聚类方法

3月机器学习在线班 邹博

2015年3月28日

预备问题

- 已知$x_1,x_2\ldots x_k$为正数，且 $x_1+x_2+\cdots+x_k=\alpha$
- 其中，α为定值；
- 求：$f(x_1,x_2\cdots x_k)=\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_k}$ 的最小值

预备问题的求解

- 显然，这是一个仅带等式约束的优化问题
- 根据 $f(x_1, x_2 \cdots x_k) = \frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_k}$
- 令 $g(x_1, x_2 \cdots x_k, \lambda) = \frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_k} + \lambda(x_1 + x_2 + \cdots + x_k - \alpha)$
- 求驻点

$$\frac{\partial g}{\partial x_i} = -\frac{1}{x_i^2} + \lambda \overset{\Delta}{=} 0 \Rightarrow x_i = \sqrt{\lambda}$$

- 带入约束条件 $x_1 = x_2 = \cdots = x_k = \frac{\alpha}{k}$
- 从而 $\min f(x_1, x_2 \cdots x_k) = \min\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_k}\right) = \frac{k^2}{\alpha}$

复习：均方误差准则 $MSE(\hat{\theta})=E\left[(\hat{\theta}-\theta)^2\right]$

- 用估计量 $\hat{\theta}$ 去估计θ，其误差是 $\hat{\theta}-\theta$，该误差显然随样本 $X_1,X_2,\cdots X_n$ 而定，因此，$\hat{\theta}-\theta$ 是随机变量，它的平方的均值，称作均方误差。这个量越小，平均误差越小，估计结果越优。

$$MSE(\hat{\theta})=E\left[(\hat{\theta}-\theta)^2\right]$$

- 显然，若 $\hat{\theta}$ 是无偏估计，则MSE即方差。

$$MSE(\hat{\theta})=E\left[(\hat{\theta}-\theta)^2\right]=E\left[(\hat{\theta}-E(\hat{\theta}))^2\right]=Var(\hat{\theta})$$

本次目标

- 掌握K-means聚类的思路和使用条件
- 了解层次聚类的思路和方法
- 理解密度聚类并能够应用于实践
  - DBSCAN
  - 密度最大值聚类
- 掌握谱聚类的算法，初步理解谱聚类的内涵

聚类的定义

- 聚类就是对大量未知标注的数据集，按数据的内在相似性将数据集划分为多个类别，使类别内的数据相似度较大而类别间的数据相似度较小
  - 无监督

相似度/距离计算方法总结

- 闵可夫斯基距离Minkowski/欧式距离

$$dist(X,Y)=\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{\frac{1}{p}}$$

- 杰卡德相似系数(Jaccard)

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}$$

- 余弦相似度(cosine similarity)

$$\cos(\theta)=\frac{a^Tb}{|a|\cdot|b|}$$

- Pearson相似系数

$$\rho_{XY}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}=\frac{\sum_{i=1}^{n}(X_i-\mu_X)(Y_i-\mu_Y)}{\sqrt{\sum_{i=1}^{n}(X_i-\mu_X)^2}\sqrt{\sum_{i=1}^{n}(Y_i-\mu_Y)^2}}$$

- 相对熵(K-L距离)

$$D(p\,\|\,q)=\sum_x p(x)\log\frac{p(x)}{q(x)}=E_{p(x)}\log\frac{p(x)}{q(x)}$$

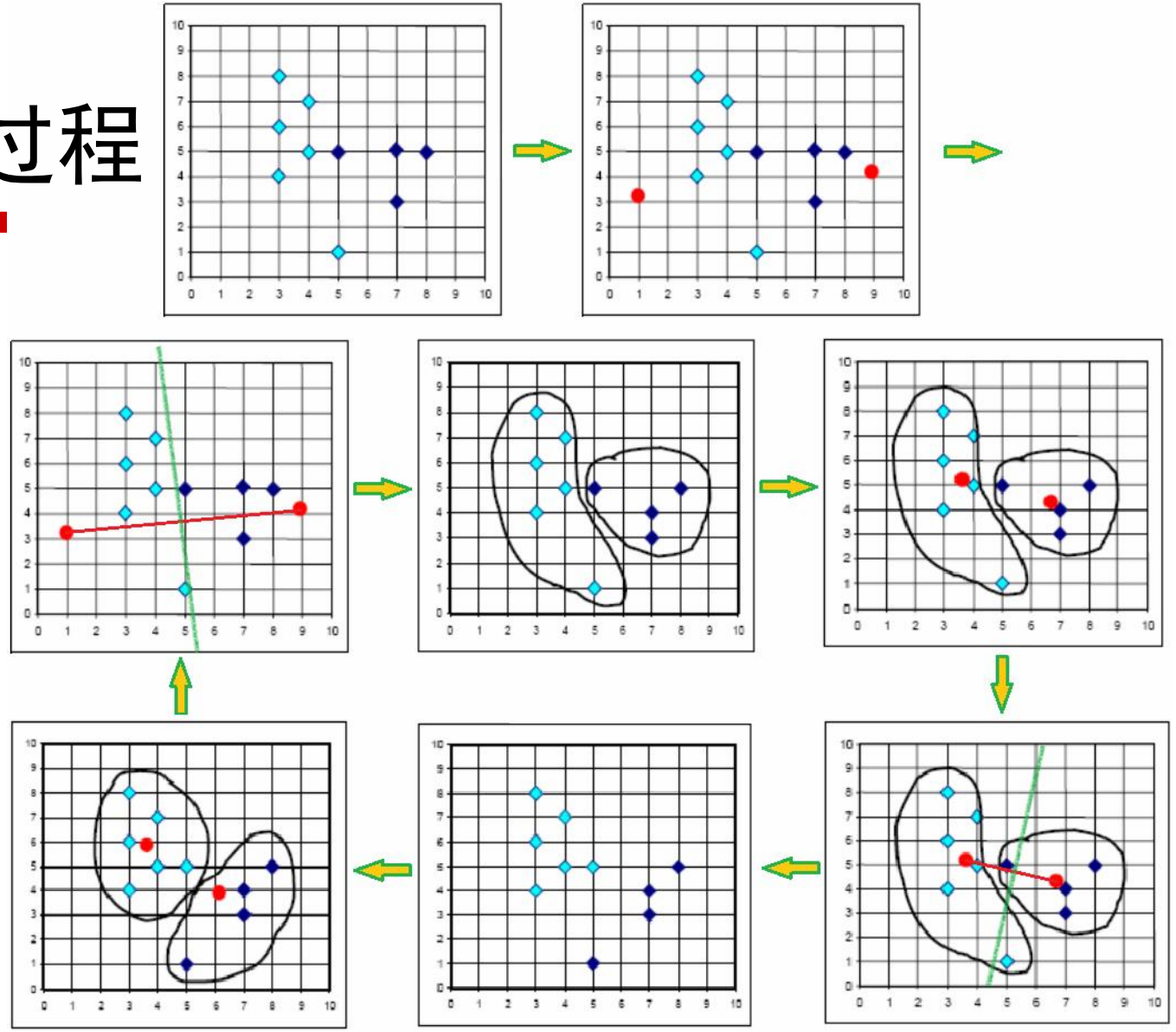
聚类的基本思想

- 给定一个有N个对象的数据集，划分聚类技术将构造数据的k个划分，每一个划分代表一个簇，k≤n。也就是说，聚类将数据划分为k个簇，而且这k个划分满足下列条件：
  - 每一个簇至少包含一个对象
  - 每一个对象属于且仅属于一个簇
- 基本思想：对于给定的k，算法首先给出一个初始的划分方法，以后通过反复迭代的方法改变划分，使得每一次改进之后的划分方案都较前一次更好。

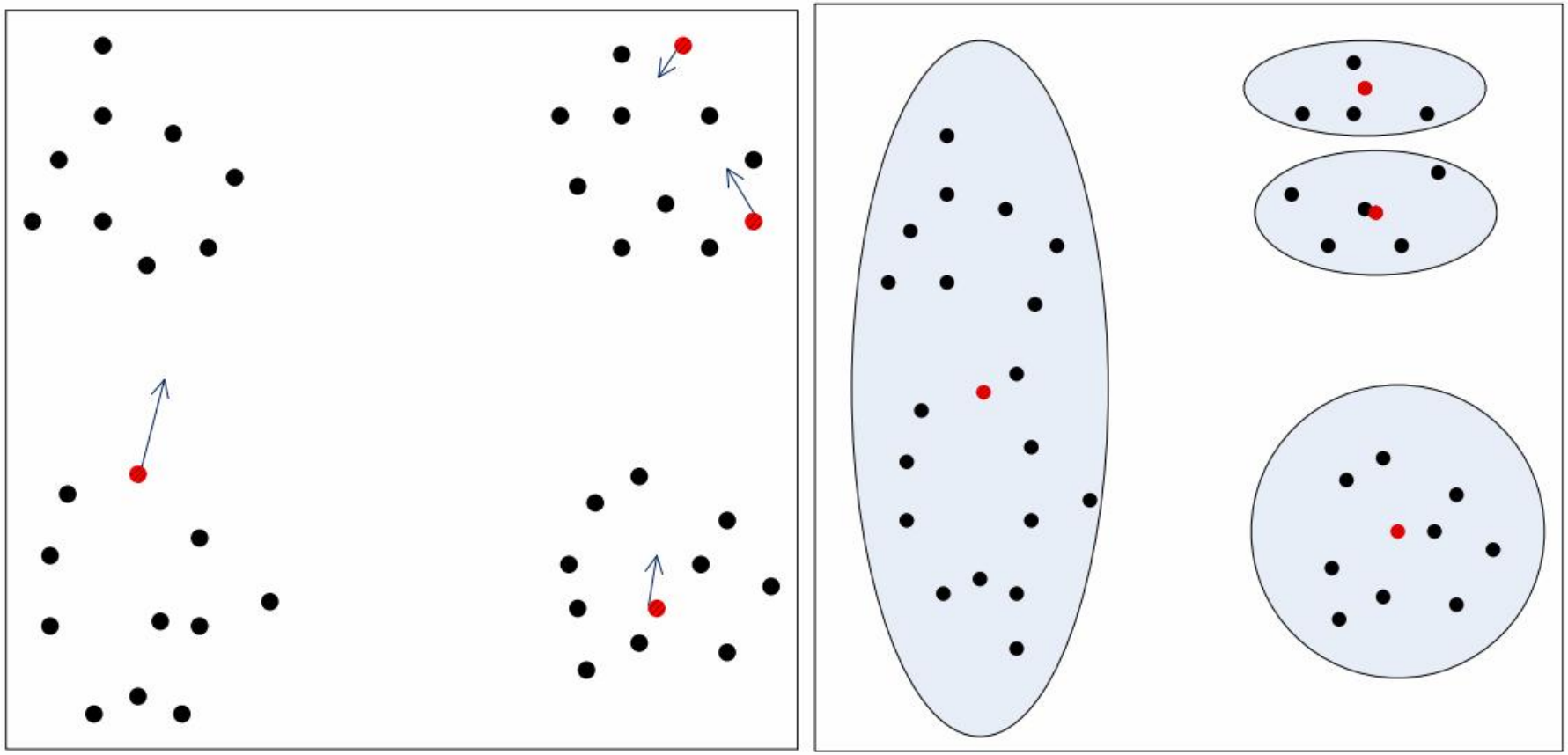
K-means算法

- K-means算法，也被称为k-平均或k-均值，是一种得到最广泛使用的聚类算法，或者成为其他聚类算法的基础。
- 算法首先随机地选择k个对象，每个对象初始地代表了一个簇的平均值或中心。对剩余的每个对象根据其与各个簇中心的距离，将它赋给最近的簇。然后重新计算每个簇的平均值。这个过程不断重复，直到准则函数收敛。
  - 准则函数常常使用最小平方误差MSE
  - Minimum Squared-Error

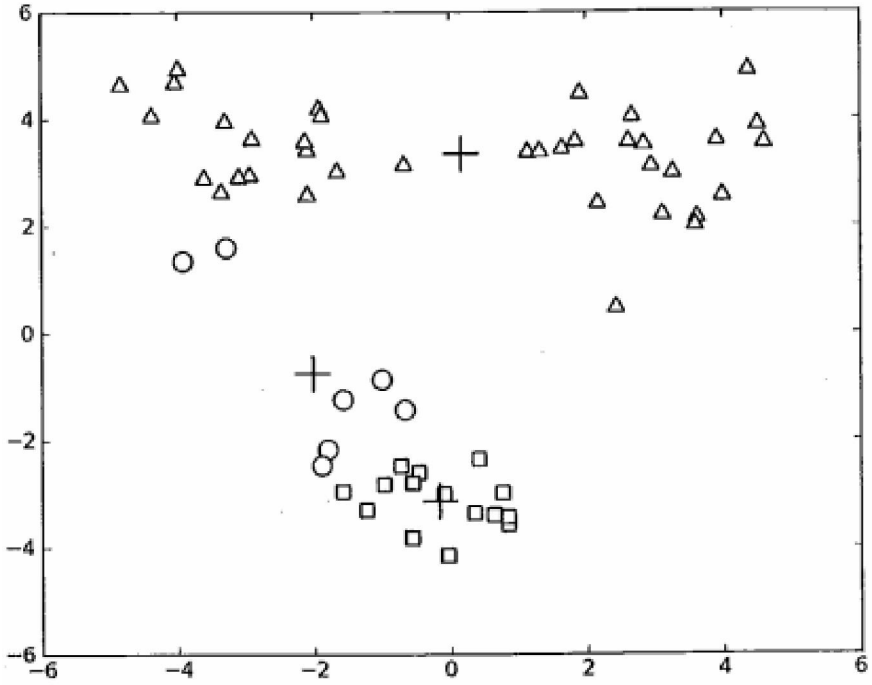
K-means过程

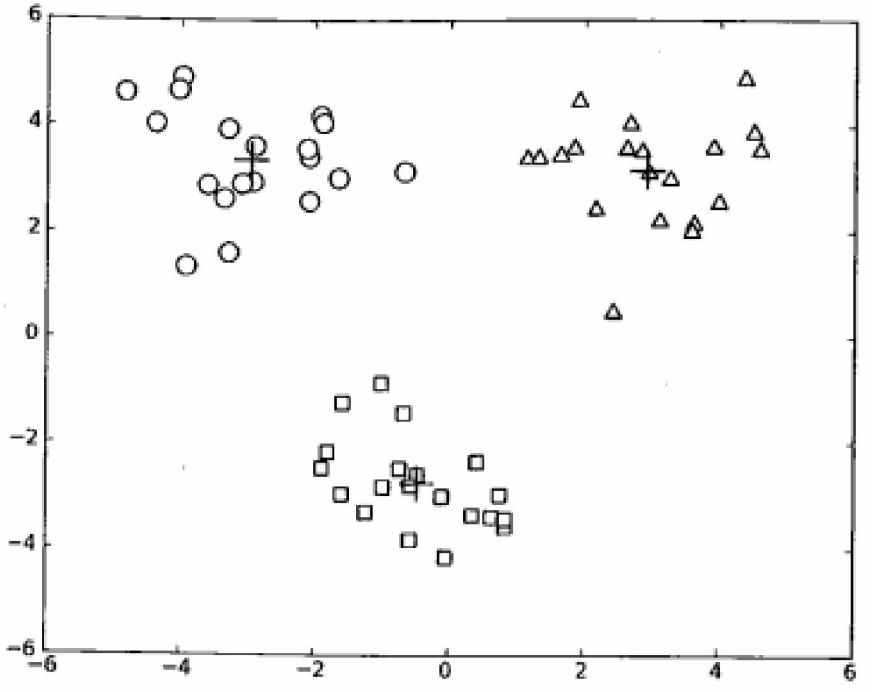


K-means是初值敏感的



二分k-均值聚类





K-means聚类方法总结

- 优点：
  - 是解决聚类问题的一种经典算法，简单、快速
  - 对处理大数据集，该算法保持可伸缩性和高效率
  - 当结果簇是密集的，它的效果较好
- 缺点
  - 在簇的平均值可被定义的情况下才能使用，可能不适用于某些应用
  - 必须事先给出k（要生成的簇的数目），而且对初值敏感，对于不同的初始值，可能会导致不同结果。
  - 不适合于发现非凸形状的簇或者大小差别很大的簇
  - 对躁声和孤立点数据敏感
- 可作为其他聚类方法的基础算法，如谱聚类

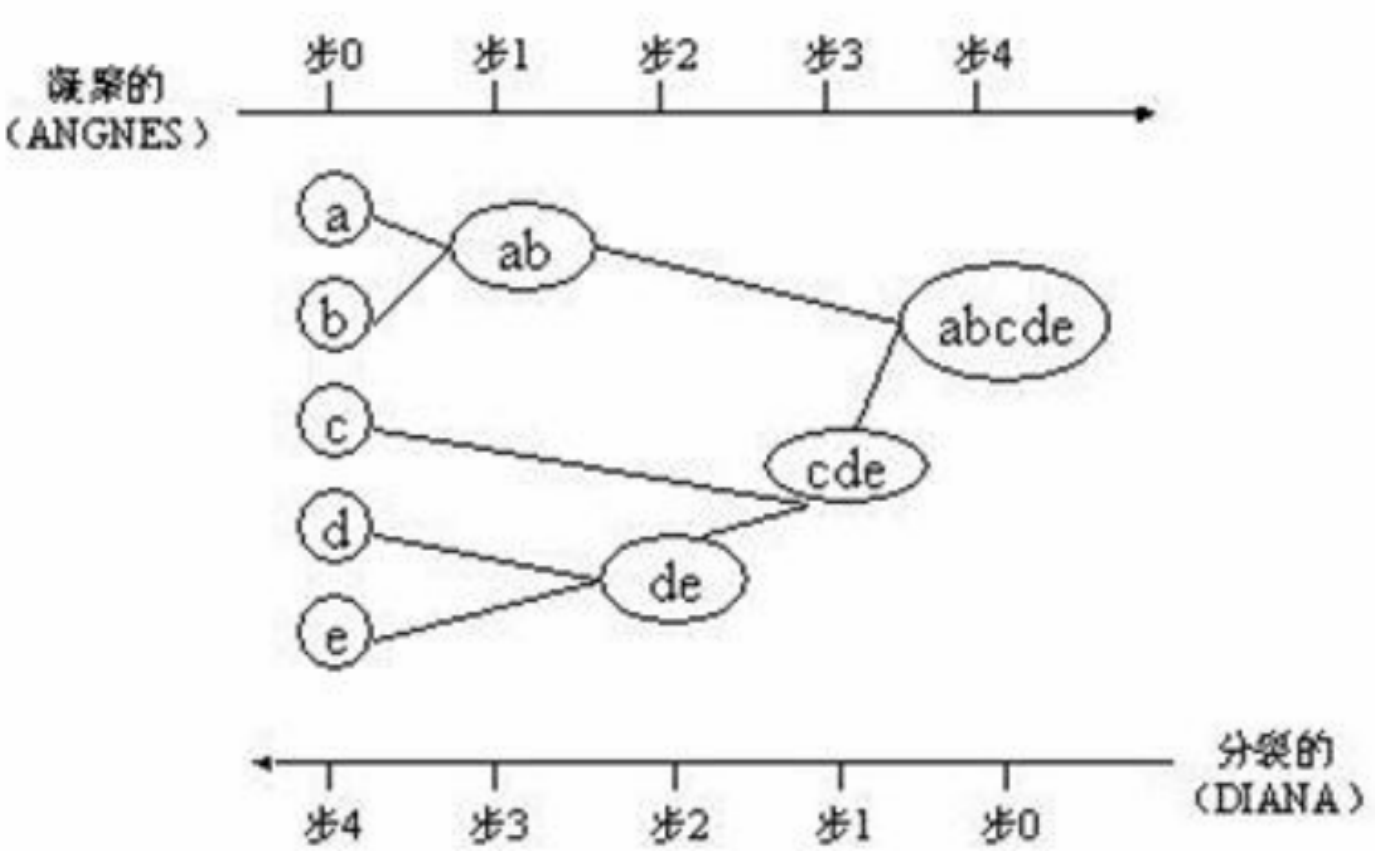
层次聚类方法

- 层次聚类方法对给定的数据集进行层次的分解，直到某种条件满足为止。具体又可分为：
- 凝聚的层次聚类：AGNES算法
  - 一种自底向上的策略，首先将每个对象作为一个簇，然后合并这些原子簇为越来越大的簇，直到某个终结条件被满足。
- 分裂的层次聚类：DIANA算法
  - 采用自顶向下的策略，它首先将所有对象置于一个簇中，然后逐渐细分为越来越小的簇，直到达到了某个终结条件。

AGNES和DIANA算法

- AGNES (AGglomerative NESting)算法最初将每个对象作为一个簇，然后这些簇根据某些准则被一步步地合并。两个簇间的距离由这两个不同簇中距离最近的数据点对的相似度来确定；聚类的合并过程反复进行直到所有的对象最终满足簇数目。
- DIANA (DIvisive ANAlysis)算法是上述过程的反过程，属于分裂的层次聚类，首先将所有的对象初始化到一个簇中，然后根据一些原则（比如最大的欧式距离），将该簇分类。直到到达用户指定的簇数目或者两个簇之间的距离超过了某个阈值。

层次聚类


凝聚的
（ANGNES）
步0
步1
步2
步3
步4
a
b
c
d
e
ab
de
cde
abcde
分裂的
（DIANA）
步4
步3
步2
步1
步0

密度聚类方法

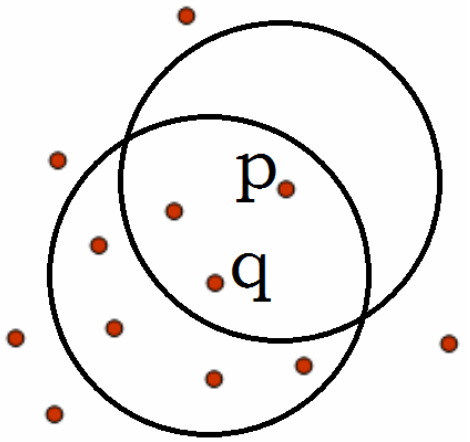
- 密度聚类方法的指导思想是，只要一个区域中的点的密度大于某个阈值，就把它加到与之相近的聚类中去。这类算法能克服基于距离的算法只能发现“类圆形”(凸)的聚类的缺点，可发现任意形状的聚类，且对噪声数据不敏感。但计算密度单元的计算复杂度大，需要建立空间索引来降低计算量。
  - DBSCAN
  - 密度最大值算法

DBSCAN算法

- DBSCAN(Density-Based Spatial Clustering of Applications with Noise)
- 一个比较有代表性的基于密度的聚类算法。与划分和层次聚类方法不同，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在有“噪声”的数据中发现任意形状的聚类。

DBSCAN算法的若干概念

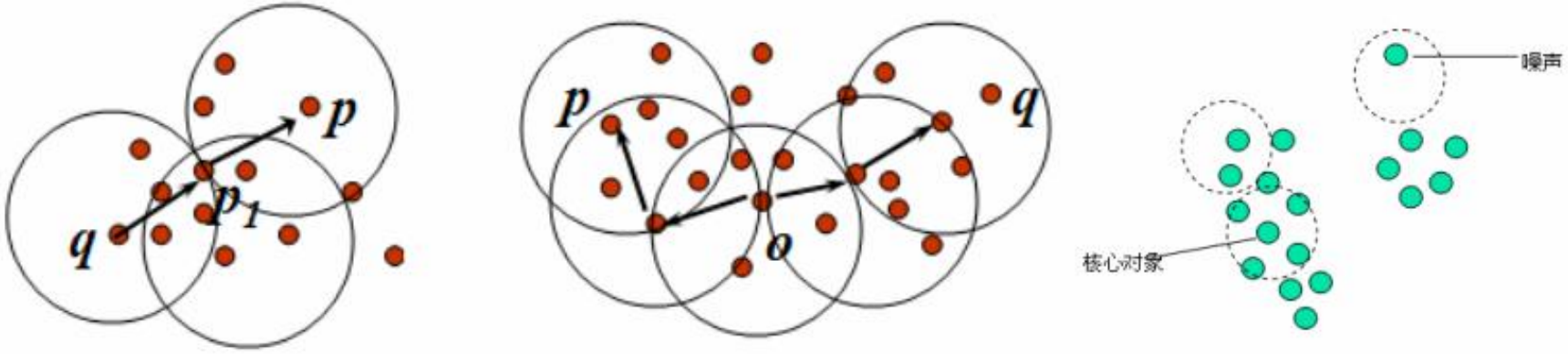
- 对象的 ε-邻域：给定对象在半径 ε 内的区域。
- 核心对象：对于给定的数目m，如果一个对象的 ε-邻域至少包含m个对象，则称该对象为核心对象。
- 直接密度可达：给定一个对象集合D，如果p是在q的 ε-邻域内，而q是一个核心对象，我们说对象p从对象q出发是直接密度可达的。



- 如图 ε=1cm，m=5，q是一个核心对象，从对象q出发到对象p是直接密度可达的。

DBSCAN算法的若干概念

- 密度可达：如果存在一个对象链$p_1p_2...p_n$，$p_1=q$，$p_n=p$，对$p_i \in D$，$(1 \leqslant i \leqslant n)$，$p_{i+1}$是从$p_i$关于 ε 和m直接密度可达的，则对象p是从对象q关于 ε 和m密度可达的。
- 密度相连：如果对象集合D中存在一个对象o，使得对象p和q是从o关于 ε 和m密度可达的，那么对象p和q是关于 ε 和m密度相连的。
- 簇：一个基于密度的簇是最大的密度相连对象的集合。
- 噪声：不包含在任何簇中的对象称为噪声。



DBSCAN算法

- DBSCAN通过检查数据集中每个对象的ε-邻域来寻找聚类。
- 如果一个点p的ε-邻域包含多于m个对象，则创建一个p作为核心对象的新簇。然后，DBSCAN反复地寻找从这些核心对象直接密度可达的对象，这个过程可能涉及密度可达簇的合并。当没有新的点可以被添加到任何簇时，该过程结束。

下面给出一个样本事务数据库（见左表），对它实施DBSCAN算法。
以下为算法的步骤（设n=12，用户输入ε =1，MinPts=4）

样本事务数据库

序号	属性 1	属性 2
1	1	0
2	4	0
3	0	1
4	1	1
5	2	1
6	3	1
7	4	1
8	5	1
9	0	2
10	1	2
11	4	2
12	1	3

DBSCAN算法执行过程示意

步骤	选择的点	在ε中点的个数	通过计算可达点而找到的新簇
1	1	2	无
2	2	2	无
3	3	3	无
4	4	5	簇C_1：{1，3，4，5，9，10}
5	5	3	已在一个簇C_1中
6	6	3	无
7	7	5	簇C_2：{2，6，7，8，11}
8	8	2	已在一个簇C_2中
9	9	3	已在一个簇C_1中
10	10	4	簇C_1：{1，3，4，5，9，10，12}
11	11	2	已在一个簇C_2中
12	12	2	已在一个簇C_1中

聚出的类为{1，3，4，5，9，11，12}，{2，6，7，8，10}。

密度最大值聚类

- 密度最大值聚类是一种简洁优美的聚类算法，可以识别各种形状的类簇，并且参数很容易确定。
- 定义：局部密度ρi

$$\rho_i = \sum_j \chi(d_{ij} - d_c), \quad 其中，\chi(x) = \begin{cases} 1 & x < 0 \\ 0 & otherwise \end{cases}$$

- dc是一个截断距离，ρi即到点i的距离小于dc的点的个数。由于该算法只对ρi的相对值敏感，

密度最大值聚类

- 密度最大值聚类是一种简洁优美的聚类算法，可以识别各种形状的类簇，并且参数很容易确定。
- 定义：局部密度 ρi

$$\rho_i = \sum_j \chi(d_{ij} - d_c), \quad \text{其中，} \chi(x) = \begin{cases} 1 & x < 0 \\ 0 & otherwise \end{cases}$$

  - d_c是一个截断距离，ρ_i即到对象i的距离小于d_c的对象的个数。由于该算法只对ρ_i的相对值敏感，所以对d_c的选择是稳健的，一种推荐做法是选择d_c，使得平均每个点的邻居数为所有点的1%-2%
- 定义：高局部密度点距离 δi
  - 简称“高密距离”（注：该称呼不具代表性）。

高局部密度点距离

- 高局部密度点距离 $\delta_i = \min_{j:\rho_j > \rho_i} \left(d_{ij} \right)$

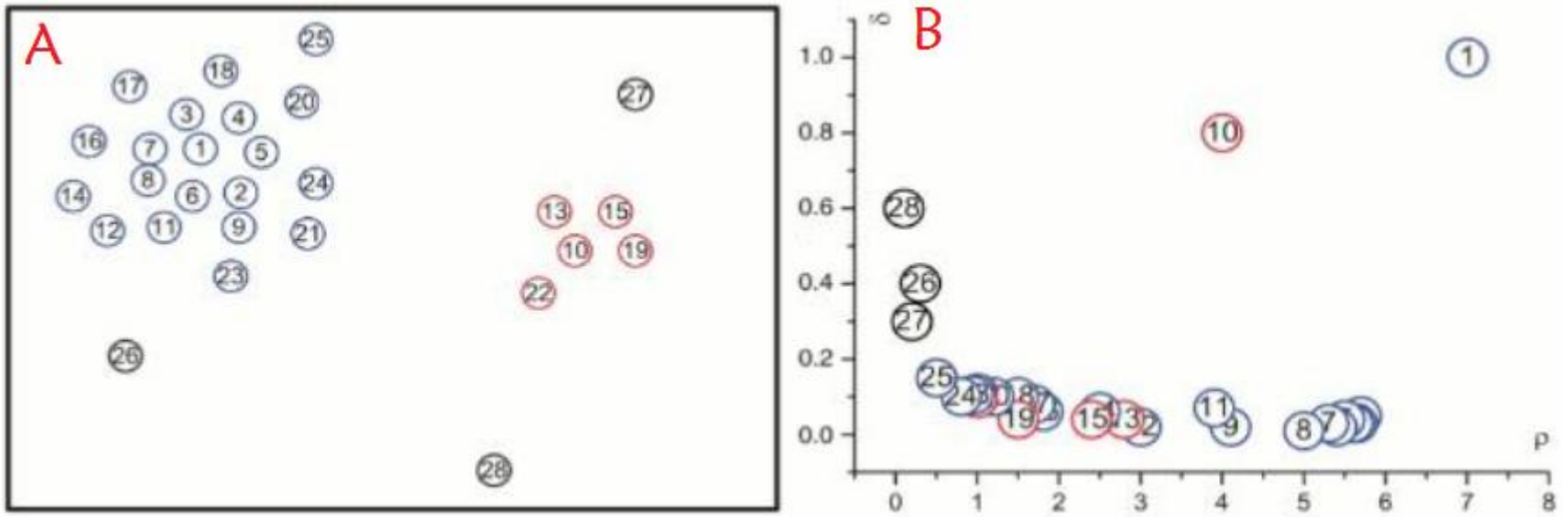
- 在密度高于对象i的所有对象中，到对象i最近的距离，即高局部密度点距离。
- 对于密度最大的对象，设置 $\delta_i = \max(d_{ij})$（即：该问题中的无穷大）。
  - 只有那些密度是局部或者全局最大的点才会有远大于正常值的高局部密度点距离。

簇中心的识别

- 那些有着比较大的局部密度ρ_i和很大的高密距离δ_i的点被认为是簇的中心；高密距离δ_i较大但局部密度ρi较小的点是异常点；
  - 确定簇中心之后，其他点按照距离已知簇的中心最近进行分类
    - 注：有些教科书如上表述。但实践中按照密度可达的方法进行分类跟为稳妥。

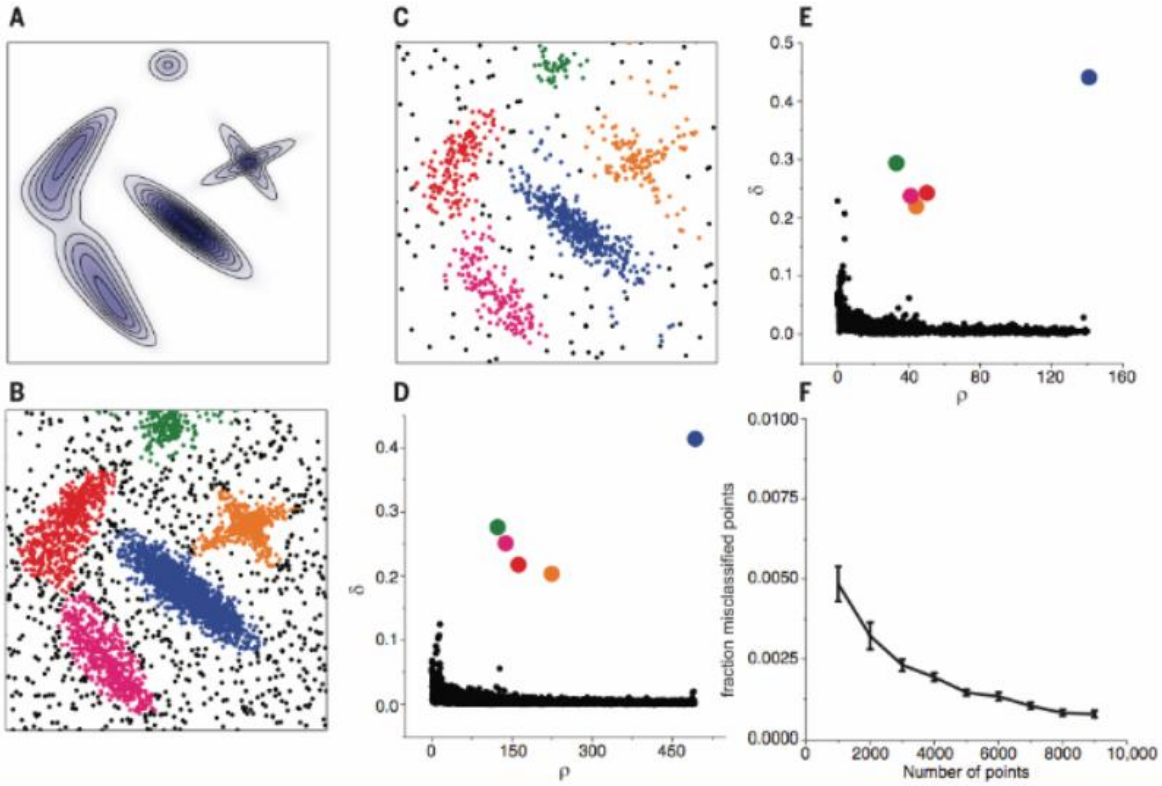
密度最大值聚类过程

- 左图是所有点在二维空间的分布，右图是以ρ为横坐标，以δ为纵坐标绘制的决策图。可以看到，1和10两个点的ρi和δi都比较大，作为簇的中心点。26、27、28三个点的δi也比较大，但是ρi较小，所以是异常点。



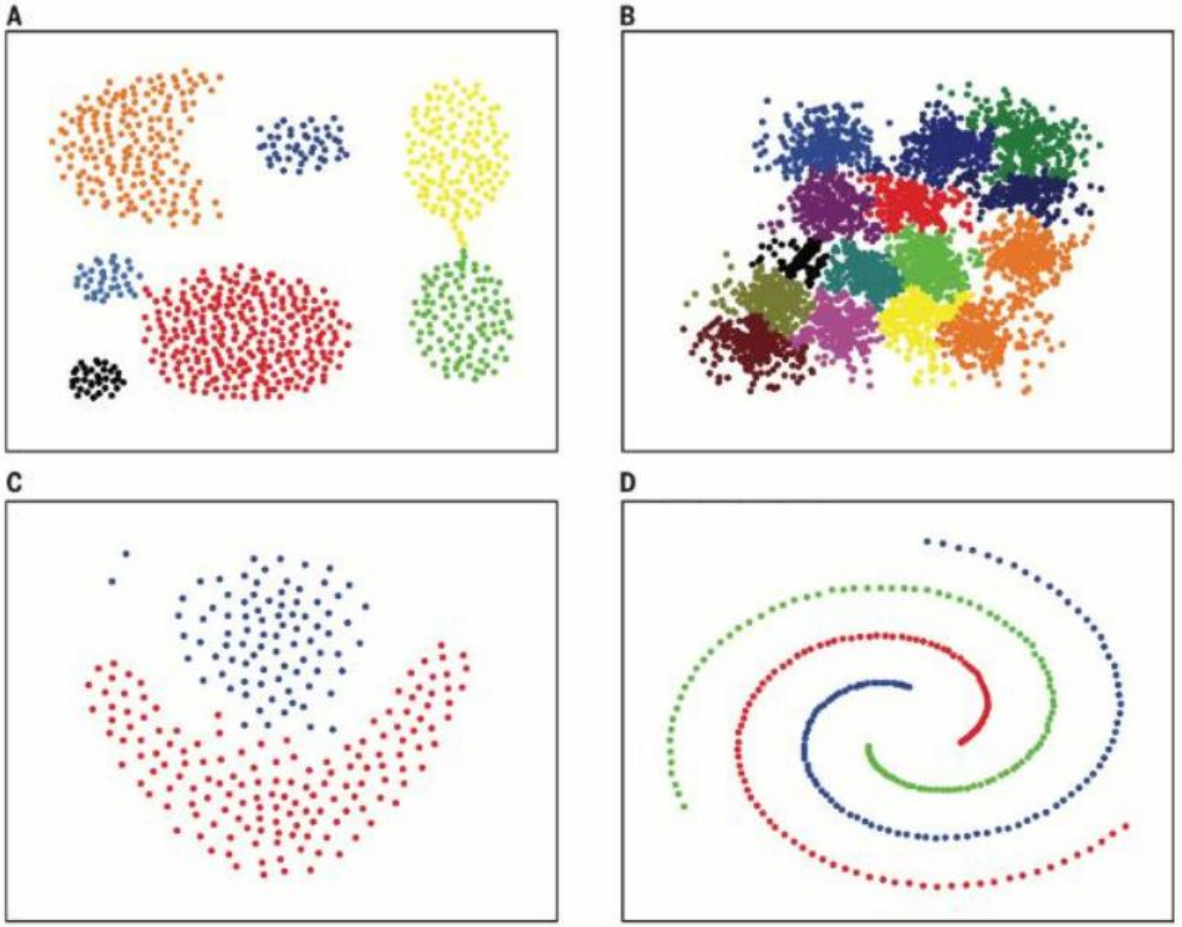
可靠性：对边界和噪声的认识

- 在聚类分析中，通常需要确定每个点划分给某个簇的可靠性；
- 在该算法中，可以首先为每个簇定义一个边界区域(border region)，亦即划分给该簇但是距离其他簇的点的距离小于d_c的点的集合。然后为每个簇找到其边界区域的局部密度最大的点，令其局部密度为ρ_h。
- 该簇中所有局部密度大于ρ_h的点被认为是簇核心的一部分(亦即将该点划分给该类簇的可靠性很大)，其余的点被认为是该类簇的光晕(halo)，亦即可以认为是噪声。
  - 注：关于可靠性问题，在EM算法中仍然会有相关涉及。



A图为生成数据的概率分布, B, C二图为分别从该分布中生成了4000, 1000个点. D, E分别是B, C两组数据的决策图(decision tree), 可以看到两组数据都只有五个点有比较大的ρ i 和很大的δ i . 这些点作为类簇的中心, 在确定了类簇的中心之后, 每个点被划分到各个类簇(彩色点), 或者是划分到类簇光晕(黑色点). F图展示的是随着抽样点数量的增多, 聚类的错误率在逐渐下降, 说明该算法是鲁棒的.

不同数据下密度最大值聚类的效果



复习1：实对称阵的特征值是实数

- 首先 $A\bar{x} = \bar{A}\bar{x} = \overline{Ax} = \overline{\lambda x} = \bar{\lambda}\bar{x}$
- 因为 $\bar{x}^T(Ax) = \bar{x}^T(Ax) = \bar{x}^T\lambda x = \lambda\bar{x}^T x$

  $\bar{x}^T(Ax) = (\bar{x}^T A^T)x = (A\bar{x})^T x = (\bar{\lambda}\bar{x})^T x = \bar{\lambda}\bar{x}^T x$
- 从而

  $$\lambda\bar{x}^T x = \bar{\lambda}\bar{x}^T x \Rightarrow (\lambda - \bar{\lambda})\bar{x}^T x = 0$$
- 而

  $$\bar{x}^T x = \sum_{i=1}^{n} \bar{x}_i x_i = \sum_{i=1}^{n} |x_i|^2 \neq 0$$
- 所以

  $$\lambda - \bar{\lambda} = 0 \Rightarrow \lambda = \bar{\lambda}$$

复习2:实对称阵不同特征值对应的特征向量正交

- 令实对称矩阵为A，它的两个不同的特征值λ_1 λ_2对应的特征向量分别是μ_1 μ_2；其中，λ_1 λ_2 μ_1 μ_2都是实数或是实向量。
- 则有：$A\mu_1=\lambda_1\mu_1$，$A\mu_2=\lambda_2\mu_2$
- $(A\mu_1)^T=(\lambda_1\mu_1)^T$，从而：$\mu_1^TA=\lambda_1\mu_1^T$
- 所以：$\mu_1^TA\mu_2=\lambda_1\mu_1^T\mu_2$
- 同时，$\mu_1^TA\mu_2=\mu_1^T(A\mu_2)=\mu_1^T\lambda_2\mu_2=\lambda_2\mu_1^T\mu_2$
- 所以，$\lambda_1\mu_1^T\mu_2=\lambda_2\mu_1^T\mu_2$
- 故：$(\lambda_1-\lambda_2)\mu_1^T\mu_2=0$
- 而$\lambda_1\neq\lambda_2$，所以$\mu_1^T\mu_2=0$，即：μ_1，μ_2正交。

谱和谱聚类

- 方阵作为线性算子，它的所有特征值的全体统称方阵的**谱**。
  - 方阵的谱半径为最大的特征值
  - 矩阵A的**谱半径**：(A^TA)的最大特征值
- 谱聚类：一般的说，是一种基于图论的聚类方法，通过对样本数据的**拉普拉斯矩阵**的**特征向量**进行聚类，从而达到对样本数据聚类的目的。

谱分析的整体过程

- 给定一组数据$x_1,x_2,...x_n$，记任意两个点之间的相似度(“距离”的减函数)为$s_{ij}=<xi,xj>$，形成相似度图(similarity graph)：G=(V,E)。如果x_i和x_j之间的相似度s_{ij}大于一定的阈值，那么，两个点是连接的，权值记做s_{ij}。
- 接下来，可以用相似度图来解决样本数据的聚类问题：找到图的一个划分，形成若干个组(Group)，使得不同组之间有较低的权值，组内有较高的权值。

若干概念

- 无向图G=(V,E)
- 邻接矩阵$W = (w_{ij})_{i,j=1,\ldots,n}$
- 顶点的度di → 度矩阵D(对角阵)

$$d_i = \sum_{j=1}^{n} w_{ij}$$

若干概念

- 子图A的指示向量

$$\mathbb{1}_A = (f_1, \ldots, f_n)' \in \mathbb{R}^n$$
$$f_i = 1 \text{ if } v_i \in A$$
$$f_i = 0 \text{ otherwise}$$

- A和B是图G的不相交子图，则定义子图的连接权：

$$W(A, B) := \sum_{i \in A, j \in B} w_{ij}$$

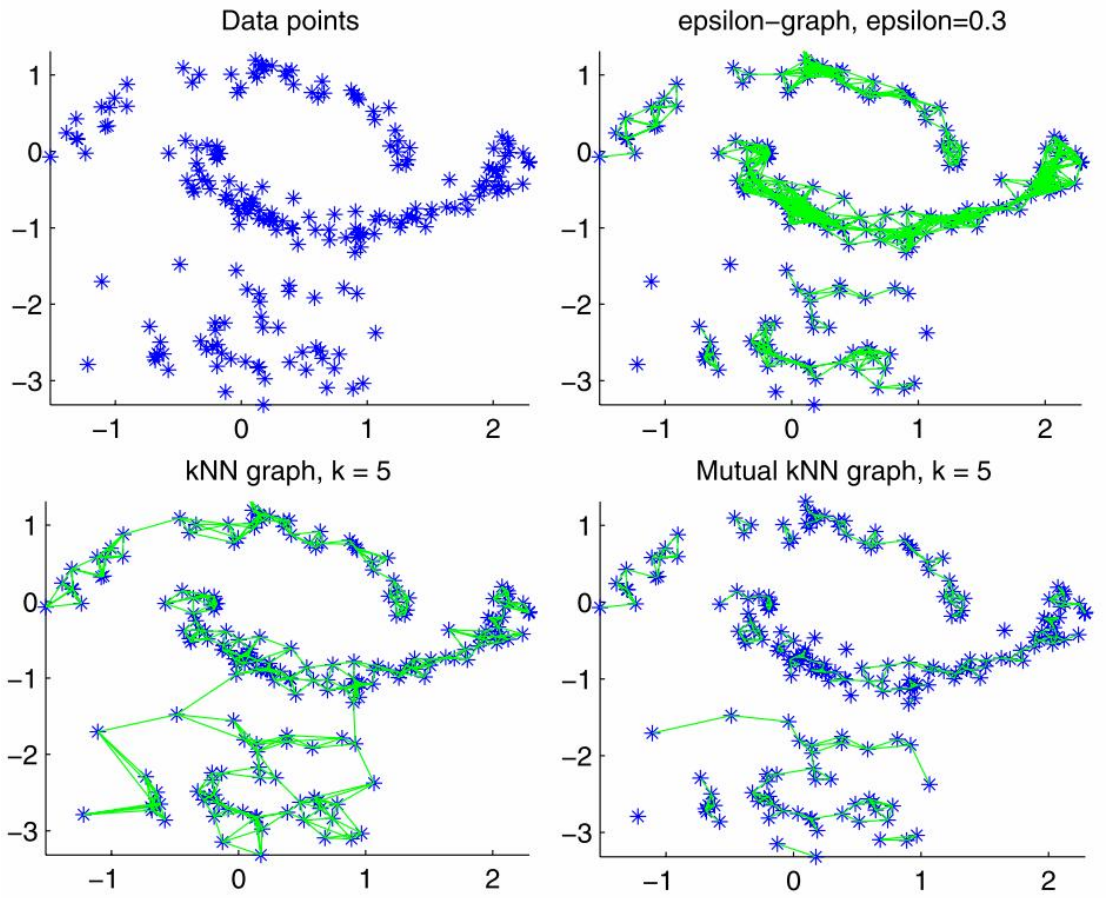
相似度图G的建立方法

- 全连接图
  - 高斯相似度函数：距离越大，相似度越小

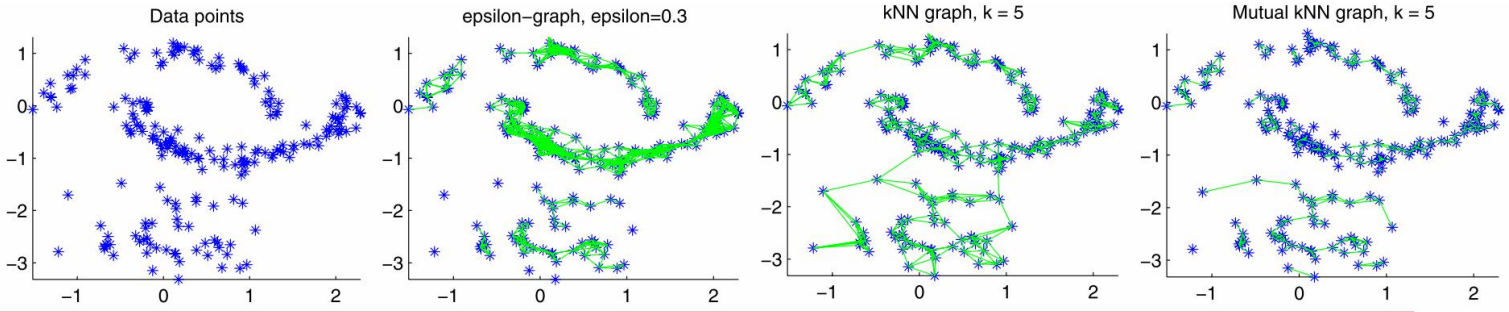
$$s(x_i, x_j) = \mathrm{e}^{-\|x_i - x_j\|^2/(2\sigma^2)}$$

- ε近邻图
  - 给定参数ε
  - 思考：如何选择ε？
    - 图G的权值的均值
    - 图G的最小生成树的最大边
- k近邻图(k-nearest neighbor graph)
  - 若vi的k最近邻包含vj，vj的k最近邻不一定包含vi：有向图
  - 忽略方向的图，往往简称“k近邻图”
  - 两者都满足才连接的图，称作“互k近邻图(mutual)”

相似度图G的举例


Data points
epsilon−graph, epsilon=0.3
kNN graph, k = 5
Mutual kNN graph, k = 5

权值比较



- ε近邻图：ε=0.3，“月牙部分”非常紧的连接了，但“高斯部分”很多都没连接。当数据有不同的“密度”时，往往发生这种问题。
- k近邻图：可以解决数据存在不同密度时有些无法连接的问题，甚至低密度的“高斯部分”与高密度的“月牙部分”也能够连接。同时，虽然两个“月牙部分”的距离比较近，但k近邻还可以把它们区分开。
- 互k近邻图：它趋向于连接相同密度的部分，而不连接不同密度的部分。这种性质介于ε近邻图和k近邻图之间。如果需要聚类不同的密度，这个性质非常有用。
- 全连接图：使用高斯相似度函数可以很好的建立权值矩阵。但缺点是建立的矩阵不是稀疏的。
- 总结：首先尝试使用k近邻图。

拉普拉斯矩阵及其性质

- 拉普拉斯矩阵：$L = D - W$

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

- L是对称半正定矩阵；
- L的最小特征值是0，相应的特征向量是 $\mathbb{1}$
- L有n个非负实特征值 $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$

拉普拉斯矩阵的性质

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

拉普拉斯矩阵的性质

- 定理：令G是权值非负的无向图，拉普拉斯矩阵L的特征值0的重数k等于图G的连通分量数。记G的连通分量为$A_1,A_2,...A_k$，则特征值0的特征向量由下列指示向量确定。

$$\mathbb{1}_{A_1},\cdots,\mathbb{1}_{A_k}$$

正则拉普拉斯矩阵的定义

- symmetric

$$L_{\text{sym}} := D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$$

- Random walk

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

正则拉普拉斯矩阵的性质

- (λ, u)是L_{rw}的特征值和特征向量，当且仅当$(\lambda, D^{1/2}u)$是L_{sym}的特征值和特征向量；
- $(0, \mathbf{1})$是L_{rw}的特征值和特征向量，$(0, D^{1/2}\mathbf{1})$是L_{sym}的特征值和特征向量；
- L_{sym}和L_{rw}是半正定的，有n个非负实特征值

$$f' L_{sym} f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2 , \quad f \in \mathbb{R}^n$$

正则拉普拉斯矩阵的性质

- 定理：令G是权值非负的无向图，正则拉普拉斯矩阵L_{sym}和L_{rw}的特征值0的重数k等于图G的连通分量数。记G的连通分量为$A_1,A_2...A_k$，则特征值0的特征向量由下列指示向量确定。

$$L_{\text{rw}} \quad \mathbb{1}_{A_1}, \ldots, \mathbb{1}_{A_k}$$

$$L_{\text{sym}} \quad D^{1/2}\mathbb{1}_{A_1}, \ldots, D^{1/2}\mathbb{1}_{A_k}$$

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 Sect. 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 $A_i = \{j | y_j \in C_i\}$.

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 Sect. 2. Let W be its weighted adjacency matrix.
- Compute the unnormalized Laplacian L.
- **Compute the first k generalized eigenvectors $u_1, \ldots, u_k$ of the generalized eigenproblem $Lu = \lambda Du$.**
- Let $U \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $u_1, \ldots, u_k$ 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 $A_i = \{j | y_j \in C_i\}$.

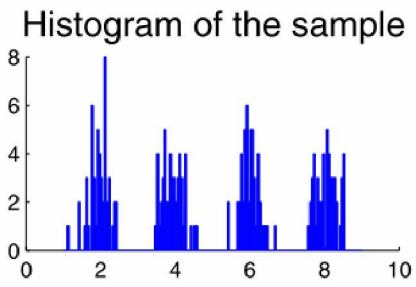
Normalized spectral clustering according to Ng et al. (2002)

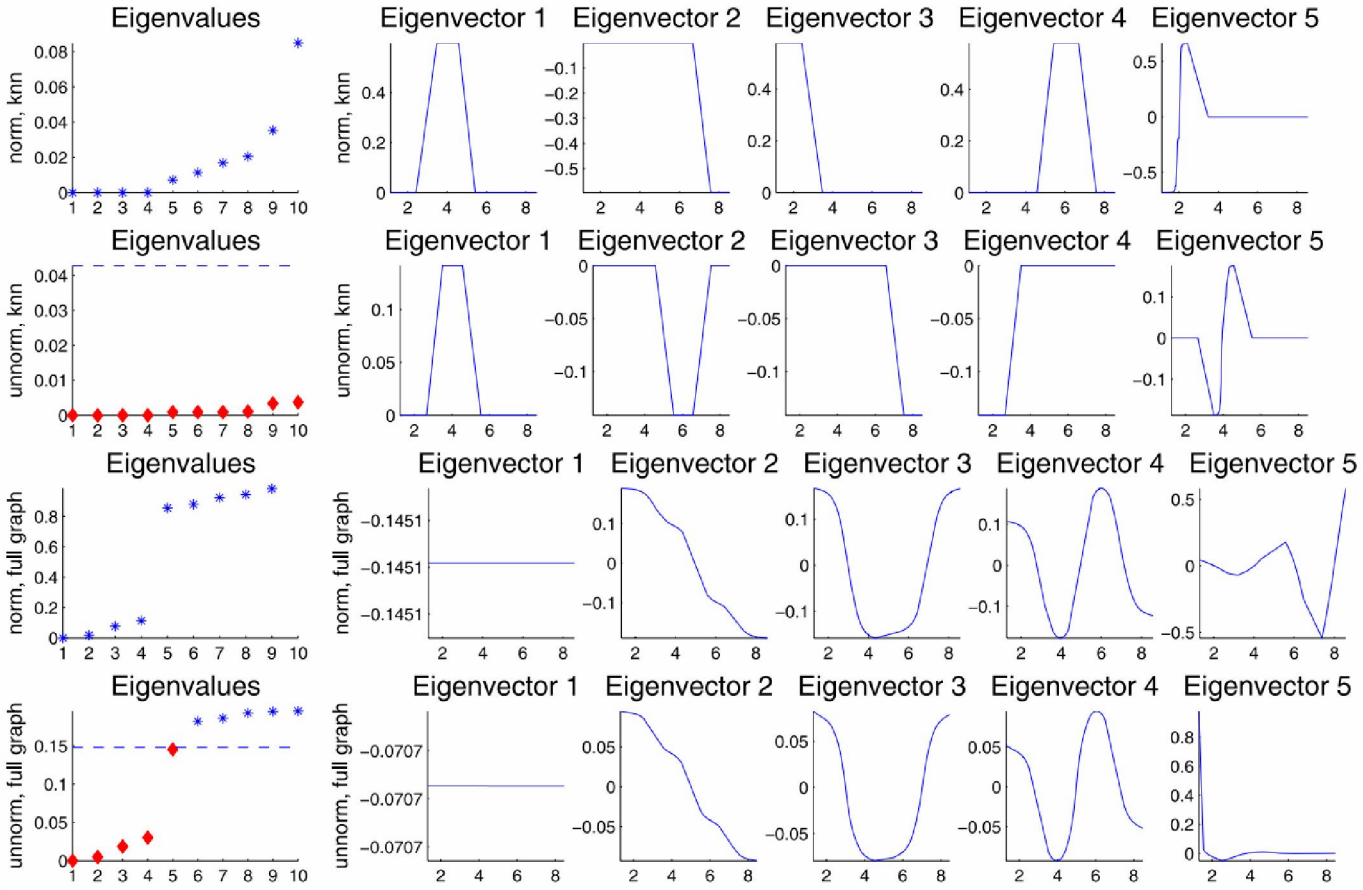
Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct.

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

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

一个实例


Histogram of the sample


Eigenvalues
Eigenvector 1
Eigenvector 2
Eigenvector 3
Eigenvector 4
Eigenvector 5
norm, knn
unnorm, knn
norm, full graph
unnorm, full graph

切割图

- 聚类问题的本质：
- 对于定值k和图G，选择一组划分：$A_1,A_2,...,Ak$，最小化下面的式子：

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

修正目标函数

- 上述的目标函数存在问题：在很多情况下，minCut的解，将图分成了一个点和其余的n-1个点。为了避免这个问题，目标函数应该显示的要求$A_1,A_2,...,Ak$足够大。

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

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

分析分母对目标函数的影响

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

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

Note that both objective functions take a small value if the clusters A_i are not too small. In particular, the minimum of the function $\sum_{i=1}^{k}(1/|A_i|)$ is achieved if all $|A_i|$ coincide, and the minimum of $\sum_{i=1}^{k}(1/\mathrm{vol}(A_i))$ is achieved if all $\mathrm{vol}(A_i)$ coincide. So what both objective functions try to achieve is that the clusters are "balanced", as measured by the number of vertices or edge weights, respectively.

当k=2时的RatioCut

- 目标函数：$\min_{A \subset V} \text{RatioCut}(A, \bar{A})$
- 定义向量f=(f1,f2,...fn)^T，

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

RatioCut与拉普拉斯矩阵的关系

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

f的约束条件

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

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

目标函数约束条件的放松relaxation

$$\min_{A \subset V} f' L f$$

subject to $f \perp \mathbb{1}$, f_i defined as in Slide 53, $\| f \| = \sqrt{n}$

$$\min_{f \in \mathbb{R}^n} f' L f \quad \text{subject to} \quad f \perp \mathbb{1}, \ \| f \| = \sqrt{n}$$

若划分为k个子集

The relaxation of the RatioCut minimization problem in the case of a general value k follows a similar principle as the one above. Given a partition of V into k sets $A_1, \ldots, A_k$, we define k indicator vectors $h_j = (h_{1,j}, \ldots, h_{n,j})'$ by

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

$$(i = 1, \ldots, n;\ j = 1, \ldots, k)$$

考察指示向量组成的矩阵

Then we set the matrix $H \in \mathbb{R}^{n \times k}$ as the matrix containing those k indicator vectors as columns. Observe that the columns in H are orthonormal to each other, that is $H'H = I$. Similar we can see that

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

$$h_i' L h_i = (H' L H)_{ii}$$

目标函数

$$\text{RatioCut}(A_1, \ldots, A_k)$$

$$= \sum_{i=1}^{k} h_i' L h_i = \sum_{i=1}^{k} (H' L H)_{ii} = \text{Tr}(H' L H)$$

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

Rayleigh-Ritz理论

By the Rayleigh-Ritz theorem (e.g., see Sect. 5.5.2 of Lütkepohl 1997) it can be seen immediately that the solution of this problem is given by the vector f which is the eigenvector corresponding to the second smallest eigenvalue of L (recall that the smallest eigenvalue of L is 0 with eigenvector $\mathbb{1}$). So we can approximate a minimizer of RatioCut by the second eigenvector of L. However, in order to obtain a partition of the graph we need to re-transform the real-valued solution vector f of the relaxed problem into a discrete indicator vector. The simplest way to do this is to use the sign of f as indicator function, that is to choose

$$\begin{cases} v_i \in A & \text{if } f_i \geq 0, \\ v_i \in \bar{A} & \text{if } f_i < 0. \end{cases}$$

随机游走和拉普拉斯矩阵的关系

- 图论中的随机游走是一个随机过程，它从一个顶点跳转到另外一个顶点。谱聚类即找到图的一个划分，使得随机游走在相同的簇中停留而几乎不会游走到其他簇。
- 转移矩阵：从顶点vi跳转到顶点vj的概率正比于边的权值wij

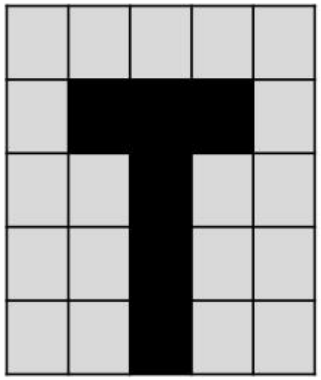
$$p_{ij} := w_{ij}/d_i \qquad P = D^{-1}W$$

谱聚类应用举例：图的像素分割

图像每个像素对应图的一个顶点

$$V = \{v_1, v_2, ..., v_{25}\}$$

$$W = \begin{vmatrix} w_{1,1} & .. & .. & .. & w_{1,25} \\ w_{2,1} & .. & .. & .. & w_{2,25} \\ : & & & & : \\ : & & & & : \\ w_{25,1} & .. & .. & .. & w_{25,25} \end{vmatrix}$$



$$w_{i,j} = e^{\frac{(x_i - x_j)^2}{-2\sigma^2}}$$ x_i, x_j为第i和j像素点的灰度值

图的像素分割过程

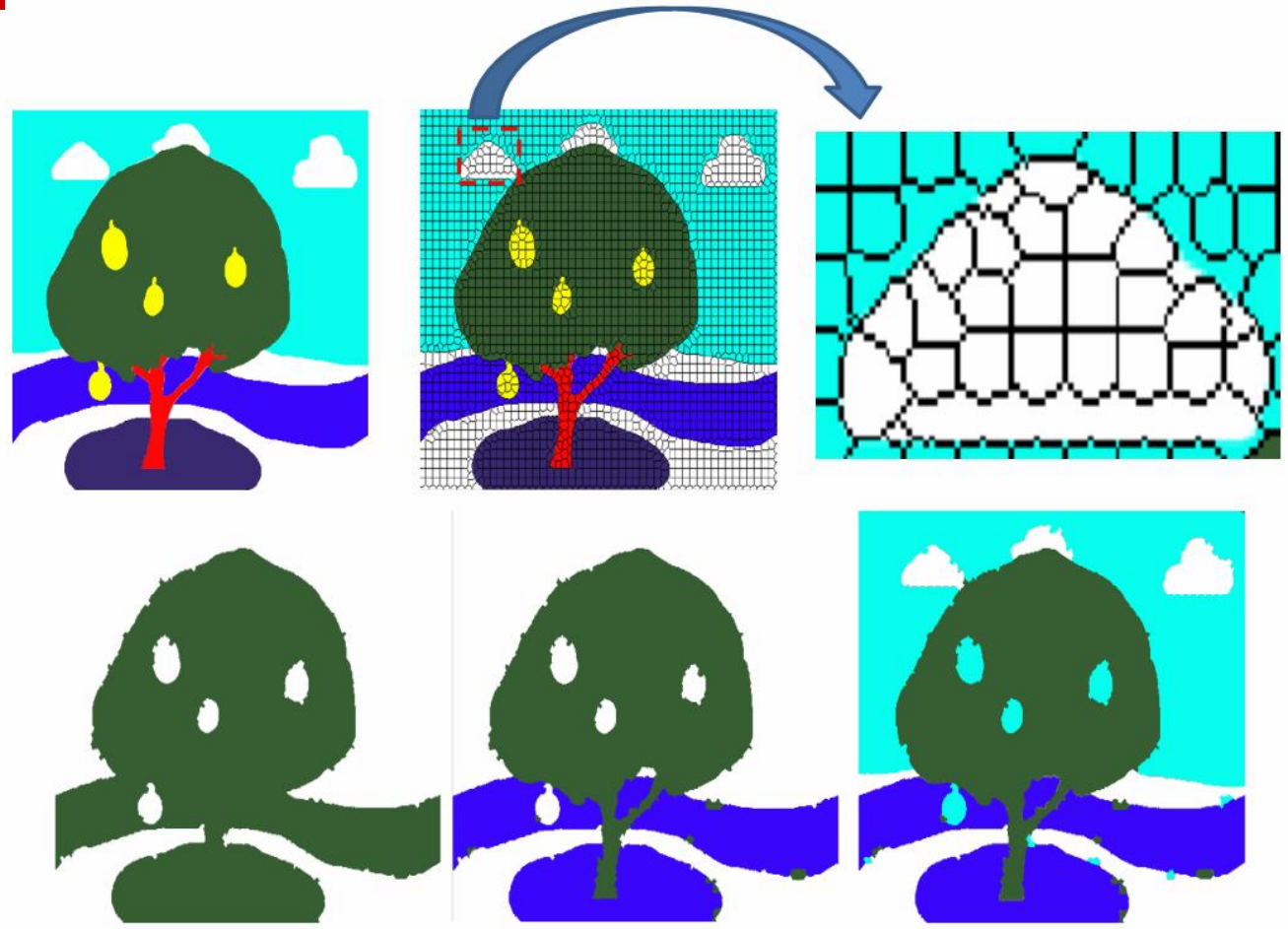
1、对图像进行超像素分割；

2、根据各超像素区域灰度平均值的相似度计算矩阵W及L；

3、计算L的特征值及特征向量 $Ve=\{v_{e1},v_{e2},\ldots,v_{en}\}$；

4、取出次小特征值对应的特征向量 v_{e2}，并对进行K-means聚类，得到2个Cluster

像素分割的处理过程



参考文献

- Clustering by fast search and find of density peak. Alex Rodriguez, Alessandro Laio(密度最大值聚类)
- A tutorial on spectral clustering，Ulrike von Luxburg, 2007)(谱聚类)
- Lütkepohl, H.: Handbook of Matrices. Wiley, Chichester (1997)(谱聚类中特征值问题)

感谢大家！

恳请大家批评指正！