Spectral Clustering
運用圖論 (Graph Theory) 進行分群

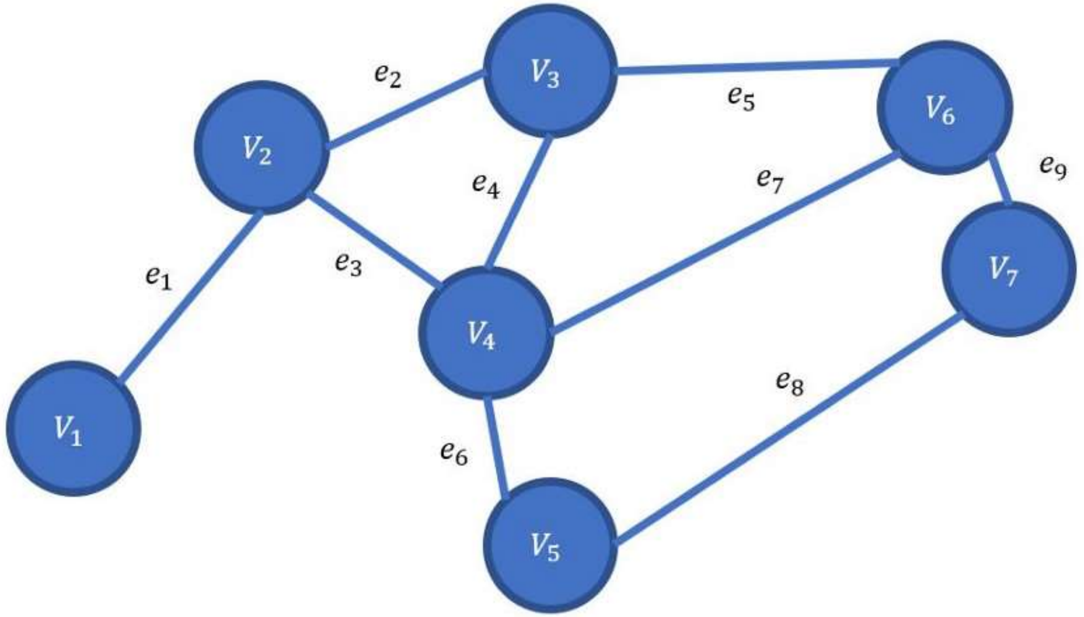
常博愛 408410086

圖（graph）

G（V, E）。

鄰接矩陣（Affinity matrix）

$$A_{ij} = \begin{cases} 1 & \text{if } \{v_i, v_j\} \in E \\ 0 & \text{else} \end{cases}$$



相似圖 (Similarity Graph) 的建立

一.k 鄰近法:（KNN）

$$A_{ij} = \begin{cases} 0 & v_i \notin knn(v_j) \ \& \ v_j \notin knn(v_i) \\ e^{-\frac{\|v_i - v_j\|^2}{2\sigma^2}} & \text{else} \end{cases}$$

二.全連接法：將所有點連接。

$$A_{ij} = \begin{cases} e^{-\frac{\|v_i - v_j\|^2}{2\sigma^2}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

高斯核相似函數 (Gaussian Kernerl Simlarity)

degree of vertex: $d_i = \sum_i^n s_{i,j}$

degree of vertex matrix: $D = diag(d_1, d_2, d_3, \ldots, d_n)$

Size:

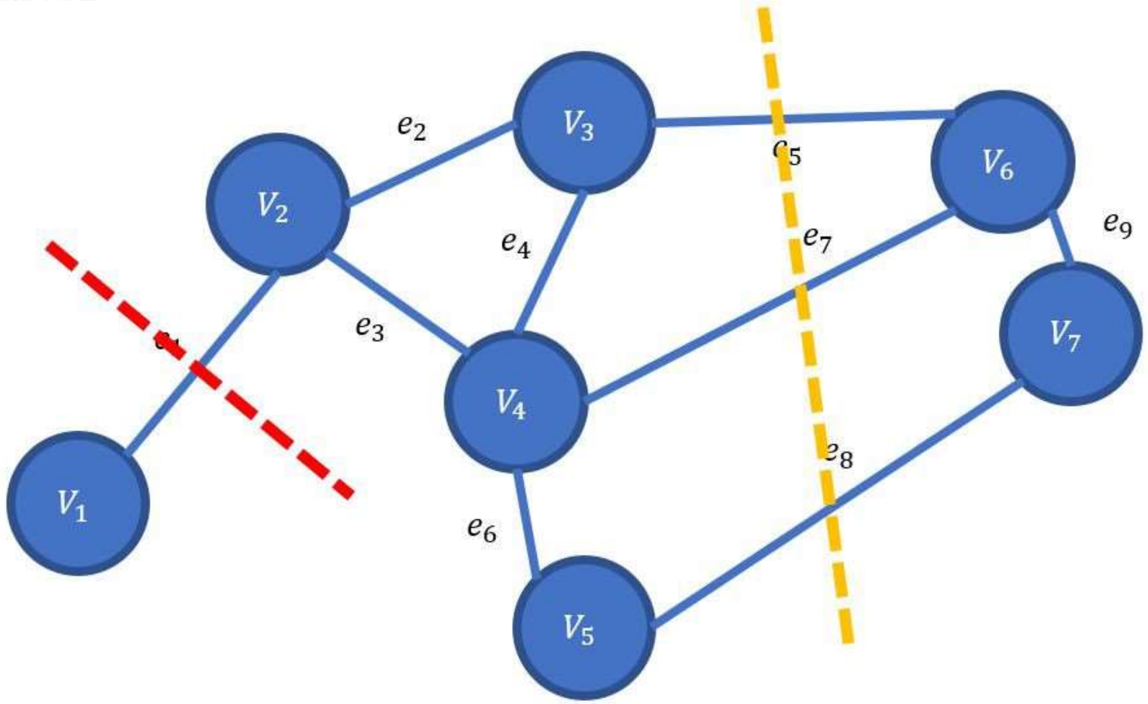
|A| := the number of vertices in A

Vol(A)

切圖（Cut）

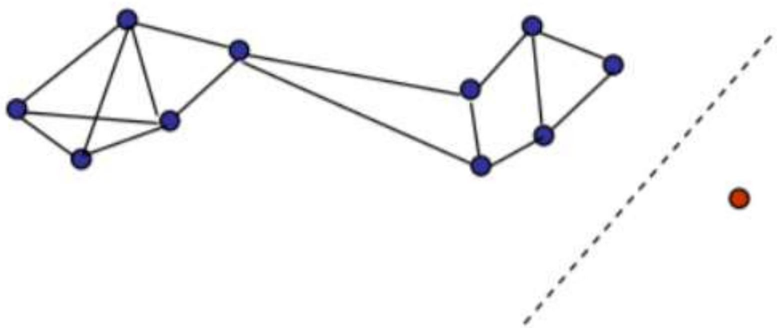
$$Cut(G_1, G_2, \ldots, G_k) = \frac{1}{2}\sum_{i=1}^{k} W(G_i, G_i^C)$$

子圖間的權重可表示為:

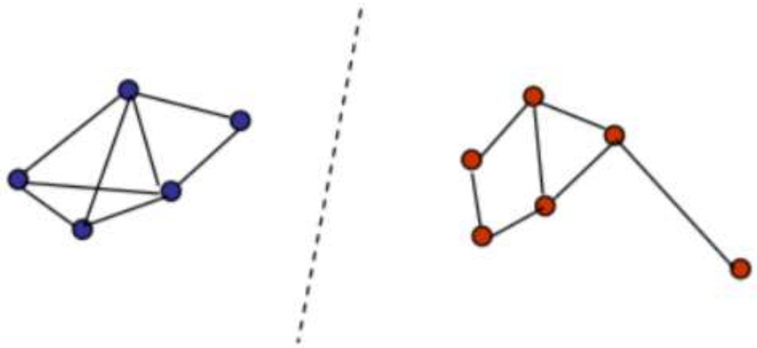


$$W(X, Y) = \sum_{i \in X,\, j \in Y} A_{ij}$$

$$\min Cut(G_1, G_2, \ldots, G_k)$$



What we get



What we want

Solutions

$|A| :=$ the number of vertices in A

$$vol(A) := \sum_{i \in A} d_i$$

- RatioCut(Hagen and Kahng, 1992)

$$RatioCut(A_1, \ldots, A_k) := \frac{1}{2} \sum_{i=1}^{k} \frac{W(A_i, \overline{A_i})}{|A_i|} = \sum_{i=1}^{k} \frac{cut(A_i, \overline{A_i})}{|A_i|}$$

- Ncut(Shi and Malik, 2000)

$$Ncut(A_1, \ldots, A_k) := \frac{1}{2} \sum_{i=1}^{k} \frac{W(A_i, \overline{A_i})}{vol(A_i)} = \sum_{i=1}^{k} \frac{cut(A_i, \overline{A_i})}{vol(A_i)}$$

Problem!!!

- NP hard

$$RatioCut(A_1, \ldots, A_k) := \frac{1}{2}\sum_{i=1}^{k}\frac{W(A_i, \overline{A_i})}{|A_i|} = \sum_{i=1}^{k}\frac{cut(A_i, \overline{A_i})}{|A_i|}$$

$$Ncut(A_1, \ldots, A_k) := \frac{1}{2}\sum_{i=1}^{k}\frac{W(A_i, \overline{A_i})}{vol(A_i)} = \sum_{i=1}^{k}\frac{cut(A_i, \overline{A_i})}{vol(A_i)}$$

Solution!!!

- Approximation

Approx imation → Spectral Clustering

Spectral Clustering → Ncut $\xrightarrow{\text{relaxing}}$ Normalized Spectral Clustering

Spectral Clustering → RatioCut $\xrightarrow{\text{relaxing}}$ Unnormalized Spectral Clustering

- Approximation RatioCut for k=2

Our goal is to solve the optimization problem:

$$\min_{A \subset V} RatioCut(A, \bar{A})$$

Rewrite the problem in a more convenient form:

Given a subset $A \subset V$, we define the vector $f = (f_1, \dots, f_n)' \in \mathbb{R}^n$ with entries

$$f_i = \begin{cases} \sqrt{|\bar{A}|/|A|}, & \text{if } v_i \in A \\ -\sqrt{|\bar{A}|/|A|}, & \text{if } v_i \in \bar{A} \end{cases}$$

$$
\begin{aligned}
RatioCut(G_i, G_i^c) &= \frac{1}{|G_i| + |G_i^c|}\left[Cut(G_i, G_i^c)\frac{|G_i| + |G_i^c|}{|G_i|} + Cut(G_i^c, G_i)\frac{|G_i| + |G_i^c|}{|G_i^c|}\right] \\
&= \frac{1}{|G_i| + |G_i^c|} Cut(G_i, G_i^c)\left(\frac{|G_i^c|}{|G_i|} + \frac{|G_i|}{|G_i^c|} + 2\right) \\
&= \frac{1}{2(|G_i| + |G_i^c|)}\left[\sum_{m \in G_i, n \in G_i^c} A_{mn}\left(\sqrt{\frac{|G_i^c|}{|G_i|}} + \sqrt{\frac{|G_i|}{|G_i^c|}}\right)^2\right. \\
&\left. + \sum_{m \in G_i^c, n \in G_i} A_{mn}\left(-\sqrt{\frac{|G_i^c|}{|G_i|}} - \sqrt{\frac{|G_i|}{|G_i^c|}}\right)^2\right]
\end{aligned}
$$

$$
f_i = \begin{cases} \sqrt{\frac{|G^C|}{|G|}} & \text{if } v_i \in G \\ -\sqrt{\frac{|G|}{|G^C|}} & \text{if } v_i \in G^C \end{cases}
$$

$$
\begin{aligned}
RatioCut(G_i, G_i^C) &= \frac{1}{2(|G_i| + |G_i^C|)} \sum_{m=1}^{N} \sum_{n=1}^{N} A_{mn}(f_m - f_n)^2 \\
&= \frac{1}{2(|G_i| + |G_i^C|)} \left(\sum_{m=1}^{N} d_m f_m^2 - \sum_{m=1}^{N} \sum_{n=1}^{N} f_m f_n A_{mn} + \sum_{n=1}^{N} dn f_n^2 \right) \\
&= \frac{1}{|G_i| + |G_i^C|} \left(\sum_{m=1}^{N} d_m f_m^2 - \sum_{m=1}^{N} \sum_{n=1}^{N} f_m f_n A_{mn} \right) \\
&= \frac{1}{|G_i| + |G_i^C|} (f' D f - f' A f) \\
&= \frac{1}{|G_i| + |G_i^C|} (f' L f)
\end{aligned}
$$

$$L = D - A$$

$$Lf = \lambda f$$

$$f'Lf = \lambda f'f = \lambda N$$

常用演算法:

1.Unnormalized spectral clustering

2.Normalized spectral clustering according to Shi and Malik

3.Normalized spectral clustering according to Ng, Jordan, and Weiss

对称正规化调和矩阵 [编辑]

$$L^{\text{sym}} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$$

$$L^{\text{sym}}_{i,j} := \begin{cases} 1 & \text{if } i = j \text{ and } \deg(v_i) \neq 0 \\ -\frac{1}{\sqrt{\deg(v_i)\deg(v_j)}} & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise.} \end{cases}$$

注意[4]

$$\lambda = \frac{\langle g, L^{\text{sym}} g\rangle}{\langle g, g\rangle} = \frac{\left\langle g, D^{-\frac{1}{2}} L D^{-\frac{1}{2}} g\right\rangle}{\langle g, g\rangle} = \frac{\langle f, Lf\rangle}{\left\langle D^{\frac{1}{2}} f, D^{\frac{1}{2}} f\right\rangle} = \frac{\sum_{u\sim v}(f(u)-f(v))^2}{\sum_v f(v)^2 d_v} \geq 0,$$

随机漫步 [编辑]

$$L^{\text{rw}} := D^{-1} L = I - D^{-1} A$$

$$L^{\text{rw}}_{i,j} := \begin{cases} 1 & \text{if } i = j \text{ and } \deg(v_i) \neq 0 \\ -\frac{1}{\deg(v_i)} & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise.} \end{cases}$$

Unnormalized spectral clustering

Unnormalized spectral clustering

Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct.

- Construct a similarity graph by one of the ways described in Section 2. Let W be its weighted adjacency matrix.
- Compute the unnormalized Laplacian L.
- **Compute the first k eigenvectors $u_1, \ldots, u_k$ of L.**
- Let $U \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $u_1, \ldots, u_k$ as columns.
- For $i = 1, \ldots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of U.
- Cluster the points $(y_i)_{i=1,\ldots,n}$ in $\mathbb{R}^k$ with the k-means algorithm into clusters $C_1, \ldots, C_k$.

Output: Clusters $A_1, \ldots, A_k$ with $A_i = \{j \mid y_j \in C_i\}$.

$$L = D - S$$

Normalized spectral clustering according to Shi and Malik （L_rw）

Normalized spectral clustering according to Shi and Malik (2000)

Input: Similarity matrix $S \in \mathbb{R}^{n\times n}$, number k of clusters to construct.

- Construct a similarity graph by one of the ways described in Section 2. Let W be its weighted adjacency matrix.
- Compute the unnormalized Laplacian L.
- **Compute the first k generalized eigenvectors $u_1, \dots, u_k$ of the generalized eigenproblem $Lu = \lambda Du$.**
- Let $U \in \mathbb{R}^{n\times k}$ be the matrix containing the vectors $u_1, \dots, u_k$ as columns.
- For $i = 1, \dots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of U.
- Cluster the points $(y_i)_{i=1,\dots,n}$ in $\mathbb{R}^k$ with the k-means algorithm into clusters $C_1, \dots, C_k$.

Output: Clusters $A_1, \dots, A_k$ with $A_i = \{j \mid y_j \in C_i\}$.

計算(D^-1)L的eigenvector

Normalized spectral clustering according to Ng, Jordan, and Weiss （L_sym ）

Normalized spectral clustering according to Ng, Jordan, and Weiss (2002)

Input: Similarity matrix $S \in \mathbb{R}^{n\times n}$, number k of clusters to construct.

- Construct a similarity graph by one of the ways described in Section 2. Let W be its weighted adjacency matrix.
- Compute the normalized Laplacian L_{sym}.
- Compute the first k eigenvectors $u_1, \dots, u_k$ of L_{sym}.
- Let $U \in \mathbb{R}^{n\times k}$ be the matrix containing the vectors $u_1, \dots, u_k$ as columns.
- Form the matrix $T \in \mathbb{R}^{n\times k}$ from U by normalizing the rows to norm 1, that is set $t_{ij} = u_{ij}/(\sum_k u_{ik}^2)^{1/2}$.
- For $i = 1, \dots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of T.
- Cluster the points $(y_i)_{i=1,\dots,n}$ with the k-means algorithm into clusters $C_1, \dots, C_k$.

Output: Clusters $A_1, \dots, A_k$ with $A_i = \{j \mid y_j \in C_i\}$.

$$L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$$

Conclusion：

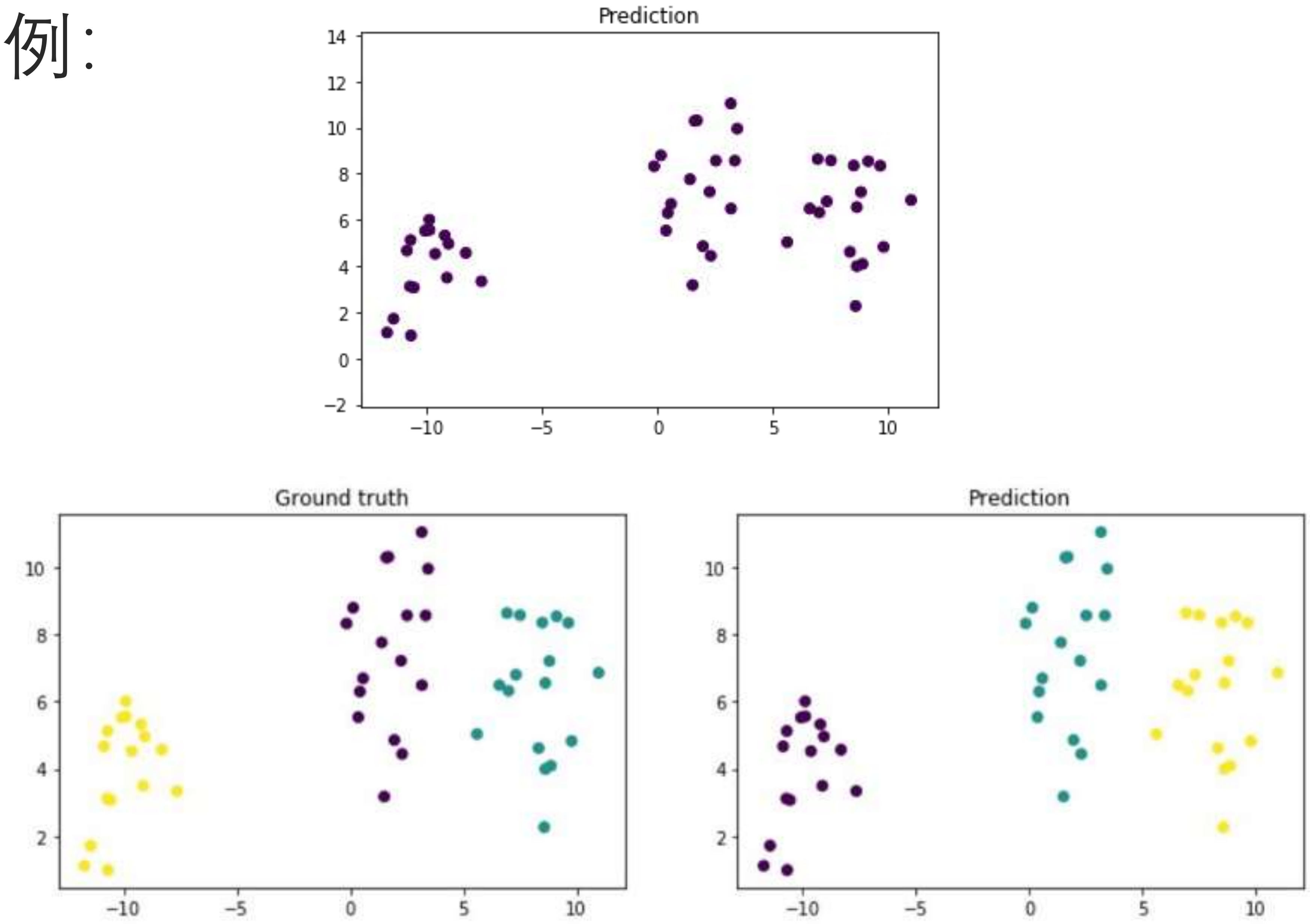
- 綜合上述的三個演算法，其實他們所做的事情，就是把資料用Laplacian Eigenmap降維，接著再以k-mean做clustering。

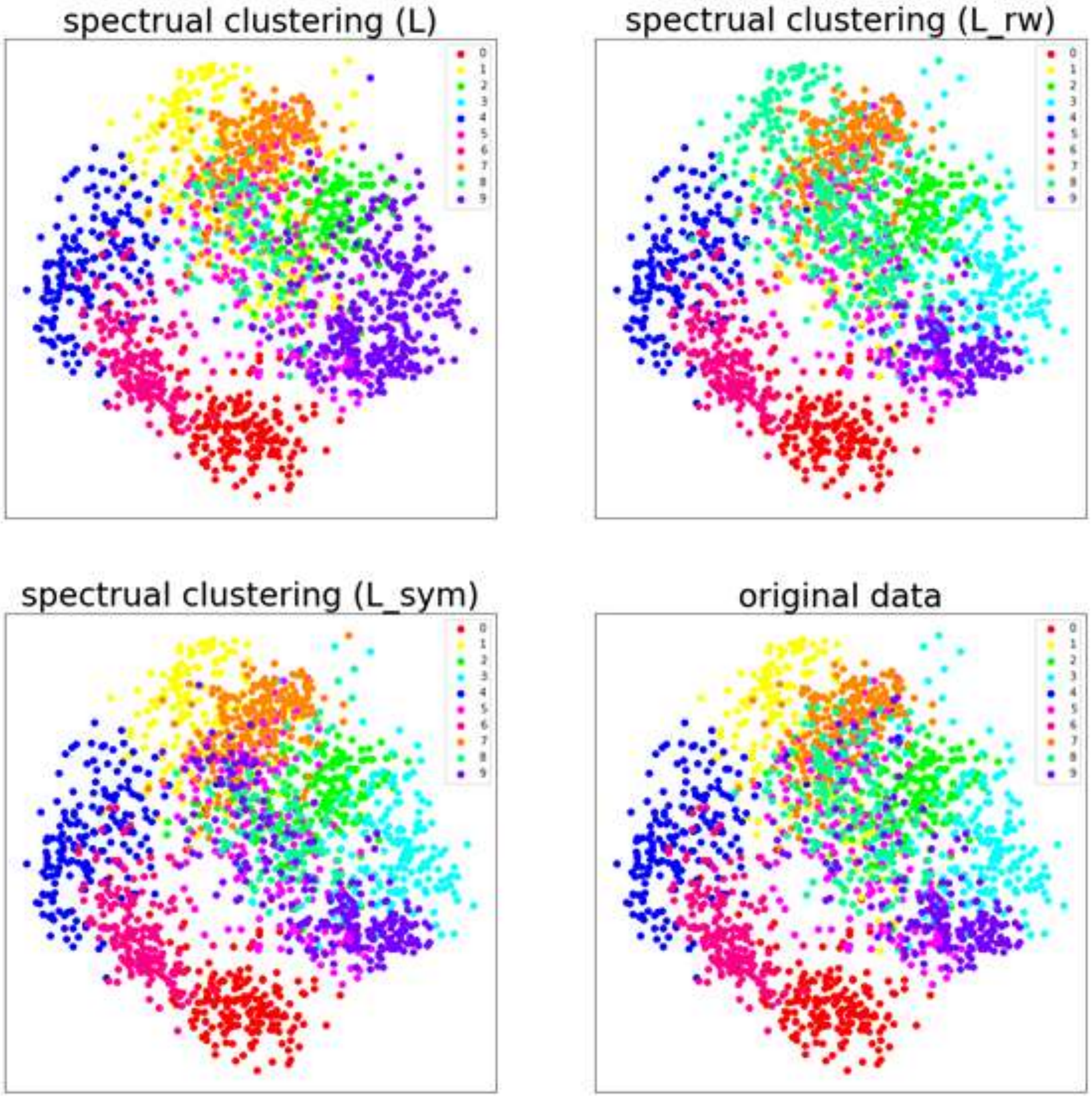
1.using normalized >>unnormalized spectral clustering,

2.Lrw >> $Lsym$

Why?

实作案例:




spectrual clustering (L)
spectrual clustering (L_rw)
spectrual clustering (L_sym)
original data