

# 聚类分析

1. 聚类分析定义:
2. 聚类方法:
3. 谱聚类:
  - 3.1 常见矩阵变换
  - 3.2 谱聚类流程
  - 3.3 谱聚类理论前提、证明
  - 3.4 图像分割实例结果
4. 总结:

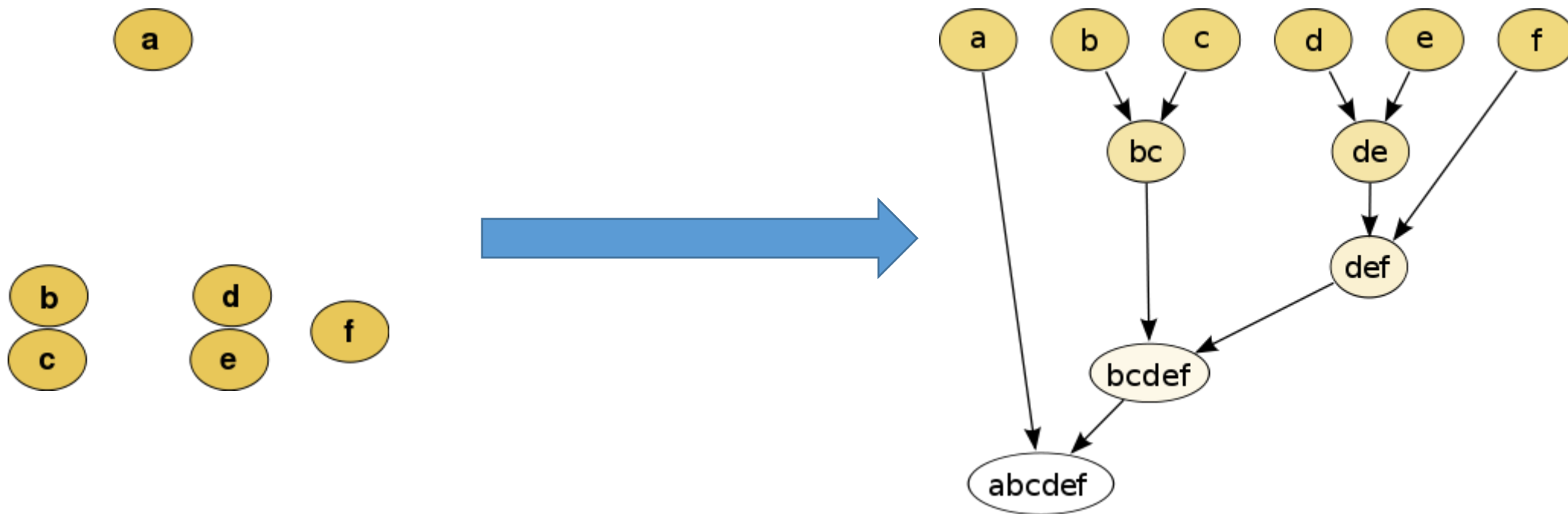
# 聚类分析：

- 聚类分析（Cluster analysis，亦称为群集分析）是对于静态[数据分析](#)的一门技术，在许多领域受到广泛应用，包括[机器学习](#)，[数据挖掘](#)，[模式识别](#)，[图像分析](#)以及[生物信息](#)。

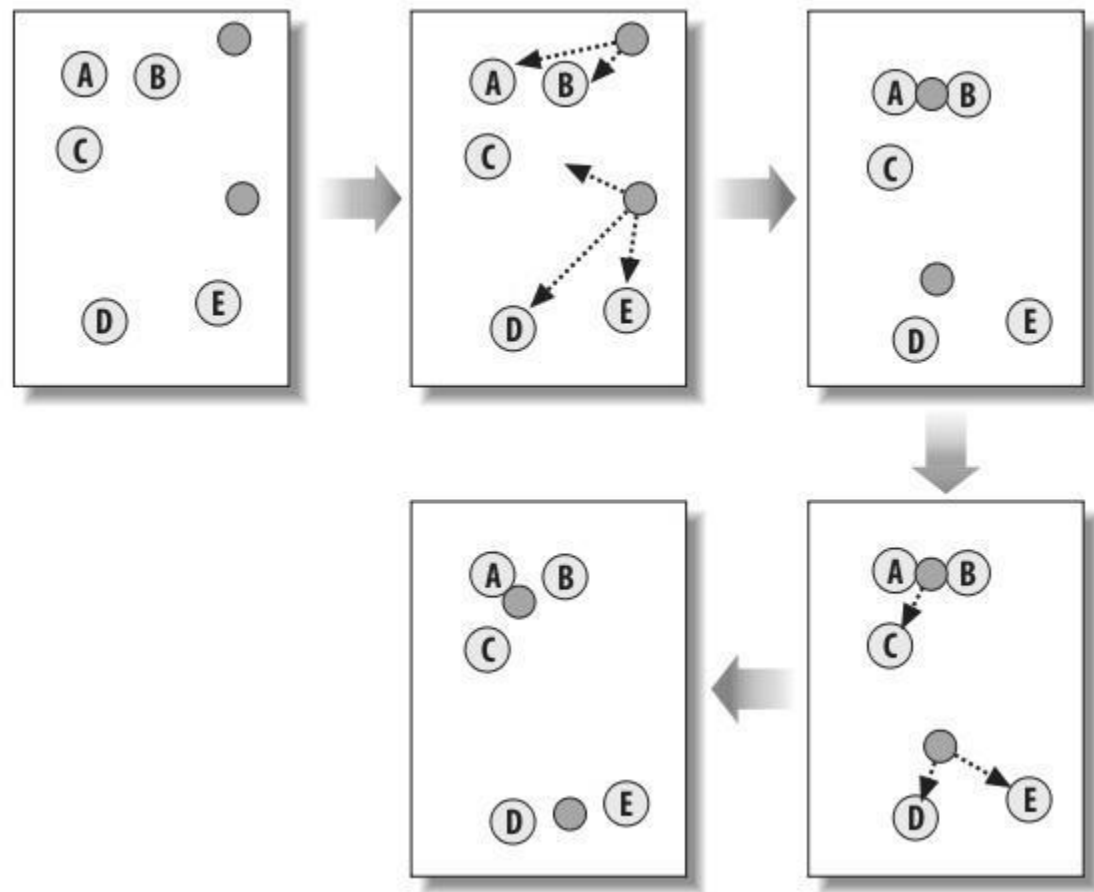
# 算法分类：

- 数据聚类算法可以分为结构性或者分散性。
  - **结构性算法** 以前成功使用过的聚类器进行分类。结构性算法可以从上至下或者从下至上双向进行计算。从下至上算法从每个对象作为单独分类开始，不断融合其中相近的对象。而从上至下算法则是把所有对象作为一个整体分类，然后逐渐分小。
  - **分散型算法** 是一次确定所有分类。K-均值法及衍生算法。
- 谱聚类（spectral clustering）

# 结构型： 层次聚类的一个例子：



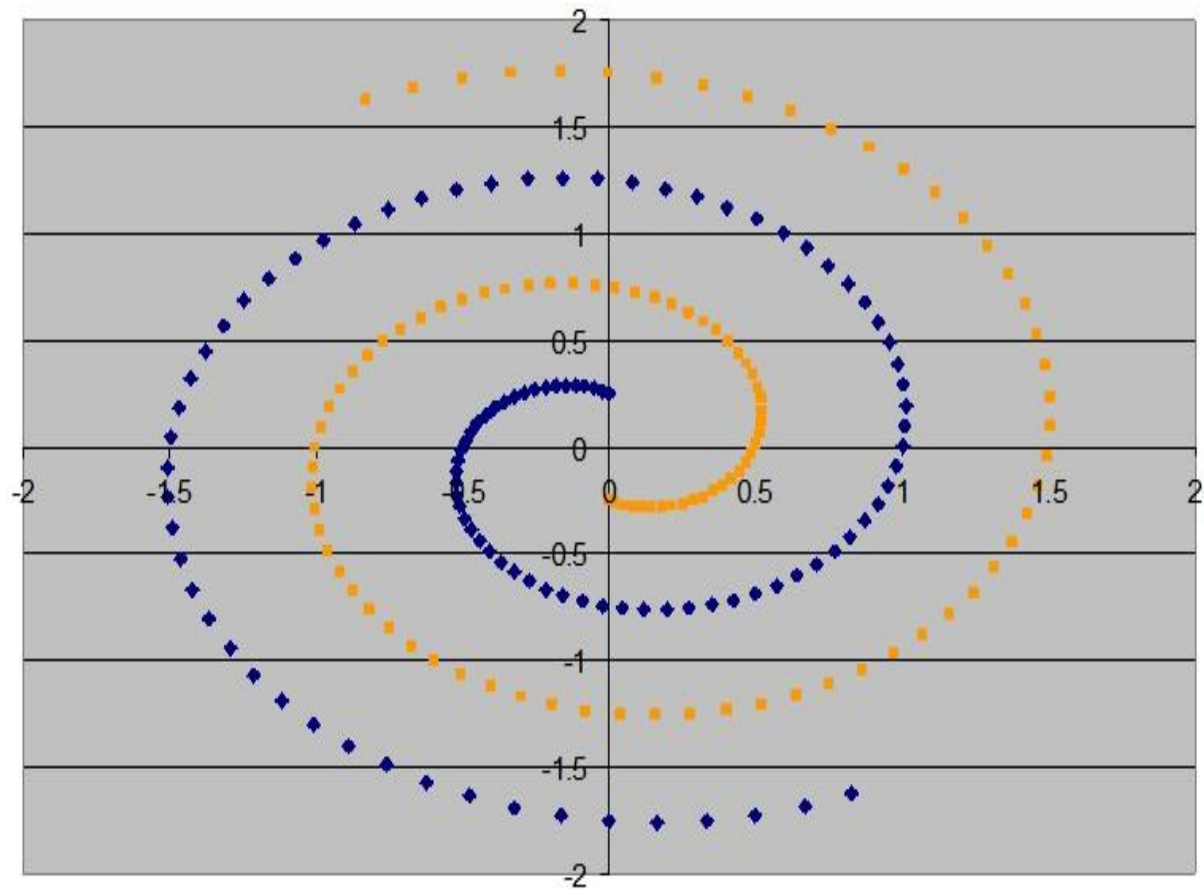
# 分散型：K-均值算法：



# 分散型k-means及其衍生算法的比较:

| K-means  | K-Medoids  |
|--|--|
| <p>K-Means算法:</p> <ol style="list-style-type: none"><li>1. 将数据分为k个非空子集</li><li>2. 计算每个类中心点 (k-means&lt;centroid&gt;中心点是所有点的average), 记为seed point</li><li>3. 将每个object聚类到最近seed point</li><li>4. 返回2, 当聚类结果不再变化的时候stop</li></ol> | <p>K-Medoids算法:</p> <ol style="list-style-type: none"><li>1. 任意选取K个对象作为medoids(<math>O_1, O_2, \dots, O_i \dots O_k</math>)。</li><li>2. 将余下的对象分到各个类中去 (根据与medoid最相近的原则);</li><li>3. 对于每个类 (<math>O_i</math>) 中, 顺序选取一个<math>O_r</math>, 计算用<math>O_r</math>代替<math>O_i</math>后的消耗<math>E(O_r)</math>。选择E最小的那个<math>O_r</math>来代替<math>O_i</math>。转到2。</li><li>4. 这样循环直到K个medoids固定下来。<br/>这种算法对于脏数据和异常数据不敏感, 但计算量显然要比K均值要大, 一般只适合小数据量。</li></ol> |

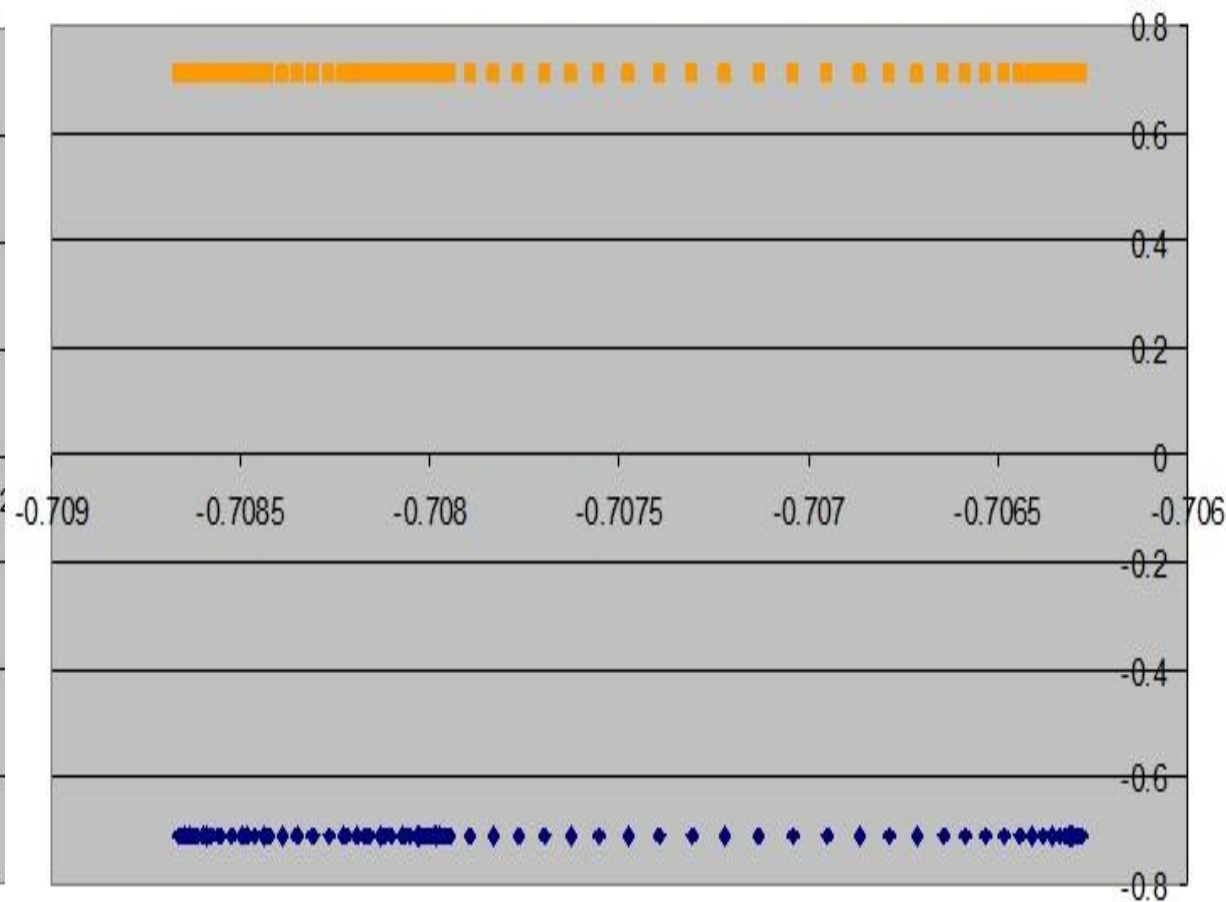
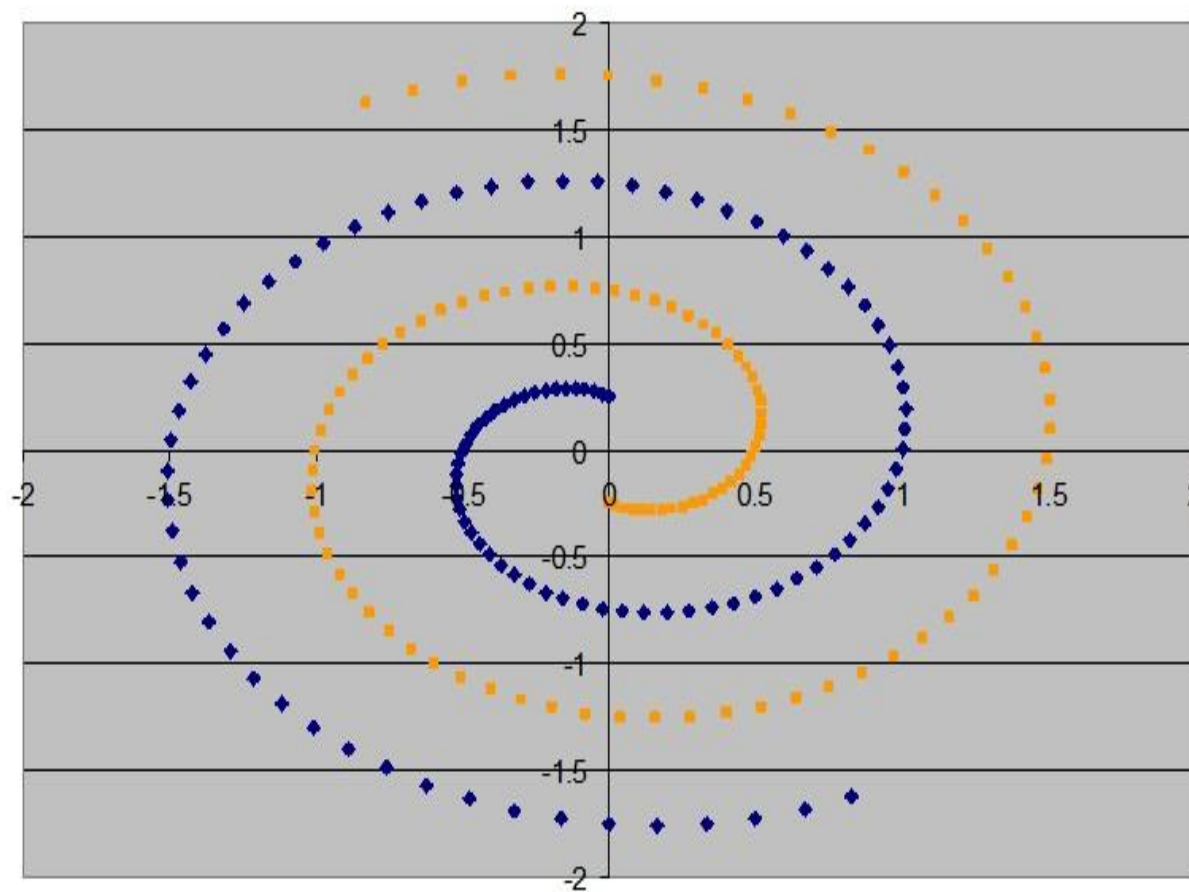
Spectral clustering (对于不同类数据相互交叉的情况, K-means无法解决。而通过空间转换则可以解决)



$$\rho = e^{\alpha\theta}$$

$$\rho = e^{\alpha(\theta-\pi)}$$

Spectral clustering (对于不同类数据相互交叉的情况, K-means无法解决。而通过空间转换则可以解决)





# 谱聚类算法流程:

1. 输入数据:  $d_1, d_2, \dots, d_n$ ;
2. 计算相似度矩阵  $W_{n \times n}$ , 其元素  $W(i,j)$  为数据  $d_i$  与  $d_j$  的相似度。(相似度计算的具体方法后面给出, 同时, 易知  $W$  为对称矩阵);
3. 计算矩阵  $D$ ,  $D$  为对角矩阵, 除对角元素外全为0,  $D$  的对角元素  $D(j,j) = \sum_{i=1}^n w_{i,j}$ 。  $D$  的对角元素为  $W_{n \times n}$  对应列的所有元素之和。
4. 计算矩阵  $L = D - W$ ;  $L$  为拉普拉斯矩阵。(不同的  $L$  矩阵定义对应不同的聚类准则, 如下文提到的  $L_{\text{sym}} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ ), 同时, 易知  $L$  为对称矩阵

# 谱聚类算法流程：

5. 求L的特征值并按照从小到大排列： $\gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_n$  (对称矩阵有n个实值的特征值)。
6. 对于K类聚类，选取前K个特征值所对应的特征向量，按列组成新的 $R = n \times k$ 维矩阵。
7. 把矩阵R的每行元素作为新的数据（共n个，每个数据为k维），使用K-means聚类。如果R的第i行元素被聚类到子类 $K_j$ ，那么原n个数据中的第i个数据属于子类j；

# 谱聚类流程的具体说明:

- 下面主要介绍:

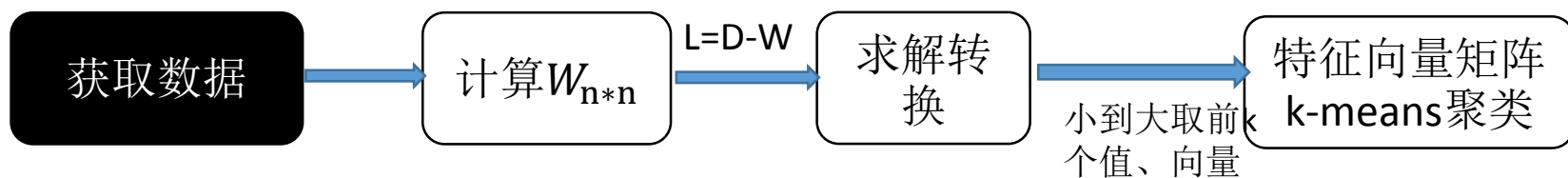
1. 相似度矩阵 $W_{n \times n}$ 的计算;

2. 聚类的准则函数类型和定义; 如:  $\min_A \text{RatioCut}(A, \bar{A}) = 1/2 \sum_{i \in A, j \in \bar{A}} W_{i,j}$

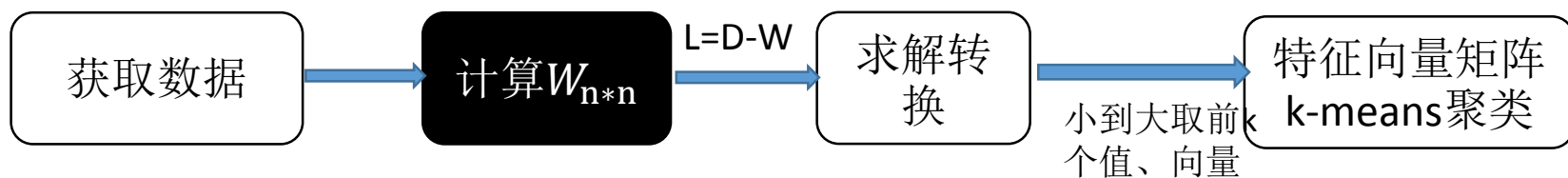
(其物理意义就是类A与非A类这两类之间的所有边的权值之和, 当边值的大、小表示点之间的相关、不相关时, 就是求一种分类A和非A, 使得这个准则函数对应的值最小。即类内相关、类间无关)。

3. 准则函数 $\text{RatioCut}(A, \bar{A})$ 的求解转换(将求解Ratiocut的较难的图分割问题转化为求较简单的f问题, 即使得下方左侧最小的f就是使得准则函数最小的分割, 具体见后面讲解)。



$$f' L f = \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2 = |V| \cdot \text{RatioCut}(A, \bar{A}).$$

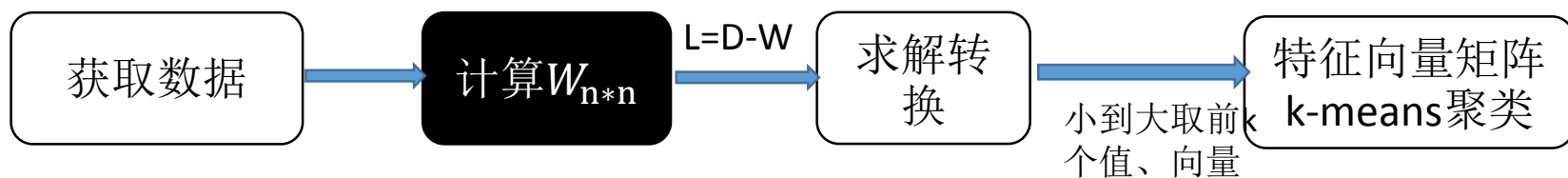


- 输入4\*6的图像image，则共计有4\*6=24个像素点，其中包括图像的像素灰度值信息、像素的位置信息等。



## 相似度矩阵计算

- 获取的图像数据中，24个像素包含灰度信息、RGB颜色及位置信息。
- 转化：
  - 24个像素  图中的点 $v_i$ 。  $i \in (1, 2, \dots, 24)$ ;
  - 像素与像素之间的信息  图中点 $v_i$ 与点 $v_j$ 之间的边，所以可以用 $24 \times 24$ 的矩阵表示图中点 $v_i$ 与点 $v_j$ 之间的关系 $w_{ij}$ 。（ $w_{ij}$ 表示图的邻接矩阵的第 $i$ 行第 $j$ 列元素）



## 相似度矩阵计算

### • 常见的转化方法:

#### 1. 高斯距离:

$$w_{ij} = \exp\left(-\frac{\|F_i - F_j\|_2^2}{\sigma_l^2}\right) \times \begin{cases} \exp\left(-\frac{\|X_i - X_j\|_2^2}{\sigma_l^2}\right) & \text{if } \|X_i - X_j\|_2 < r \\ 0 & \text{others} \end{cases}$$

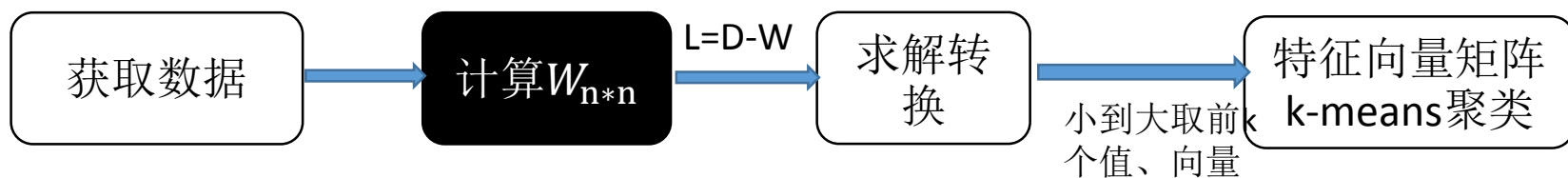
(高斯距离:  $F$ 为像素点灰度值,  $X$ 为像素坐标,  $r$ 为限定范围)

#### 2. 曼哈顿距离:

$$d_{12} = \sum_{k=1}^n |x_{1k} - x_{2k}|$$

#### 3. 欧式距离:

$$d_{12} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$



## 相似度矩阵计算

像素的灰度值、位置越相近，则边值越大，相关性就越强。

### • 常见的转化方法：

#### 1. 高斯距离：

$$w_{ij} = \exp\left(-\frac{\|F_i - F_j\|_2^2}{\sigma_l^2}\right) \times \begin{cases} \exp\left(-\frac{\|X_i - X_j\|_2^2}{\sigma_l^2}\right) & \text{if } \|X_i - X_j\|_2 < r \\ 0 & \text{others} \end{cases}$$

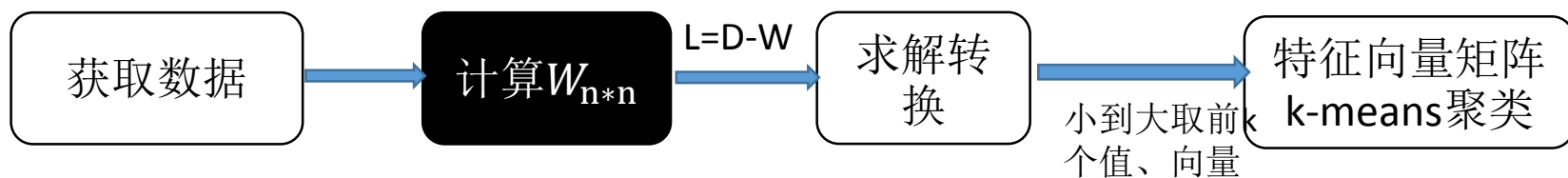
#### 2. 曼哈顿距离：

$$d_{12} = \sum_{k=1}^n |x_{1k} - x_{2k}|$$

#### 3. 欧式距离：

$$d_{12} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

曼哈顿距离和欧式距离区别于高斯距离，两点越相关距离越小。越无关距离越大。所以高斯距离准则函数求极小而曼哈都欧式距离求准则函数最大值。具体见后面的准则函数定义。

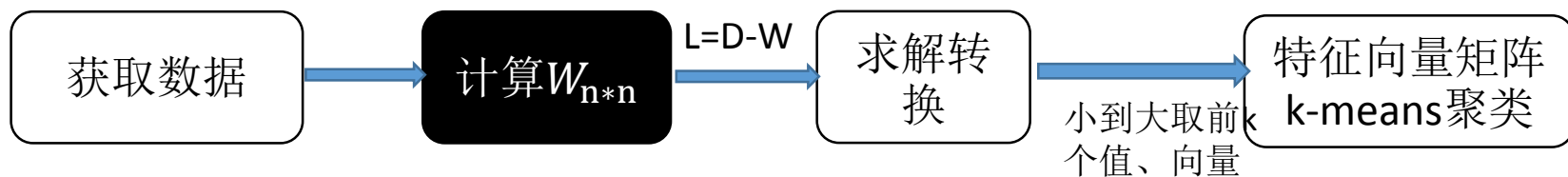


## 相似度矩阵计算

- 转化到图后：

1. 邻接矩阵  $W_{24*24}$ ;
2.  $D_{24*24}$ ; ( $W$ 矩阵的 $i$ 列所有元素之和作为 $D$ 的 $(i,i)$ 元素, 即  $D_{ii} = \sum_{j=1}^{24} w_{ji}$ , 实际上 $W$ 为对称的。),  $D$ 的非对角元素全为0;
3. Laplacian矩阵:  $L = D - W$ ;
4. 3中的 $L$ 矩阵, 是其中一种方法, 还有其他的 $L$ 的定义, 如:
  1.  $L_{\text{sym}} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ ; symmetric矩阵,  $I$ 表示单位矩阵。
  2.  $L_{\text{rw}} := D^{-1} L D^{\frac{1}{2}} = I - D^{-1} W$ ; random walk矩阵,  $I$ 表示单位矩阵。

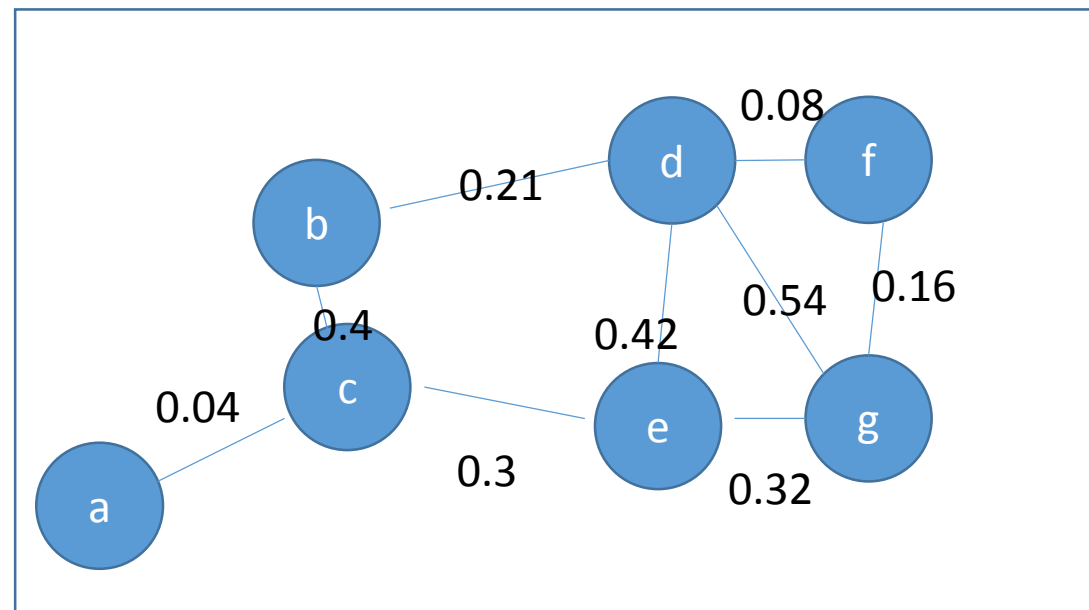


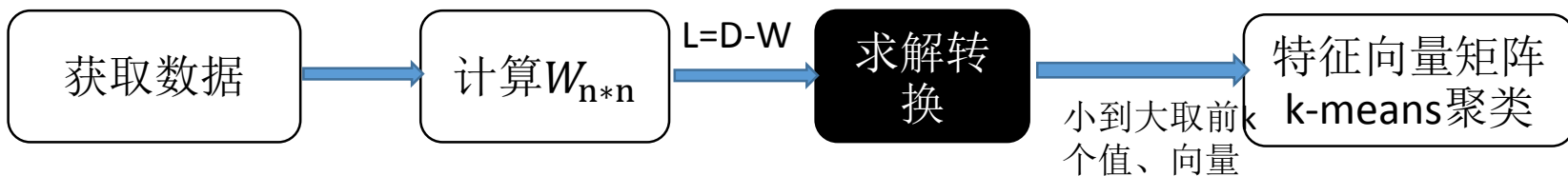


## 相似度矩阵计算

- 右图的一个 $W$ 矩阵实例：

|   | a    | b    | c    | d    | e    | f    | g    |
|---|------|------|------|------|------|------|------|
| a | 1    | 0    | 0.04 | 0    | 0    | 0    | 0    |
| b | 0    | 1    | 0.4  | 0.21 | 0    | 0    | 0    |
| c | 0.04 | 0.4  | 1    | 0    | 0.3  | 0    | 0    |
| d | 0    | 0.21 | 0    | 1    | 0.42 | 0.08 | 0.54 |
| e | 0    | 0    | 0.3  | 0.42 | 1    | 0    | 0.32 |
| f | 0    | 0    | 0    | 0.08 | 0    | 1    | 0.16 |
| g | 0    | 0    | 0    | 0.54 | 0.32 | 0.16 | 1    |

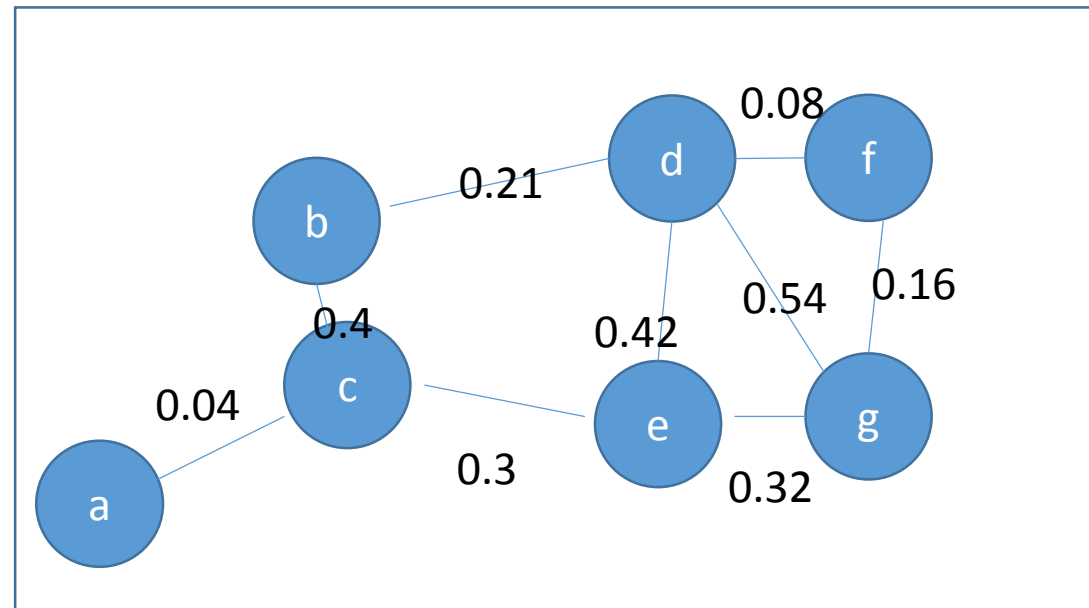




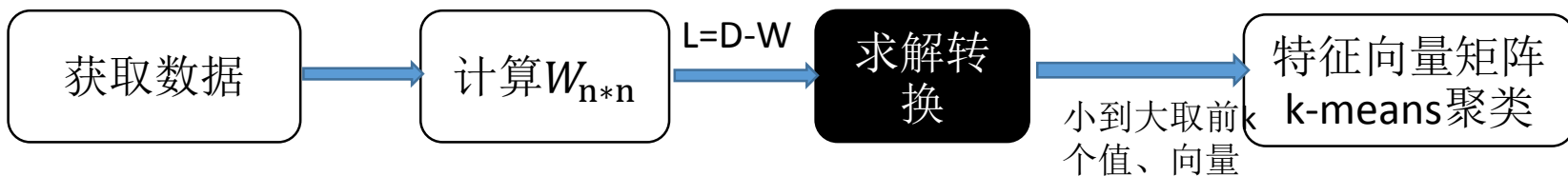
## 图分割、准则函数及准则函数最小化问题转换

- **准则函数1:**
- Min cut: 图G分割为两类A,  $\bar{A}$  ( $\bar{A}$ 表示不属于A子图的点);
- $\text{Min cut}(A, \bar{A}) = \sum_{i \in A, j \in \bar{A}} w_{ij}$

$$\text{Min cut}(A_1, \dots, A_k) := \frac{1}{2} \sum_{i=1}^k W(A_i, \bar{A}_i).$$



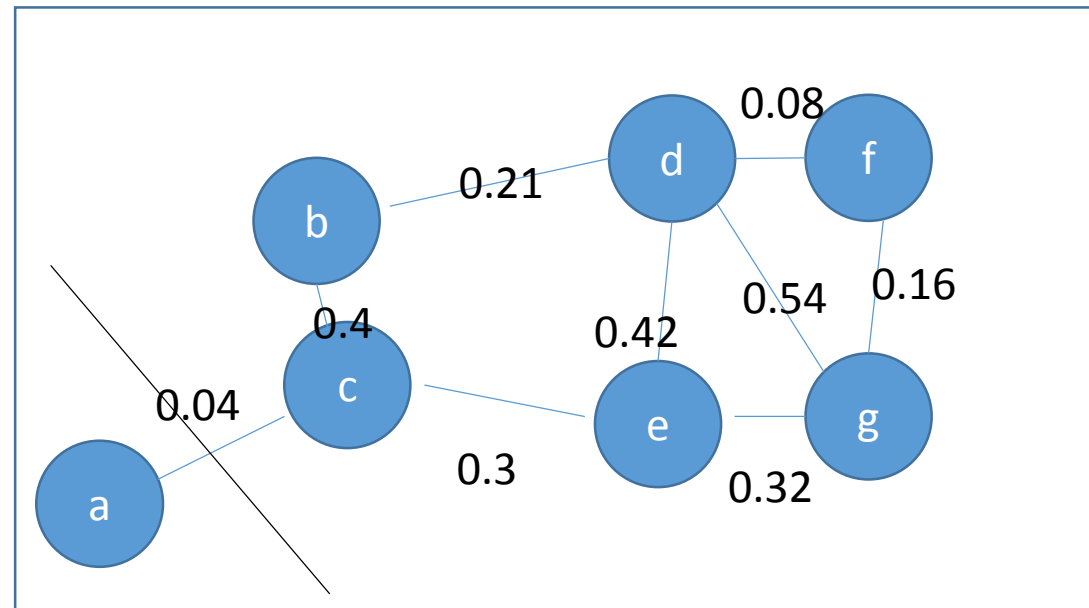
Min Cut分割。



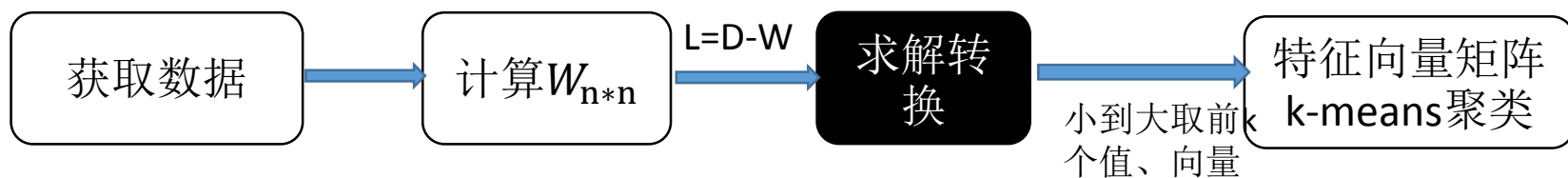
## 图分割、准则函数及准则函数最小化问题转换

- **准则函数1:**
- Min cut: 图G分割为两类A,  $\bar{A}$  ( $\bar{A}$ 表示不属于A子图的点);
- $\text{Min cut}(A, \bar{A}) = \sum_{i \in A, j \in \bar{A}} w_{ij}$

$$\text{Min cut}(A_1, \dots, A_k) := \frac{1}{2} \sum_{i=1}^k W(A_i, \bar{A}_i).$$



Min Cut分割。

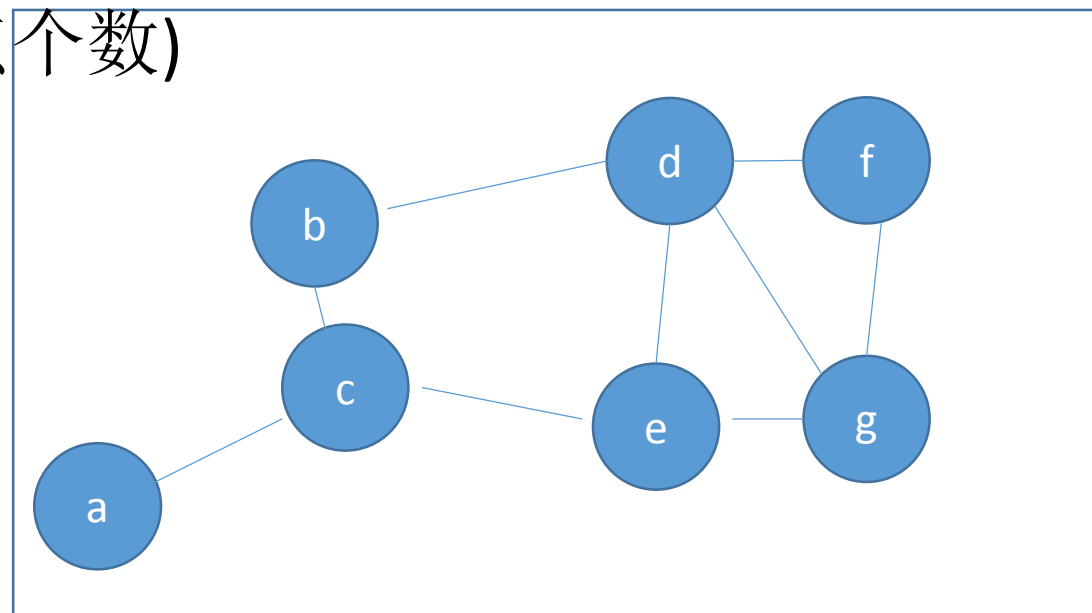


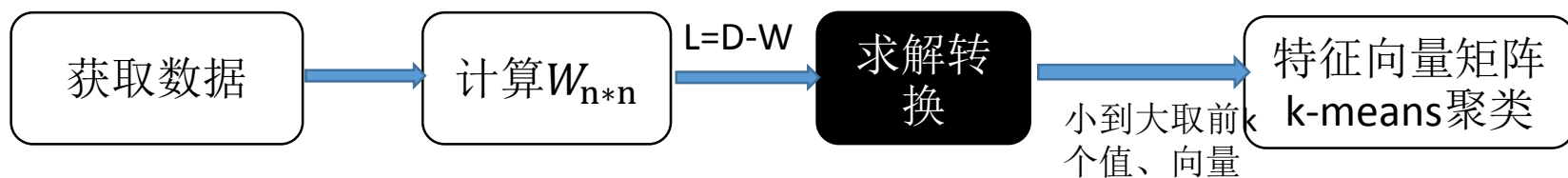
## 图分割、准则函数及准则函数最小化问题转换

**准则函数2:**

$\text{RatioCut}(A, \bar{A})$  ( $|A_i|$  表示子图  $A_i$  中的顶点个数)

$$\min \text{RatioCut}(A_1, \dots, A_k) := \sum_{i=1}^k \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|}$$



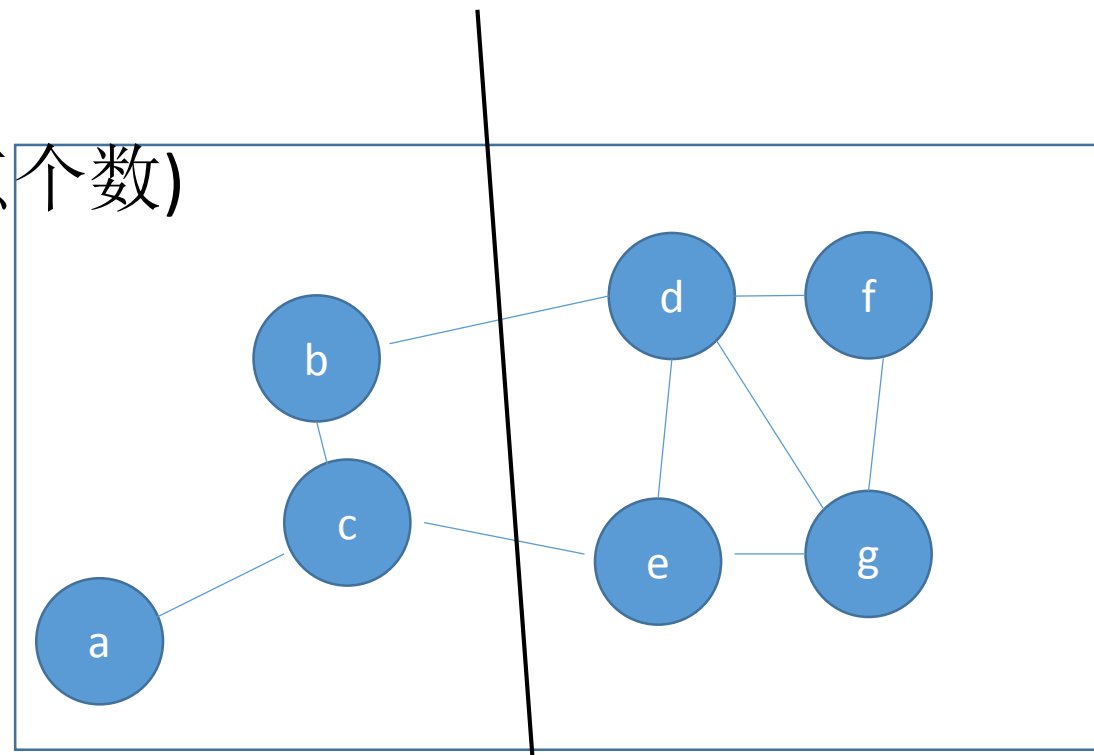


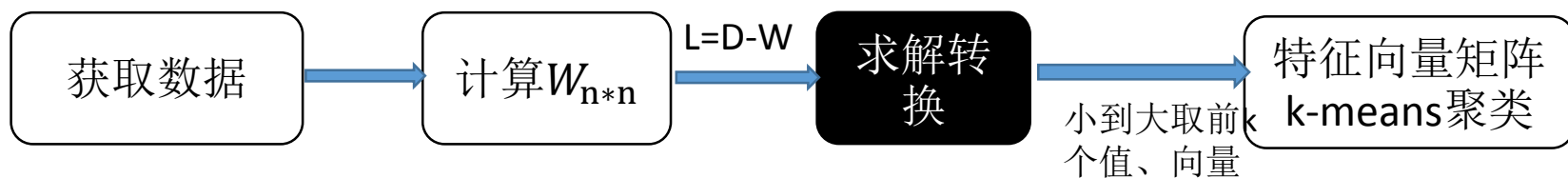
## 图分割、准则函数及准则函数最小化问题转换

**准则函数2:**

$\text{RatioCut}(A, \bar{A})$  ( $|A_i|$  表示子图  $A_i$  中的顶点个数)

$$\min \text{RatioCut}(A_1, \dots, A_k) := \sum_{i=1}^k \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|}$$





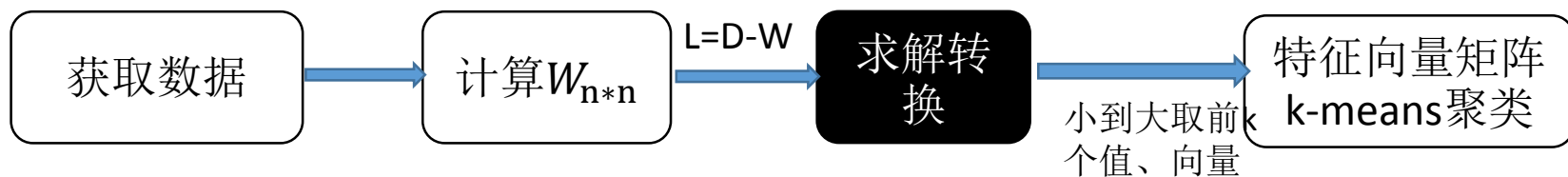
## 图分割、准则函数及准则函数最小化问题转换

**准则函数3:**  $Ncut(A, \bar{A})$

$$\min Ncut(A_1, \dots, A_k) := \sum_{i=1}^k \frac{cut(A_i, \bar{A}_i)}{vol(A_i)}.$$

其中， $vol(A_i)$ 表示的是子图 $A_i$ 中的每个顶点上所有边的权值和。

$$d_i = \sum_{j=1}^n w_{ij}. \quad vol(A) := \sum_{i \in A} d_i.$$



## 图分割、准则函数及准则函数最小化问题转换

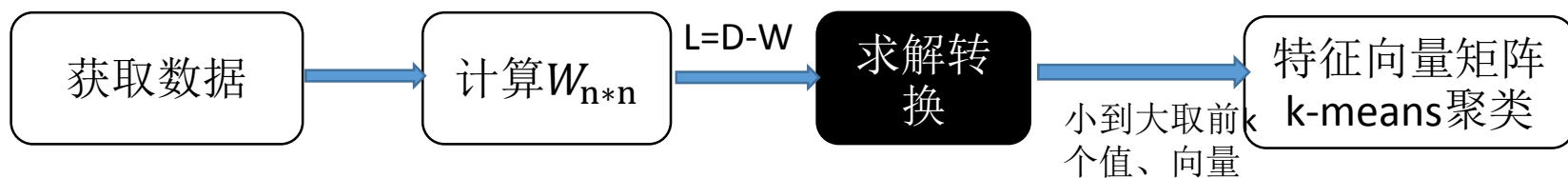
• 以准则函数2为例:

$$\min \text{RatioCut}(A_1, \dots, A_k) := \sum_{i=1}^k \frac{\text{cut}(A_i, \overline{A_i})}{|A_i|}$$

对于含有7个顶点的图（见上前面的图，对于4\*6的图像，则可以转为到24个点的图，图的边表示两个像素之间的相似度，故后面就以7个点的图为例），如果分为两类，那么：

1. 当一类只含有1个元素时有： $C_7^1$ 种可能。（排列组合）
2. 当一类只含有2个元素是有： $C_7^2$ 种可能；
3. 、 、 、

每一种情况对应一个准则函数的值准则函数最小时对应的分类就是聚类结果，但是在数据个数较大时，是个非线性时间复杂度问题，也就是NP问题。



## 图分割、准则函数及准则函数最小化问题转换

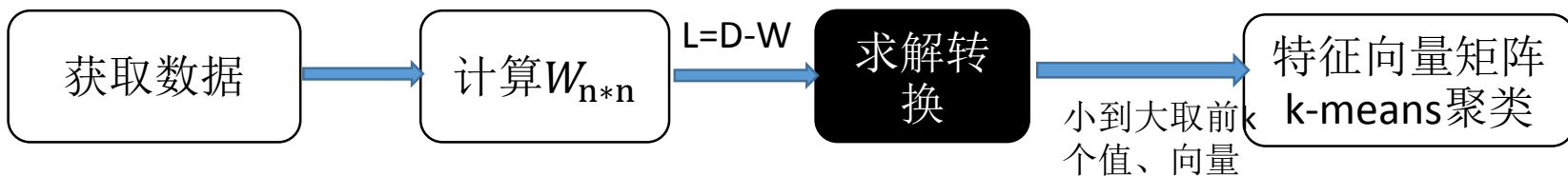
- 为了求最好的分割将准则函数用另一种形式表示（即问题的转化）

- 对于前面得到的对称矩阵L:

$$f' L f = \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2. \quad \text{这里对于任意 } f(f_1, f_2, \dots, f_7) \text{ 成立。 (附录1)}$$

- 如果把f的每个元素分别对应到图的7点，那么，当第i个顶点 $v_i$ 属于子图A时，那么取 $f_i$ 的值为 $\sqrt{|\bar{A}|/|A|}$ ，否则顶点 $v_i$ 就是属于 $\bar{A}$ ，此时 $f_i$ 取的值则为 $-\sqrt{|A|/|\bar{A}|}$ 。





## 图分割、准则函数及准则函数最小化问题转换

- 例如：七个顶点的图分割共有  $C_7^1 + C_7^2 + \dots + C_7^6$  种分割情况；
- 如果点a、c被分为一类其余点被分为另一类，则对应的f设置为：

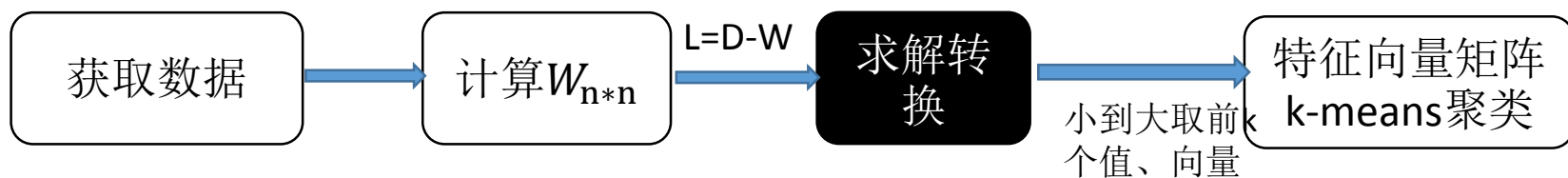
- $F(\sqrt{5/2}, -\sqrt{\frac{2}{5}}, \sqrt{5/2}, -\sqrt{\frac{2}{5}}, -\sqrt{\frac{2}{5}}, -\sqrt{\frac{2}{5}}, -\sqrt{\frac{2}{5}});$

其中  $(\sqrt{|\bar{A}|/|A|} = \sqrt{5/2}, -\sqrt{|A|/|\bar{A}|} = -\sqrt{\frac{2}{5}})$

- 如果点a、b、d被分为一类其余被分为另一类，则对应的f为：

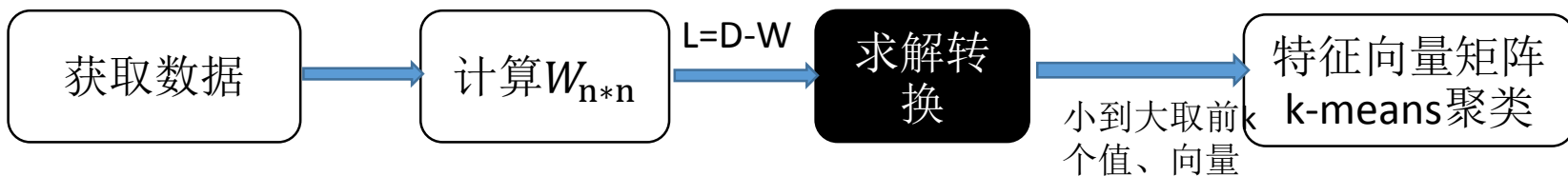
- $F(\sqrt{4/3}, \sqrt{4/3}, -\sqrt{\frac{3}{4}}, \sqrt{4/3}, -\sqrt{\frac{3}{4}}, -\sqrt{\frac{3}{4}}, -\sqrt{\frac{3}{4}});$

其中  $(\sqrt{|\bar{A}|/|A|} = \sqrt{4/3}, -\sqrt{|A|/|\bar{A}|} = -\sqrt{3/4})$



## 图分割、准则函数及准则函数最小化问题转换

- 为了求最好的分割将准则函数用另一种形式表示（即问题的转化）
- 这样，对于含有7个点的图分割问题，是个NP问题，对于每种可能的分割，就对应着一个 $f(f_1, f_2, \dots, f_{24})$ 的值：即如果某个顶点 $v_i$ 属于子图A，那么我们就知道其对应的 $f_i$ 值为 $\sqrt{|\bar{A}|/|A|}$ 。
- 反过来，每一个f就对应着一个分割：若第i个元素是 $\sqrt{|\bar{A}|/|A|}$ ，那么我们就可以知道点 $v_i$ 属于子图A，如果是 $-\sqrt{A/|\bar{A}|}$ ，那么点 $v_i$ 就属于子图 $\bar{A}$ 。
- 其实，通过以上的步骤我们已经转换了问题；

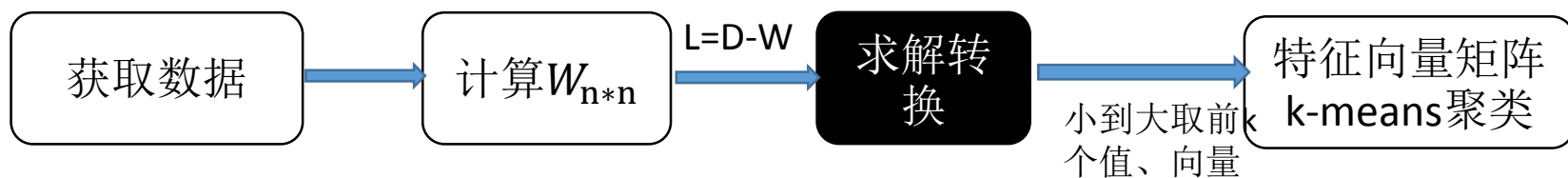


## 图分割、准则函数及准则函数最小化问题转换

- 转换的原理在于：把 $f$ 带入下式：

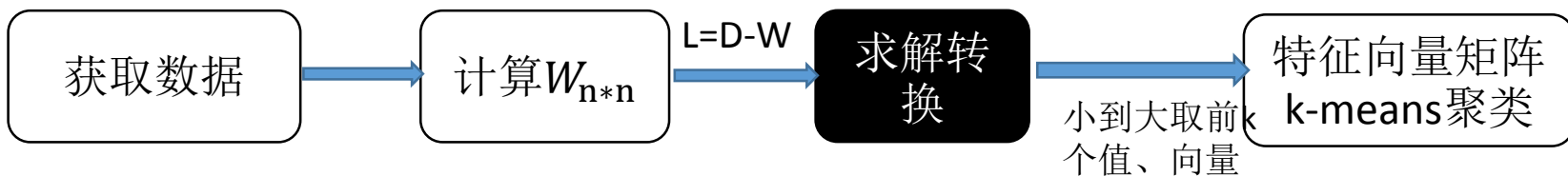
$$f^T L f = \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2 = |V| \cdot \text{RatioCut}(A, \bar{A}).$$

- 也就是当 $f$ 的元素满足前面的约束时 $f^T L f$ 就是图分割的第二种准则函数，所以，求图分割的准则函数的最小值就转化为求 $f$ 的值，使得 $f^T L f$ 值最小。而实际上每一个 $f$ 也就是一种分割，可以通过 $f$ 的第 $i$ 个元素的值是 $\sqrt{|\bar{A}|/|A|}$ 或 $-\sqrt{A/|\bar{A}|}$ 而推导出图的第 $i$ 个点属于或不属于类 $A$ 。



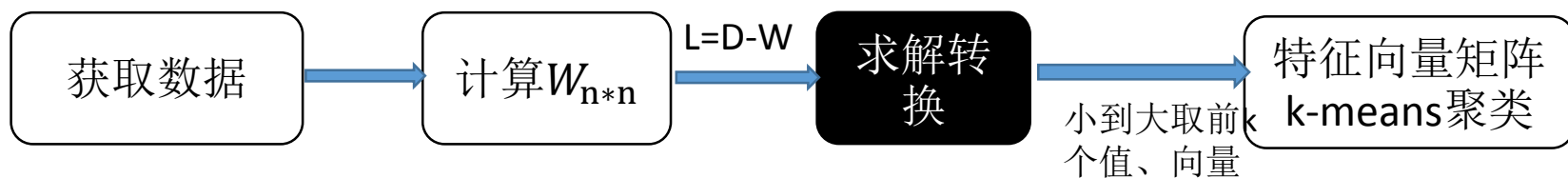
## 图分割、准则函数及准则函数最小化问题转换

- 而实际上 $f$ 的取值也是个NP问题（因为每个 $f$ 对应一个图分割，而图分割是个NP问题，所以 $f$ 的取值也对应有一样的取值种类）。
- 那为什么还要引入 $f$ 呢？
- 因为对于图的分割我们难以解决，而转换到 $f$ 后，我们把 $f$ 的约束（元素值为 $\sqrt{|\bar{A}|/|A|}$ 或 $-\sqrt{A/|\bar{A}|}$ ）放宽为实数，若 $f$ 的第 $i$ 个值大于0（相当于元素值为 $\sqrt{|\bar{A}|/|A|}$ ）就认为第 $i$ 个顶点属于A类，否则小于0（相当于元素值为 $-\sqrt{A/|\bar{A}|}$ ），则认为第 $i$ 个点属于非A类。
- F元素值的放宽的原因是 $f$ 放宽到实数域后就可以很方便得到解答。



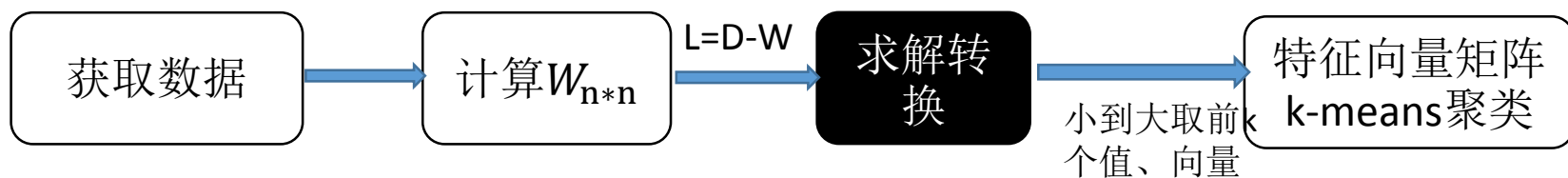
## 图分割、准则函数及准则函数最小化问题转换

- $\min_A \text{RatioCut}(A, \bar{A}) = \min_A \frac{f^T L f}{f^T f} = \frac{f^T L f}{|V|} = > \min_{f \in \mathbb{R}^n} f^T L f;$ 
  - 其中  $f$  垂直于向量  $\mathbf{1}$ ,  $\|f\|^2 = |V|$ ;
- $\min_{f \in \mathbb{R}^n} \text{RatioCut}(A, \bar{A}) = \min_{f \in \mathbb{R}^n} f^T L f;$ 
  - 其中  $f$  垂直于向量  $\mathbf{1}$ ,  $\|f\|^2 = |V|$ ;
- *Rayleigh quotient*(转换后求解  $f$  用到的原理)
  - $F(f) = \frac{f^T L f}{f^T f} = \frac{f^T L f}{|V|}$
  - $F(f)$  最大值和最小值分别在  $f$  取值为  $L$  的最大、最小特征值所对应的特征向量时取得。



## 图分割、准则函数及准则函数最小化问题转换

- 实际计算发现， $L$ 的最小特征值为0，所对应的特征向量为常向量。所以，实际中取的是第二小特征值所对应的特征向量。
- 第二小特征值对应的特征向量的元素并非满足要么 $\sqrt{|\bar{A}|/|A|}$ 要么 $-\sqrt{A/|\bar{A}|}$ 的特定大小，而是放宽了限制，允许元素为实数即可，。
- 没放宽之前，若特征向量的第 $i$ 个元素是 $\sqrt{|\bar{A}|/|A|}$ 则对应的第 $i$ 个顶点属于子图 $A$ ，否则属于 $\bar{A}$ 。放宽后，可以设定若元素为正数则对应属于子图 $A$ ，否则属于 $\bar{A}$ 。



- 前面讲解的是二聚类，对于K（K>2）聚类的转换同理：

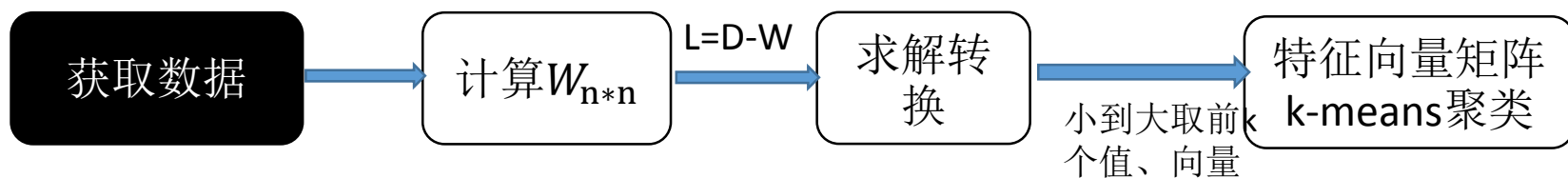
$$\text{RatioCut}(A_1, \dots, A_k) := \frac{1}{2} \sum_{i=1}^k \frac{W(A_i, \bar{A}_i)}{|A_i|} = \sum_{i=1}^k \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|}$$

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

其中 $f_i$ 是向量 $f$ 的第 $i$ 个元素;  
 $f$ 垂直于向量 $\mathbf{1}$ ,  $\|f\|^2 = n$ ;

$$h_{i,j} = \begin{cases} 1/\sqrt{|A_j|} & \text{if } v_i \in A_j \\ 0 & \text{otherwise} \end{cases}$$

其中 $h_{i,j}$ 是第 $j$ 个类别的第 $i$ 个元素;  
 $H$ 是由 $h_j$ 组成的 $n * k$ 矩阵。 $H^T H = I$



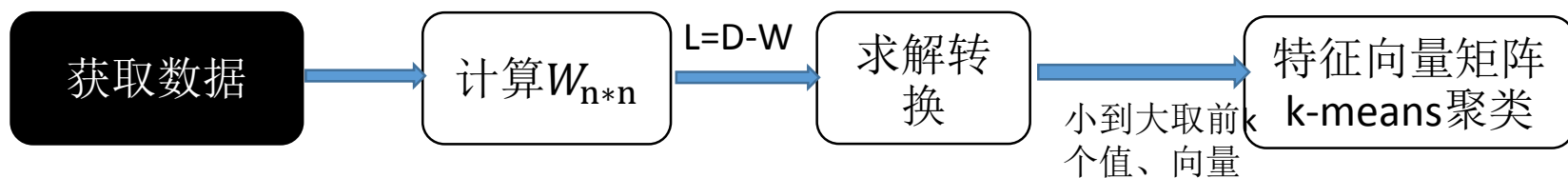
## K聚类

$$h_{i,j} = \begin{cases} 1/\sqrt{|A_j|} & \text{if } v_i \in A_j \\ 0 & \text{otherwise} \end{cases} \quad (i = 1, \dots, n; j = 1, \dots, k).$$

$$h'_i L h_i = \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|}. \quad h'_i L h_i = (H' L H)_{ii}.$$

$$\text{RatioCut}(A_1, \dots, A_k) = \sum_{i=1}^k h'_i L h_i = \sum_{i=1}^k (H' L H)_{ii} = \text{Tr}(H' L H),$$





$$\min_{A_1, \dots, A_k} \text{Tr}(H' L H) \text{ subject to } H' H = I,$$

$$\min_{H \in \mathbb{R}^{n \times k}} \text{Tr}(H' L H) \text{ subject to } H' H = I.$$

根据Rayleigh-Ritz定理，最小化问题的解为L的前K个特征值的对应特征向量按列组成的K列矩阵。

# RatioCut聚类→Ncut分类

$$\text{RatioCut}(A_1, \dots, A_k) := \frac{1}{2} \sum_{i=1}^k \frac{W(A_i, \bar{A}_i)}{|A_i|} = \sum_{i=1}^k \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|}$$

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

其中 $f_i$ 是向量 $f$ 的第 $i$ 个元素;  
 $f$ 垂直于向量 $\mathbf{1}$ ,  $\|f\|^2 = n$ ;

$$h_{i,j} = \begin{cases} 1/\sqrt{|A_j|} & \text{if } v_i \in A_j \\ 0 & \text{otherwise} \end{cases}$$

其中 $h_{i,j}$ 是第 $j$ 个类别的第 $i$ 个元素;  
 $H$ 是由 $h_j$ 组成的 $n * k$ 矩阵。 $H^T H = I$

# Ratiocut聚类→Ncut分类

$$f_i = \begin{cases} \sqrt{\frac{\text{vol}(\bar{A})}{\text{vol}(A)}} & \text{if } v_i \in A \\ -\sqrt{\frac{\text{vol}(A)}{\text{vol}(\bar{A})}} & \text{if } v_i \in \bar{A}. \end{cases}$$

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

其中 $f_i$ 是向量 $f$ 的第 $i$ 个元素;  
 $f$ 垂直于向量 $\mathbf{1}$ ,  $\|f\|^2 = n$ ;

$$(Df)' \mathbf{1} = 0 \quad f' D f = \text{vol}(V)$$

$$h_{i,j} = \begin{cases} 1/\sqrt{\text{vol}(A_j)} & \text{if } v_i \in A_j \\ 0 & \text{otherwise} \end{cases}$$

$$h_{i,j} = \begin{cases} 1/\sqrt{|A_j|} & \text{if } v_i \in A_j \\ 0 & \text{otherwise} \end{cases}$$

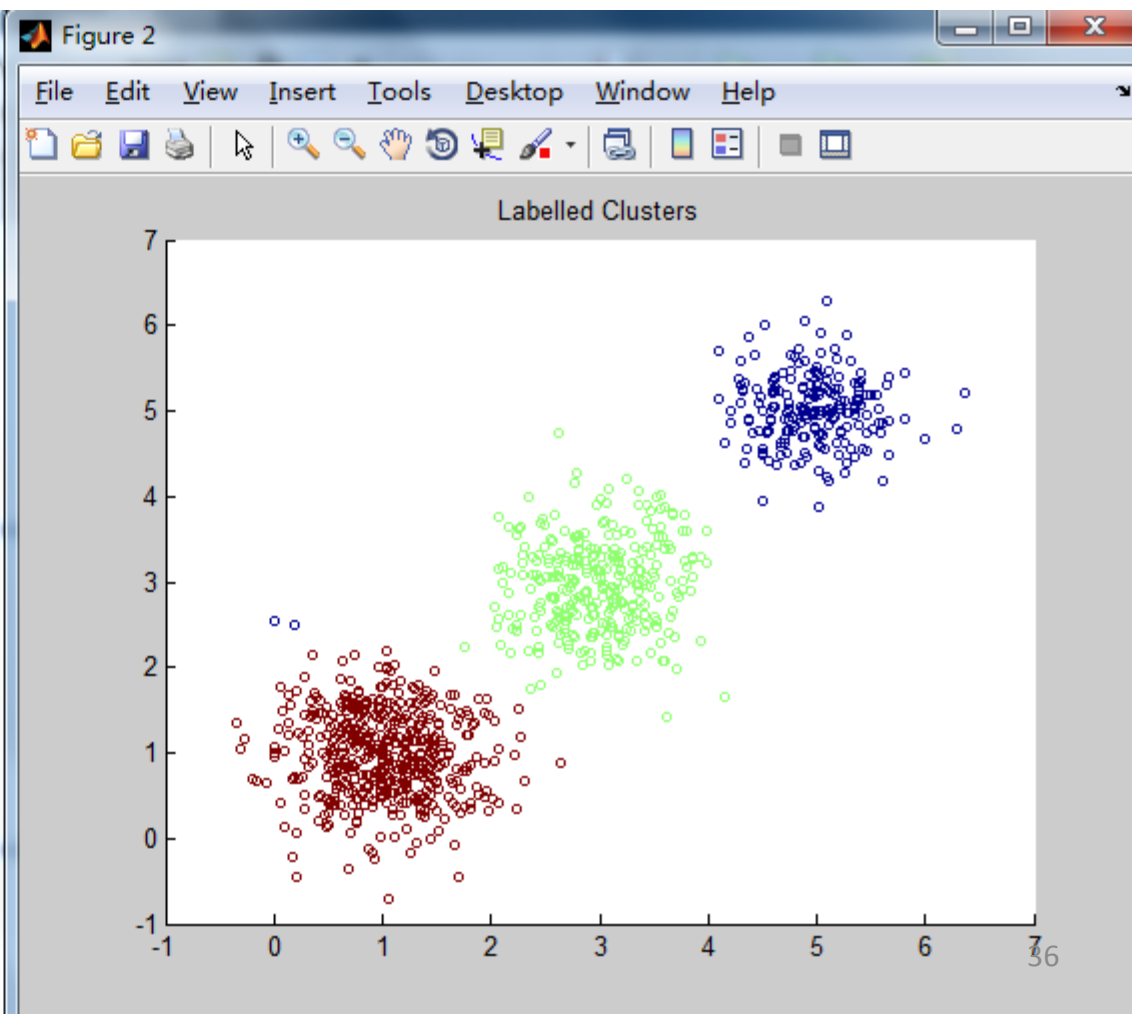
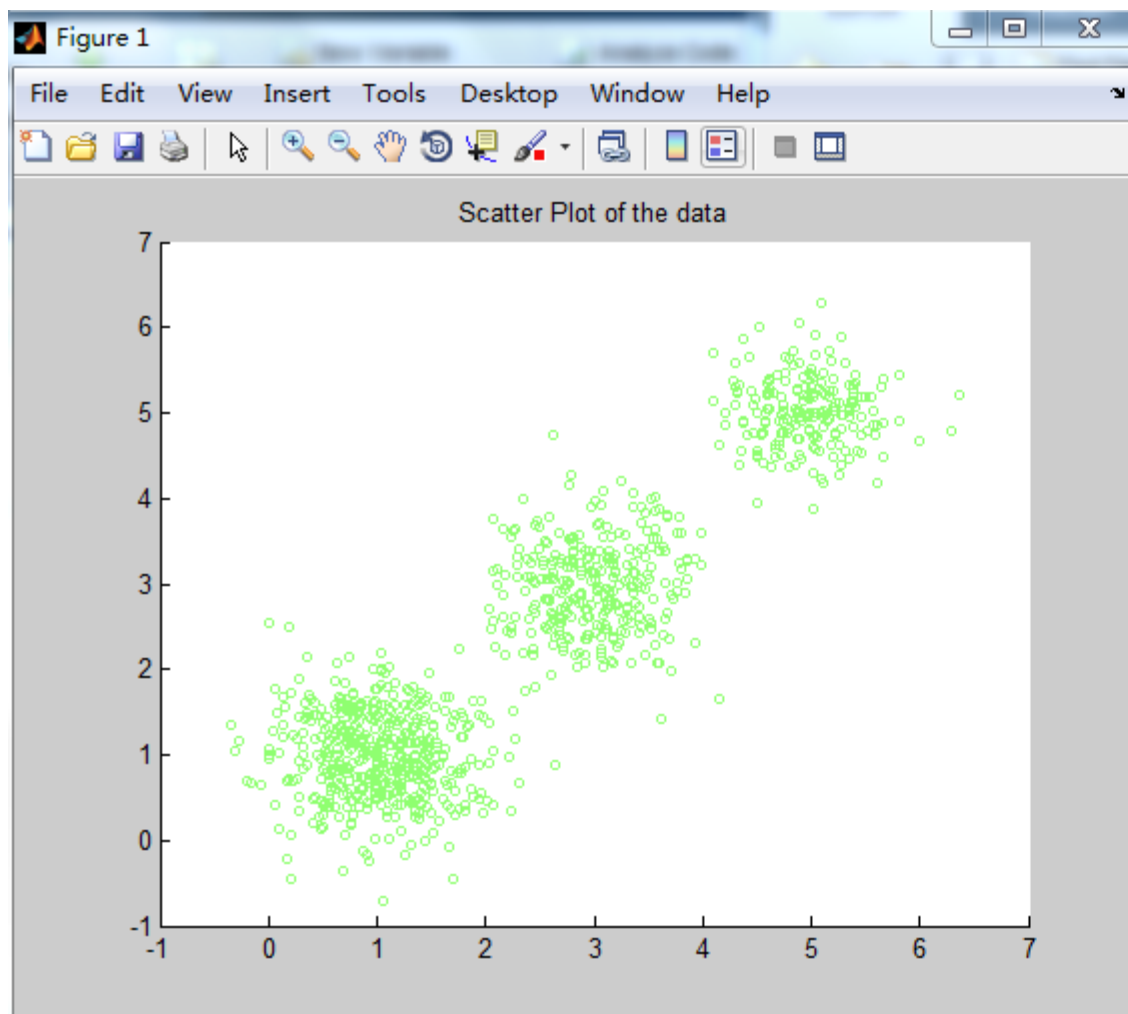
其中 $h_{i,j}$ 是第 $j$ 个类别的第 $i$ 个元素;  
 $H$ 是由 $h_j$ 组成的 $n * k$ 矩阵。  $H^T H = I$

$$H' H = I,$$

$$h'_i D h_i = 1,$$

$$h'_i L h_i = \text{cut}(A_i, \bar{A}_i) / \text{vol}(A_i).$$

## 三个混合高斯的分类:



Original Image



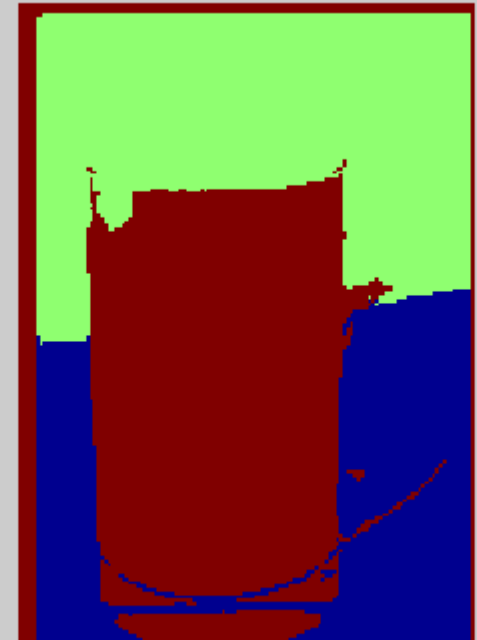
原图

Partition Image



K=2

Partition Image



K=3

Original Image



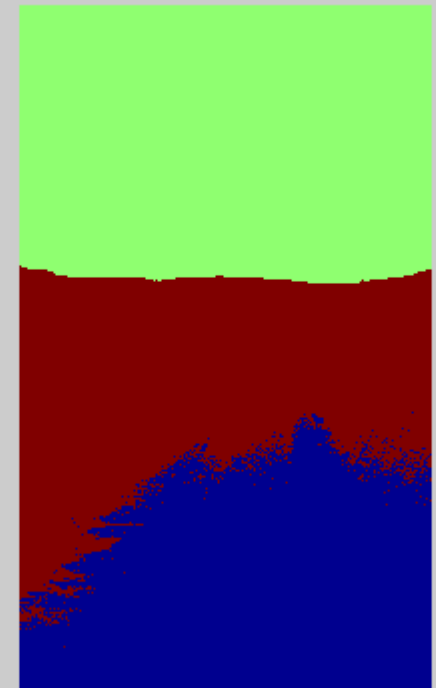
原图

Partition Image



K=2

Partition Image



K=3

# 总结:

1. K-means在数据维度和数据量较大时复杂度为 $O(kmnt)$ ,  $t$ 为迭代次数,  $k$ 为类别个数,  $n$ 为数据的个数,  $m$ 数据的维度。其次, K-means对数据分布有局限, 需要数据满足凸性屏, 谱聚类对数据的要求则较少。
2. 谱聚类的所有操作都是基于数据点之间的相似度矩阵 $W$ 。(相似度选择很重要)
3. 谱聚类的操作相当于将数据映射到图, 图可以用矩阵表示, 矩阵最重要的参数就是特征值和特征向量, 矩阵的特征向量类似光的光谱。
4. 根据2的性质, 其实谱聚类也类似等效于高维数据的降维。

## 参考文献:

- Von Luxburg, Ulrike. "A tutorial on spectral clustering." *Statistics and computing* 17.4 (2007): 395-416.
- 北京大学数学系几何与代数教研室前代数小组编. 高等代数. 第三版. 北京: 高等教育出版社, 2003. 9: 273-298.
- [http://en.wikipedia.org/wiki/Rayleigh\\_quotient](http://en.wikipedia.org/wiki/Rayleigh_quotient).

# 附录： 1

$$\begin{aligned} f' L f &= f' D f - f' W f = \sum_{i=1}^n d_i f_i^2 - \sum_{i,j=1}^n f_i f_j w_{ij} \\ &= \frac{1}{2} \left( \sum_{i=1}^n d_i f_i^2 - 2 \sum_{i,j=1}^n f_i f_j w_{ij} + \sum_{j=1}^n d_j f_j^2 \right) = \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2. \end{aligned}$$

$$\sum_{i=1}^n f_i = \sum_{i \in A} \sqrt{\frac{|\bar{A}|}{|A|}} - \sum_{i \in \bar{A}} \sqrt{\frac{|A|}{|\bar{A}|}} = |A| \sqrt{\frac{|\bar{A}|}{|A|}} - |\bar{A}| \sqrt{\frac{|A|}{|\bar{A}|}} = 0.$$

$$\|f\|^2 = \sum_{i=1}^n f_i^2 = |A| \frac{|\bar{A}|}{|A|} + |\bar{A}| \frac{|A|}{|\bar{A}|} = |\bar{A}| + |A| = n.$$



## 附录： 2

$$\begin{aligned} f'Lf &= \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2 \\ &= \frac{1}{2} \sum_{i \in A, j \in \bar{A}} w_{ij} \left( \sqrt{\frac{|\bar{A}|}{|A|}} + \sqrt{\frac{|A|}{|\bar{A}|}} \right)^2 + \frac{1}{2} \sum_{i \in \bar{A}, j \in A} w_{ij} \left( -\sqrt{\frac{|\bar{A}|}{|A|}} - \sqrt{\frac{|A|}{|\bar{A}|}} \right)^2 \\ &= \text{cut}(A, \bar{A}) \left( \frac{|\bar{A}|}{|A|} + \frac{|A|}{|\bar{A}|} + 2 \right) \\ &= \text{cut}(A, \bar{A}) \left( \frac{|A| + |\bar{A}|}{|A|} + \frac{|A| + |\bar{A}|}{|\bar{A}|} \right) \\ &= |V| \cdot \text{RatioCut}(A, \bar{A}). \end{aligned}$$

# 附录： 3

$$h'_i L h_i = \frac{\text{cut}(A_i, \bar{A}_i)}{|A|}$$

$$f' L f = \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2$$

$$\left(\sqrt{\frac{1}{|A|}} + 0\right)^2$$

$$= \frac{1}{2} \sum_{i \in A, j \in \bar{A}} w_{ij} \left( \sqrt{\frac{|\bar{A}|}{|A|}} + \sqrt{\frac{|A|}{|\bar{A}|}} \right)^2 + \frac{1}{2} \sum_{i \in \bar{A}, j \in A} w_{ij} \left( -\sqrt{\frac{|\bar{A}|}{|A|}} - \sqrt{\frac{|A|}{|\bar{A}|}} \right)^2$$

$$= \text{cut}(A, \bar{A}) \left( \frac{|\bar{A}|}{|A|} + \frac{|A|}{|\bar{A}|} + 2 \right)$$

$$= \text{cut}(A, \bar{A}) \left( \frac{|A| + |\bar{A}|}{|A|} + \frac{|A| + |\bar{A}|}{|\bar{A}|} \right)$$

$$= |V| \cdot \text{RatioCut}(A, \bar{A}).$$

$$\left(\sqrt{0 + \frac{1}{|A|}}\right)^2$$

# 附录： 4

$$h'_i L h_i = \frac{\text{cut}(A_i, \overline{A})}{|A_i|}$$

$$(h_1^t, h_2^t, \dots, h_i^t, \dots, h_k^t) L (h_1, h_2, \dots, h_i, \dots, h_k) = \begin{matrix} h_1^t L h_1 & h_1^t L h_i & \dots \\ \dots & h_i^t L h_i & \dots \\ \dots & \dots & h_k^t L h_k \end{matrix}$$

$$f' L f = \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2$$

$$= \frac{1}{2} \sum_{i \in A, j \in \overline{A}} w_{ij} \left( \sqrt{\frac{|\overline{A}|}{|A|}} + \sqrt{\frac{|A|}{|\overline{A}|}} \right)^2 + \frac{1}{2} \sum_{i \in \overline{A}, j \in A} w_{ij} \left( -\sqrt{\frac{|\overline{A}|}{|A|}} - \sqrt{\frac{|A|}{|\overline{A}|}} \right)^2$$

$$= \text{cut}(A, \overline{A}) \left( \frac{|\overline{A}|}{|A|} + \frac{|A|}{|\overline{A}|} + 2 \right)$$

$$= \text{cut}(A, \overline{A}) \left( \frac{|A| + |\overline{A}|}{|A|} + \frac{|A| + |\overline{A}|}{|\overline{A}|} \right)$$

$$= |V| \cdot \text{RatioCut}(A, \overline{A}).$$

$$h_{i,j} = \begin{cases} 1/\sqrt{\text{vol}(A_j)} & \text{if } v_i \in A_j \\ 0 & \text{otherwise} \end{cases}$$

因此最小化 $\text{Ratio}(A_1, A_2, \dots, A_k)$ 问题转化为:

$$\min_{A_1, \dots, A_k} \text{Tr}(H' L H)$$



$$\min_{T \in \mathbb{R}^{n \times k}} \text{Tr}(T' D^{-1/2} L D^{-1/2} T)$$

其中:  $H' D H = I,$



$$T = D^{1/2} H$$

其中:  $T' T = I.$

这是标准的最小化矩阵迹问题, 根据Rayleigh—Ritz theorem另一个定理: 通过选取L矩阵的前K个最小特征值对应的特征向量以列的方式组成H矩阵即为该问题的解。然后, 对H的行用k-means聚类 聚类数为k;

(P19)