to generate the similarity matrix for drugs

	Drug_1	…	Drug_{141}	…	Drug_{915}
Generic Name	…	…	CARBAMAZEPINE C.R	…	…
Trade Name	…	…	TEGRETOL C.R. TAB	…	…
Pharmacology	…	…	6910	…	…
Location	…	…	架C	…	…
Dose Amount	…	…	200MG	…	…
Fifth Number	…	…	2	…	…
Dose Form Unit	…	…	TAB	…	…
Shape	…	…	凸長圓形	…	…
Color	…	…	橙色	…	…
Size	…	…	5.5*12	…	…

?

Similarity Matrix

	D_1	D_2	…	D_m
D_1	–	…	…	…
D_2	…	–	…	…
…	…	…	–	…
D_m	…	…	…	–

Data Preprocessing

	Drug1	Drug2
ID	141	143
Generic Name	CARBAMAZEPINE C.R	CARBAMAZEPINE (TEGRETOL)
Trade Name	TEGRETOL C.R. TAB	SINEMET 25/100
Pharmacology	6910	6954
Location	架C	少4C
Dose Amount	200MG	125MG
Fifth Number	2	2
Dose Form Unit	TAB	TAB
Shape	凸長圓形	扁橢型
Color	橙色	黃色
Size	5.5*12	7*13

	Pair N
T1	0.204
T2	0
T4	0.150
NED1	0.719
NED2	0.882
NED4	0.804
Pharma	1
Loca	0
Dose	0
Form	2
Shape	0
Color	0
Size	0

T1, 2, and 4 > 0.116

NED1, 2, and 4 < 0.659

Data Preprocessing (Cont.)

	Pair N
T1	0.204
T2	0
T4	0.150
NED1	0.719
NED2	0.882
NED4	0.804
Pharma	1
Loca	0
Dose	0
Form	2
Shape	0
Color	0
Size	0

T1, 2, and 4 > 0.116

NED1, 2, and 4 < 0.659

	Pair N
T1	1
T2	0
T4	0
NED1	0
NED2	0
NED4	0
Pharma	1
Loca	0
Dose	0
Forms	2
Shape	0
Color	0
Size	0

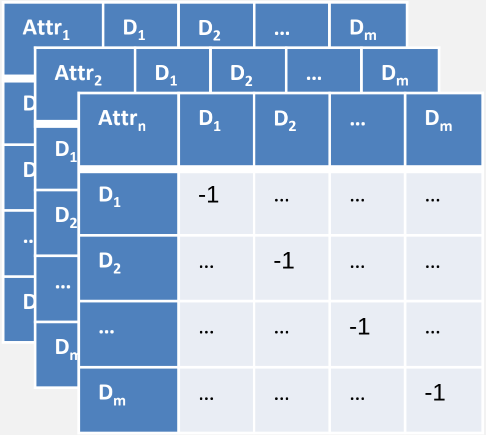
Data Preprocessing (Cont.)

Drug Database

	D1	…	D141	…	D915
Generic Name	…	…	CARBAMAZEPINE C.R	…	…
Trade Name	…	…	TEGRETOL C.R. TAB	…	…
Pharmacology	…	…	6910	…	…
Location	…	…	架C	…	…
Dose Amount	…	…	200MG	…	…
Fifth Number	…	…	2	…	…
Dose Form Unit	…	…	TAB	…	…
Shape	…	…	凸長圓形	…	…
Color	…	…	橙色	…	…
Size	…	…	5.5*12	…	…

T1, NED1, T2,
NED2, T4, NED4,
Loca, Dose,
Pharma, Form,
Size, Shape, Color

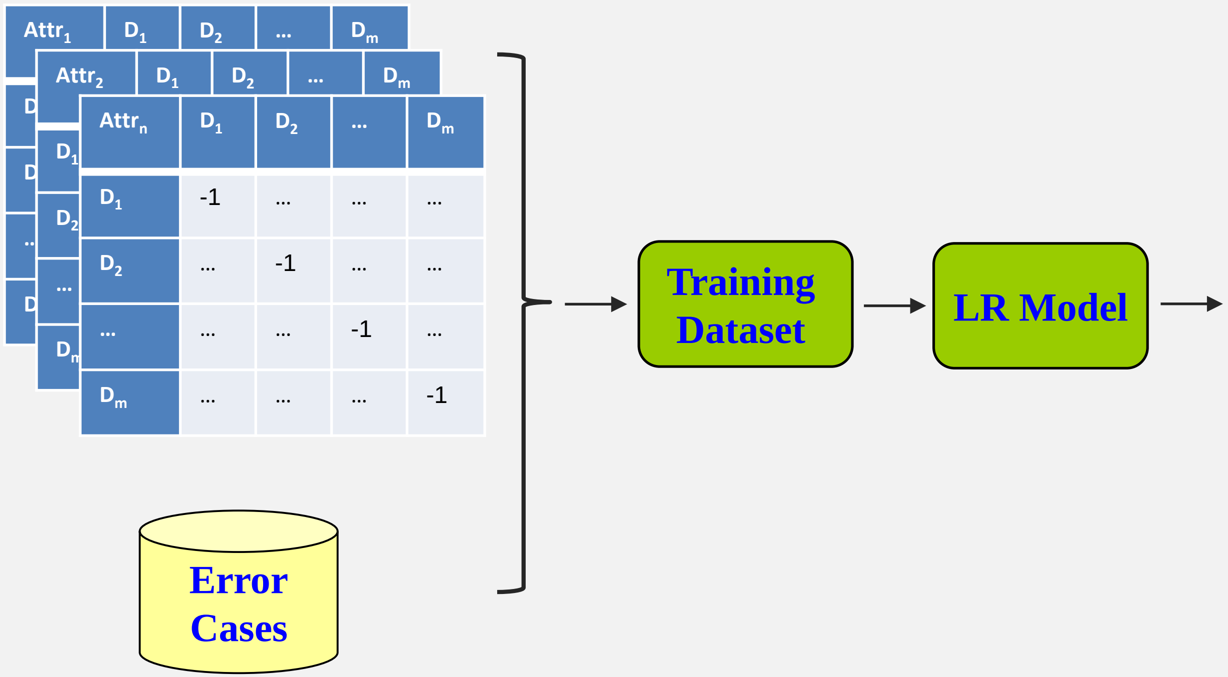
13 Drug-Drug Matrices



$Attr_n$	D_1	D_2	…	D_m
D_1	-1	…	…	…
D_2	…	-1	…	…
…	…	…	-1	…
D_m	…	…	…	-1

Logic Regression Model

13 Drug-Drug Matrices

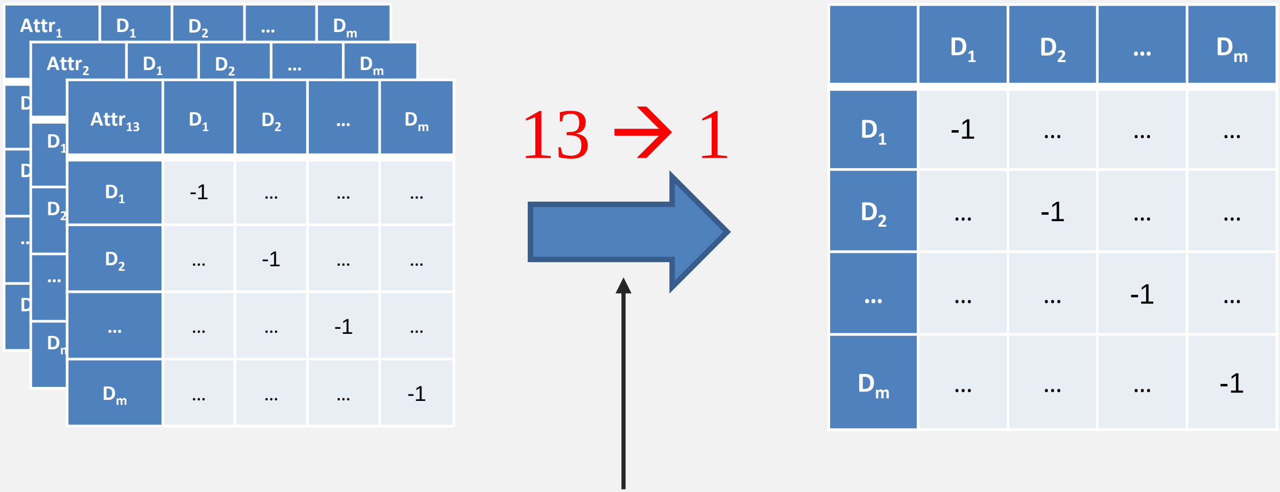


Attr$_n$	D$_1$	D$_2$	…	D$_m$
D$_1$	-1	…	…	…
D$_2$	…	-1	…	…
…	…	…	-1	…
D$_m$	…	…	…	-1

Regression Function

Value = 1.5382*(T1) + 0.8398*(NED1) –
0.2054*(T2) + 0.1425*(NED2) +
0.3912*(T4) + 0.4466*(NED4) +
0.9301*(Pharma) + 0.8858*(Loca) +
1.2563*(Dose) + 1.1272*(Form) –
0.3184*(Shape) + 0.0533*(Color) +
0.3505*(Size) - 2.4398

Use LR to Generate the Similarity Matrix



13 Drug-Drug Matrices

Attr_{13}	D_1	D_2	…	D_m
D_1	-1	…	…	…
D_2	…	-1	…	…
…	…	…	-1	…
D_m	…	…	…	-1

13 → 1

Similarity Matrix

	D_1	D_2	…	D_m
D_1	-1	…	…	…
D_2	…	-1	…	…
…	…	…	-1	…
D_m	…	…	…	-1

Regression Function

Value = 1.5382*(T1) + 0.8398*(NED1) - 0.2054*(T2) + 0.1425*(NED2) + 0.3912*(T4) + 0.4466*(NED4) + 0.9301*(Pharma) + 0.8858*(Loca) + 1.2563*(Dose) + 1.1272*(Form) - 0.3184*(Shape) + 0.0533*(Color) + 0.3505*(Size) - 2.4398

An Example

Drug141= CARBAMAZEPINE C.R
Drug143= CARBAMAZEPINE (TEGRETOL)

	T1	T2	T4	N1	N2	N4	Ph	Lo	Do	Fo	Sh	Co	Si
i	1	0	0	0	0	0	1	0	0	2	0	0	0

Value = 1.5382*(T1)+0.8398*(NED1)-0.2054*(T2)+0.1425*(NED2)+0.3912*(T4)
+0.4466*(NED4)+0.9301*(Pharma)+0.8858*(Loca)+1.2563*(Dose)
+1.1272*(Form)-0.3184*(Shape)+0.0533*(Color)+0.3505*(Size)-2.4398

Value = 2.2829

	D_1	…	D_{141}	D_m
D_1	-1	…	…	…
…	…	-1	…	…
D_{143}	…	…	2.28	…
D_m	…	…	…	-1

Discretization & Clustering Result

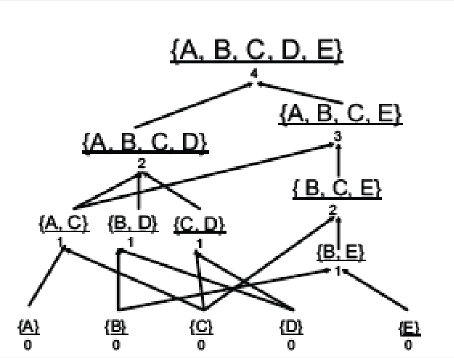
	1	2	3	…	915
1	-1	1.08	2.36	…	-1.51
2	1.08	-1	4.21	…	3.12
3	2.36	4.21	-1	…	6.18
…	…	…	…	-1	…
915	-1.51	3.12	6.18	…	-1

Level	Low	Up
1	-2	0
2	0	2
3	2	3
4	3	7

Final Similarity Matrix

	1	2	3	…	915
1	-1	2	3	…	1
2	2	-1	4	…	3
3	3	4	-1	…	4
…	…	…	…	-1	…
915	1	3	4	…	-1

Clustering Result

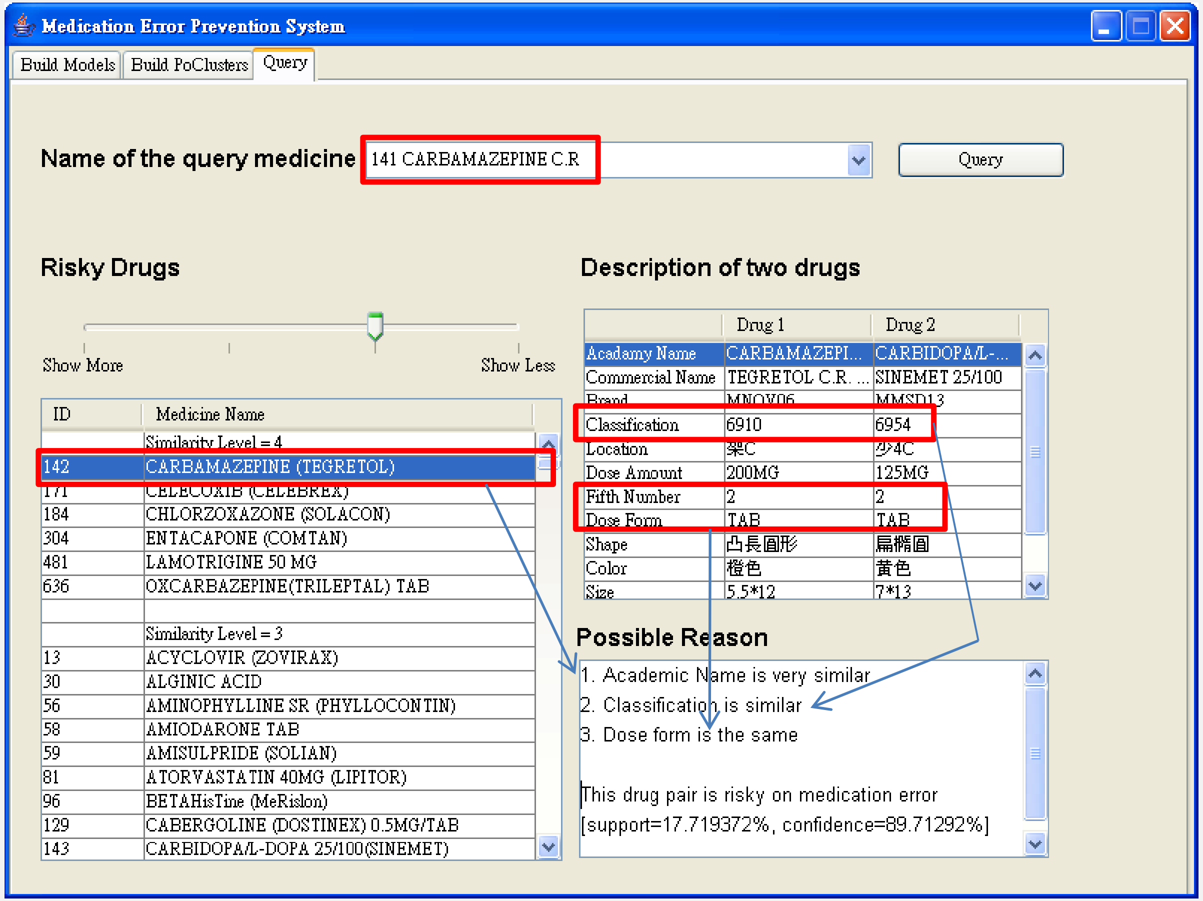


PoCluster Algorithm

Clustering Analysis

- ✓ Take “CARBAMAZEPINE C.R” as an example
- ✓ The following two slides show that
  - First, it can definitely find out the drug which is very similar to the queried one in Medicine Name
  - Second, we find that some drugs with low similarity with the queried one in Medicine Name, but similar in environmental attributes, such as Dose form, Classification and etc.

Clustering Analysis (Cont.)



Name of the query medicine: 141 CARBAMAZEPINE C.R

Risky Drugs

ID	Medicine Name
	Similarity Level = 4
142	CARBAMAZEPINE (TEGRETOL)
171	CELECOXIB (CELEBREX)
184	CHLORZOXAZONE (SOLACON)
304	ENTACAPONE (COMTAN)
481	LAMOTRIGINE 50 MG
636	OXCARBAZEPINE(TRILEPTAL) TAB
	Similarity Level = 3
13	ACYCLOVIR (ZOVIRAX)
30	ALGINIC ACID
56	AMINOPHYLLINE SR (PHYLLOCONTIN)
58	AMIODARONE TAB
59	AMISULPRIDE (SOLIAN)
81	ATORVASTATIN 40MG (LIPITOR)
96	BETAHisTine (MeRislon)
129	CABERGOLINE (DOSTINEX) 0.5MG/TAB
143	CARBIDOPA/L-DOPA 25/100(SINEMET)

Description of two drugs

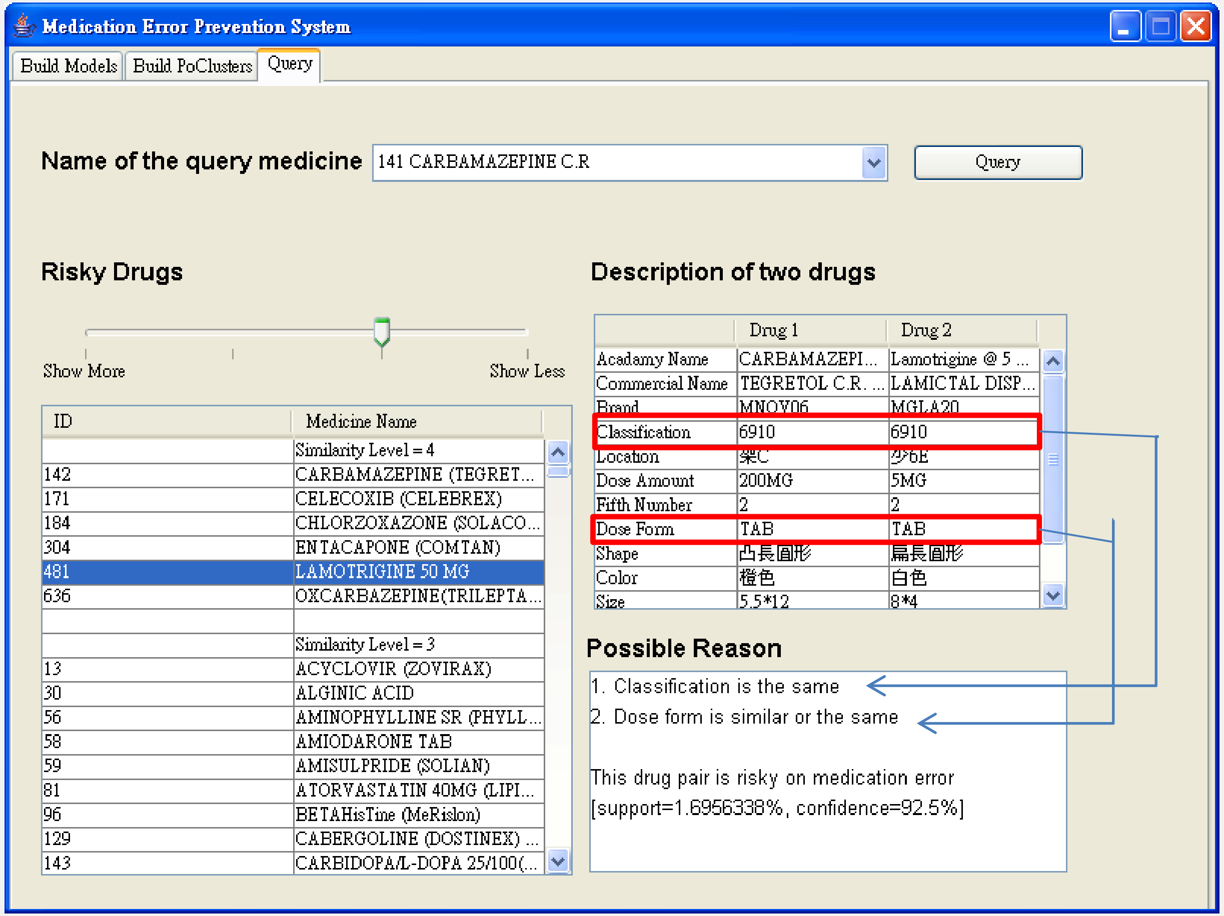
	Drug 1	Drug 2
Acadamy Name	CARBAMAZEPI...	CARBIDOPA/L-...
Commercial Name	TEGRETOL C.R. ...	SINEMET 25/100
Brand	MNOV06	MMSD13
Classification	6910	6954
Location	架C	少4C
Dose Amount	200MG	125MG
Fifth Number	2	2
Dose Form	TAB	TAB
Shape	凸長圓形	扁橢圓
Color	橙色	黃色
Size	5.5*12	7*13

Possible Reason

1. Academic Name is very similar
2. Classification is similar
3. Dose form is the same

This drug pair is risky on medication error [support=17.719372%, confidence=89.71292%]

Clustering Analysis (Cont.)



Name of the query medicine 141 CARBAMAZEPINE C.R

Risky Drugs

ID	Medicine Name
	Similarity Level = 4
142	CARBAMAZEPINE (TEGRET...
171	CELECOXIB (CELEBREX)
184	CHLORZOXAZONE (SOLACO...
304	ENTACAPONE (COMTAN)
481	LAMOTRIGINE 50 MG
636	OXCARBAZEPINE(TRILEPTA...
	Similarity Level = 3
13	ACYCLOVIR (ZOVIRAX)
30	ALGINIC ACID
56	AMINOPHYLLINE SR (PHYLL...
58	AMIODARONE TAB
59	AMISULPRIDE (SOLIAN)
81	ATORVASTATIN 40MG (LIPI...
96	BETAHisTine (MeRislon)
129	CABERGOLINE (DOSTINEX) ...
143	CARBIDOPA/L-DOPA 25/100(...

Description of two drugs

	Drug 1	Drug 2
Acadamy Name	CARBAMAZEPI...	Lamotrigine @ 5 ...
Commercial Name	TEGRETOL C.R. ...	LAMICTAL DISP...
Brand	MNOV06	MGLA20
Classification	6910	6910
Location	架C	少6E
Dose Amount	200MG	5MG
Fifth Number	2	2
Dose Form	TAB	TAB
Shape	凸長圓形	扁長圓形
Color	橙色	白色
Size	5.5*12	8*4

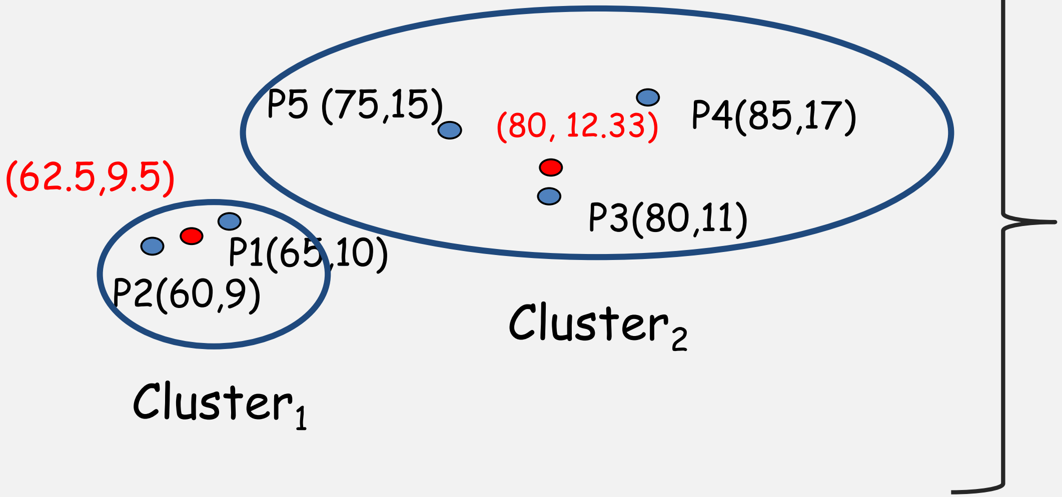
Possible Reason

1. Classification is the same
2. Dose form is similar or the same

This drug pair is risky on medication error
[support=1.6956338%, confidence=92.5%]

Solution II - Soft (Fuzzy) Clustering

✓*k*-means clustering



Membership Matrix M

	Cluster_1	Cluster_2
P1	**1**	0
P2	**1**	0
P3	0	**1**
P4	0	**1**
P5	0	**1**

Soft (Fuzzy) Clustering (Cont.)

✓Fuzzy k-means (Dunn, 1973)

- Data can belong to two or more clusters

	Cluster$_1$	Cluster$_2$
P1	**1**	0
P2	**1**	0
P3	0	**1**
P4	0	**1**
P5	0	**1**

0 or 1

→

	Cluster$_1$	Cluster$_2$
P1	0.9	0.1
P2	0.85	0.15
P3	0.15	0.85
P4	0.10	0.90
P5	0.05	0.95

[0, 1]

An Example

- ✓Assume
  - 5 basketball players: P1, P2, P3, P4, P5
  - Two attributes: Speed and Weight

	P1	P2	P3	P4	P5
Speed	10sec(100m)	9sec(100m)	11sec(100m)	20sec(100m)	13sec(100m)
Weight	65kg	60kg	80kg	99kg	70kg

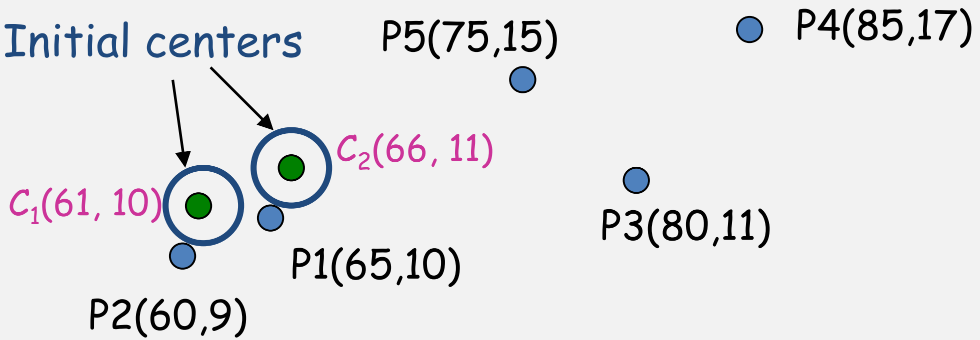
- ✓Illustrate how k-means clustering algorithm work
  - Number of cluster *k* = 2

Step 1

✓Initialize *k* centers

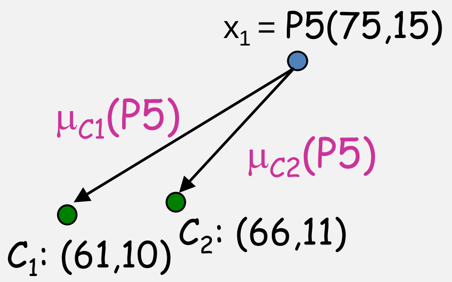
- Randomly generate
- *e.g.* $C_1(61, 10)$ and $C_2(66, 11)$

	P1	P2	P3	P4	P5
Speed	10sec(100m)	9sec(100m)	11sec(100m)	17sec(100m)	15sec(100m)
Weight	65kg	60kg	80kg	85kg	75kg



Step 2

- ✓ Calculate fuzzy value of each object to every group
  - e.g. $\mu_{C1}(P5)$



$$\mu_{C_1}(x_1) = \frac{1}{\sum_{j=1}^{2}\left(\frac{\| x_1 - c_1 \|^2}{\| x_1 - c_j \|^2}\right)} = \frac{1}{\left(\frac{\| x_1 - c_1 \|^2}{\| x_1 - c_1 \|^2}\right) + \left(\frac{\| x_1 - c_1 \|^2}{\| x_1 - c_2 \|^2}\right)}$$

$$= \frac{1}{\frac{221}{221} + \frac{221}{97}} = \frac{1}{1 + 2.783} = 0.305$$

$$\mu_{C_2}(x_1) = \frac{1}{\sum_{j=1}^{2}\left(\frac{\| x_1 - c_2 \|^2}{\| x_1 - c_j \|^2}\right)} = \frac{1}{\frac{97}{221} + \frac{97}{97}}$$

$$= \frac{1}{0.439 + 1} = 0.695$$

$||x_1 - c_1||^2 = (75-61)^2+(15-10)^2 = 221$

$||x_1 - c_2||^2 = (75-66)^2+(15-11)^2 = 97$

	Cluster_1	Cluster_2
P5	0.305	0.695

After Step 2

- ✓ Form the $U^{(0)}$ matrix

	Cluster_1	Cluster_2
P5	0.305	0.695
P2	0.952	0.048
P1	0.111	0.889
P3	0.351	0.649
P4	0.388	0.612

Step 3

✓ Calculate new centers

	Cluster_1	Cluster_2
P5	0.305	0.695
P2	0.952	0.048
P1	0.111	0.889
P3	0.351	0.649
P4	0.388	0.612

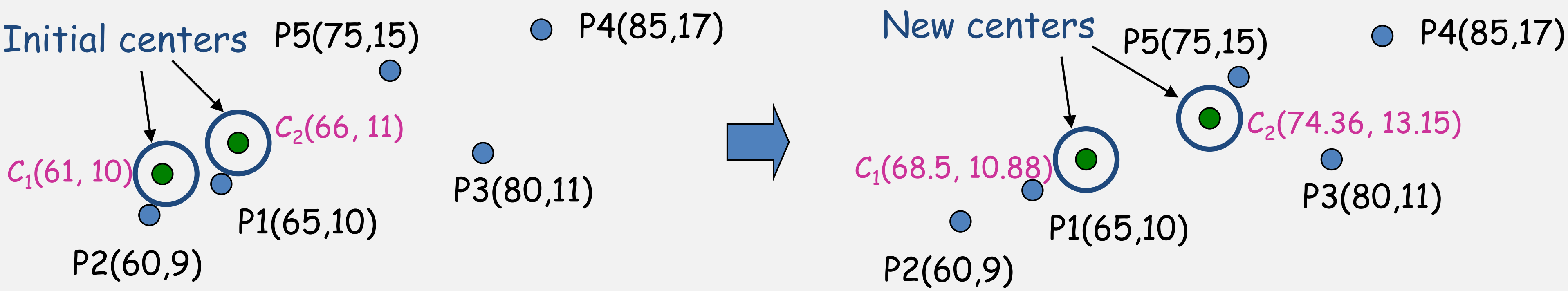
$$C_1 = \frac{\sum_{i=1}^{5} (\mu_{C_1}(x_i))^2 \times x_i}{\sum_{i=1}^{5} (\mu_{C_1}(x_i))^2}$$

$$C_1 = \frac{0.305^2(60,9) + 0.952^2(65,10) + 0.111^2(75,15) + 0.351^2(80,11) + 0.388^2(85,17)}{0.305^2 + 0.952^2 + 0.111^2 + 0.351^2 + 0.388^2}$$
$$= (\frac{87.96}{1.284}, \frac{13.98}{1.284}) = (68.50, 10.88)$$

$$C_2 = \frac{0.695^2(60,9) + 0.048^2(65,10) + 0.889^2(75,15) + 0.649^2(80,11) + 0.612^2(85,17)}{0.695^2 + 0.048^2 + 0.889^2 + 0.649^2 + 0.612^2}$$
$$= (\frac{153.93}{2.07}, \frac{27.22}{2.07}) = (74.36, 13.15)$$

After Step 3

- ✓ New Centers



Steps 4 & 5

- ✓ Step 4: Calculate $U^{(1)}$ matrix
  - The same as Step 2
- ✓ Step 5: Reach the stop criterion or not
  - If $||U^{(r+1)} - U^{(r)}|| < \varepsilon$, Then STOP.
  - Otherwise repeat Steps 3 and 4

Fuzzy *k*-means Algorithm

- ✓ Five Steps
  - 1. Initialize *k* centers
  - 2. Calculate $U^{(r)}$ ($=[u_{ij}]$) matrix
  - 3. Calculate the new centers
  - 4. Calculate $U^{(r+1)}$ ($=[u_{ij}]$) matrix
  - 5. If $||U^{(r+1)} - U^{(r)}|| < \varepsilon$, Then STOP; Otherwise repeat Steps 3 and 4

Discussions

What are the key points of the fuzzy c-means algorithm?

The Formulas

$$\mu_{xk} = \frac{1}{\sum_{j\in K}\left(\frac{d^2(x,k)}{d^2(x,j)}\right)^{\frac{1}{m-1}}}$$

Calculate
Membership Values

$$k = \frac{\sum_{x\in X}\mu_{xk}^{m}x}{\sum_{x\in X}\mu_{xk}^{m}}$$

Calculate New Center

In The Example ..

✓ The parameter *m* was set at 2

$$\mu_{xk} = \frac{1}{\sum_{j \in K} \left(\frac{d^2(x,k)}{d^2(x,j)} \right)^{\frac{1}{m-1}}}$$

$$\Downarrow$$

$$\mu_{xk} = \frac{1}{\sum_{j \in K} \left(\frac{d^2(x,k)}{d^2(x,j)} \right)}$$

$$k = \frac{\sum_{x \in X} \mu_{xk}^{m} x}{\sum_{x \in X} \mu_{xk}^{m}}$$

$$\Downarrow$$

$$k = \frac{\sum_{x \in X} \mu_{xk}^{2} x}{\sum_{x \in X} \mu_{xk}^{2}}$$

When Variable *m* with Large Value

✓ *m* = 10

$$\mu_{C_1}(x_1) = \frac{1}{\sum_{j=1}^{2}\left(\frac{\| x_1 - c_1 \|^2}{\| x_1 - c_j \|^2}\right)} = \frac{1}{\left(\frac{\| x_1 - c_1 \|^2}{\| x_1 - c_1 \|^2}\right) + \left(\frac{\| x_1 - c_1 \|^2}{\| x_1 - c_2 \|^2}\right)}$$

$$= \frac{1}{\frac{221}{221} + \frac{221}{97}} = \frac{1}{1 + 2.783} = 0.305$$

$$\Downarrow$$

$$\mu_{C_1}(x_1) = \frac{1}{\sum_{j=1}^{2}\left(\frac{\| x_1 - c_1 \|^2}{\| x_1 - c_j \|^2}\right)^{\frac{1}{10-1}}} = \frac{1}{\left(\frac{\| x_1 - c_1 \|^2}{\| x_1 - c_1 \|^2}\right)^{\frac{1}{10-1}} + \left(\frac{\| x_1 - c_1 \|^2}{\| x_1 - c_2 \|^2}\right)^{\frac{1}{10-1}}}$$

$$= \frac{1}{(\frac{221}{221})^{\frac{1}{10-1}} + (\frac{221}{97})^{\frac{1}{10-1}}} = \frac{1}{1 + (2.783)^{\frac{1}{10-1}}} = > 0.305$$

$$\mu_{xk} = \frac{1}{\sum_{j \in K} (\frac{d^2(x,k)}{d^2(x,j)})^{\frac{1}{m-1}}}$$

For u_{xk},
The parameter *m* is used to control **the sensitive of the distance between data point and centers**

When Variable *m* with Large Value (Cont.)

✓ $m = 10$

$$C_1 = \frac{0.305^2(60,9)+0.952^2(65,10)+0.111^2(75,15)+0.351^2(80,11)+0.388^2(85,17)}{0.305^2+0.952^2+0.111^2+0.351^2+0.388^2}$$

$$=(\frac{87.96}{1.284},\frac{13.98}{1.284})=(68.50,10.88)$$

$$\Downarrow$$

$$C_1 = \frac{0.305^{10}(60,9)+0.952^{10}(65,10)+0.111^{10}(75,15)+0.351^{10}(80,11)+0.388^{10}(85,17)}{0.305^{10}+0.952^{10}+0.111^{10}+0.351^{10}+0.388^{10}}$$

$$=(<68.50,<10.88)$$

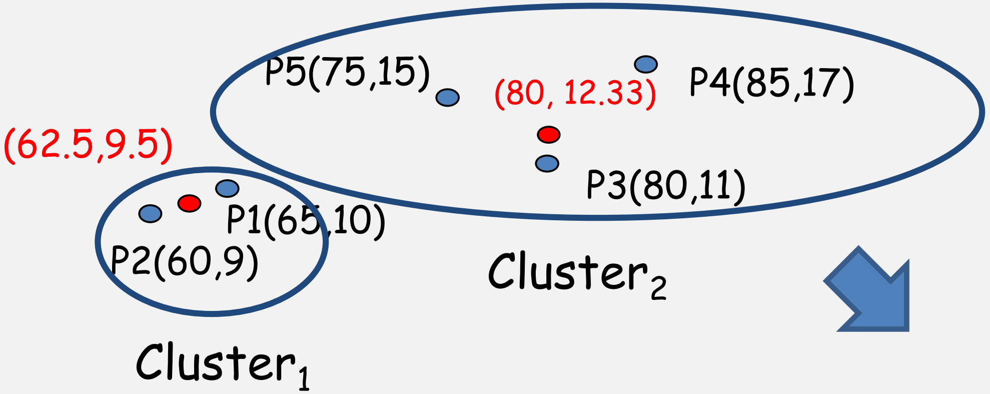
$$k = \frac{\sum_{x\in X}\mu_{xk}^m x}{\sum_{x\in X}\mu_{xk}^m}$$

For k,

The parameter *m* is used to control **the changing sensitive of the centers**

Discussion 2

- ✓Clustering results with *k* = 2



Do we have any approach to get clusters without the *k* value?

Cluster Affinity Search Technique

✓Input

- S： a symmetric $n \times n$ Similarity Matrix， $S(i, j) \in [0, 1]$
- t： Affinity Threshold ($0 < t < 1$)

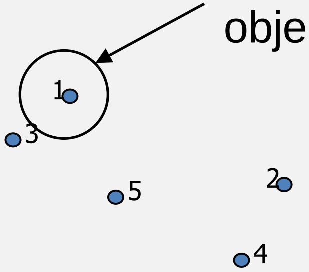
✓Method

- 1. Choose a seed for generating a new cluster
- 2. ADD: add qualified items to the cluster
- 3. REMOVE: remove unqualified items from the stable cluster
- 4. Repeat Steps 1-3 till no more clusters can be generated

Step 1: Select A Seed

✓ Assume we have five data objects

- U = {1, 2, 3, 4, 5}
- Affinity Threshold $t = 0.5$



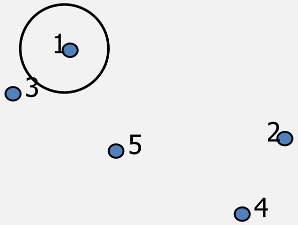
Assume we select data object 1 as a seed

Similarity matrix

	1	2	3	4	5
1	1	0.2	0.8	0.1	0.5
2		1	0.4	0.7	0.4
3			1	0.3	0.6
4				1	0.5
5					1

Step 2: ADD Phase

Before



a(1) = 1

a(2) = 0.2

a(3) = 0.8 **max** $\geq t^*|C_{open}|$

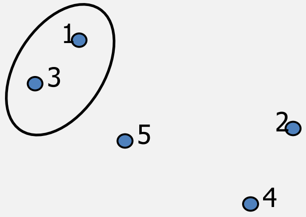
$= 0.5^*1$

$= 0.5$

a(4) = 0.1

a(5) = 0.5

After

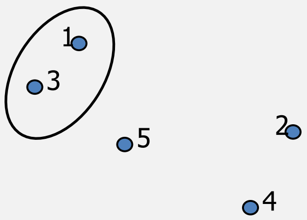


Similarity matrix

	1	**2**	**3**	**4**	**5**
1	**1**	**0.2**	**0.8**	**0.1**	**0.5**
2		**1**	**0.4**	**0.7**	**0.4**
3			**1**	**0.3**	**0.6**
4				**1**	**0.5**
5					**1**

Step 2: ADD Phase (Cont.)

Before



$\mathbf{a(2) = 0.2 + 0.4}$

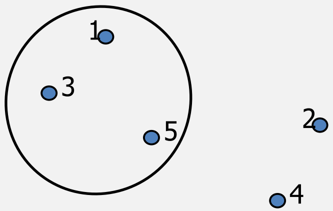
$\mathbf{a(4) = 0.1 + 0.3}$

$\mathbf{a(5) = 0.5 + 0.6}$ $\mathbf{max} \geq t|C_{open}|$

$= 0.5*2$

$=1.0$

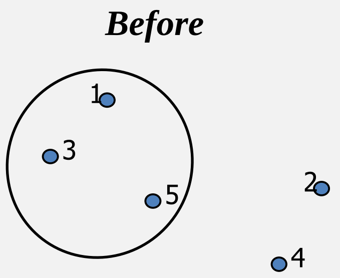
After



Similarity matrix

	1	2	3	4	5
1	1	0.2	0.8	0.1	0.5
2		1	0.4	0.7	0.4
3			1	0.3	0.6
4				1	0.5
5					1

Step 2: ADD Phase (Cont.)



$a(2) = 0.2 + 0.4 + 0.4$

$a(4) = 0.1 + 0.3 + 0.5$

max $< t|C_{open}|$
$= 0.5*3$
$=1.5$

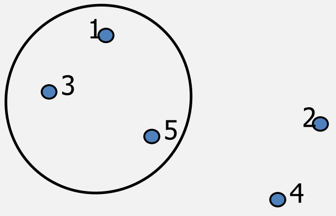
Stop

Similarity matrix

	1	2	3	4	5
1	1	0.2	0.8	0.1	0.5
2		1	0.4	0.7	0.4
3			1	0.3	0.6
4				1	0.5
5					1

Step 3: Remove Phase

Before



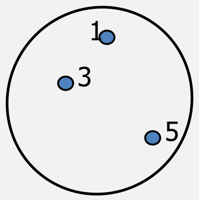
a(1) = 1 + 0.8 + 0.5 = 2.3

a(3) = 0.8 + 1 + 0.6 = 2.4

a(5) = 0.5 + 0.6 + 1 = 2.1

$$\min > t|C_{open}| = 0.5*3 = 1.5$$

Cluster$_1$

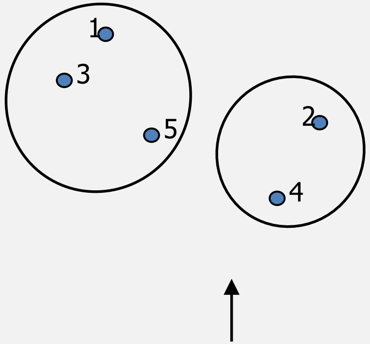


Similarity matrix

	1	2	3	4	5
1	1	0.2	0.8	0.1	0.5
2		1	0.4	0.7	0.4
3			1	0.3	0.6
4				1	0.5
5					1

Step 4: Repeat Steps 1-3

✓ The Final Clustering Results



Two Clusters

Repeat Add and Remove

Discussion

✓ Input

- S： a symmetric $n \times n$ Similarity Matrix，$S(i, j) \in [0, 1]$
- t： Affinity Threshold ($0 < t < 1$)

✓ Method

- 1. Choose a seed for generating a new cluster
- 2. ADD: add qualified items to the cluster
- 3. REMOVE: remove unqualified items from the stable cluster
- 4. Repeat Steps 1-3 till no more clusters can be generated

What are the key points of the CAST algorithm?

Conclusions

- ✓ Advanced Clustering Algorithms
  - ● Objects can belong multiple Clusters
    - ✓ PoCluster algorithm
    - ✓ Case Study: Drug Dispensing Error
    - ✓ Fuzzy c-means clustering algorithm
  - ● Clustering without number of groups
    - ✓ CAST