聚类分析

耿修瑞

中国科学院空天信息创新研究院

gengxr@sina.com.cn

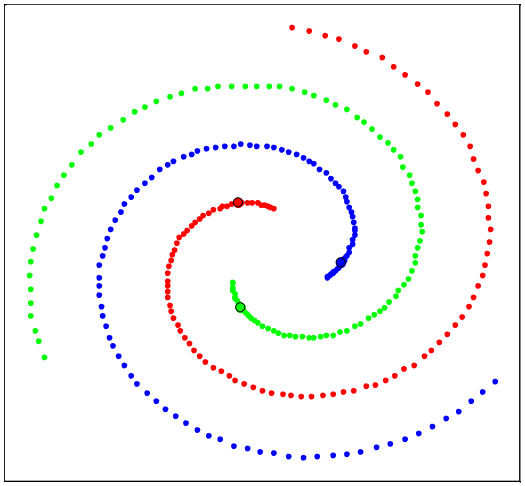
2025.5

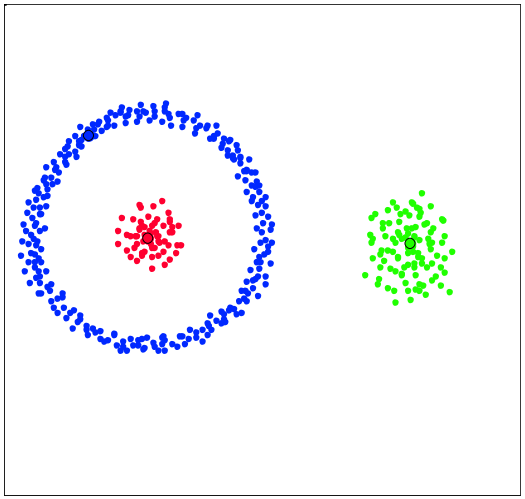
提纲

- 问题背景
- K-means
- 万有引力
- Mean-shift
- 谱聚类
- 连通中心演化(CCE)

问题背景

- 定义：聚类分析指将物理或抽象对象的集合分组为由类似的对象组成的多个类的分析过程。它是一种重要的人类行为。
- 应用：特征提取，数据压缩，图像分割，三维建模，社交网络分析等
- 关键点：
  - 类别数
  - 特征选择
  - 相似性度量：距离，密度，连通度
  - 聚类方法：K-means，万有引力，mean-shift，spectral clustering





K-means

- 聚类准则：使得类内点到该类均值的均方误差最小。

假设观测数据为 $\mathbf{V}=(\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_m)$，类别数为 r，则K-means的目标函数为

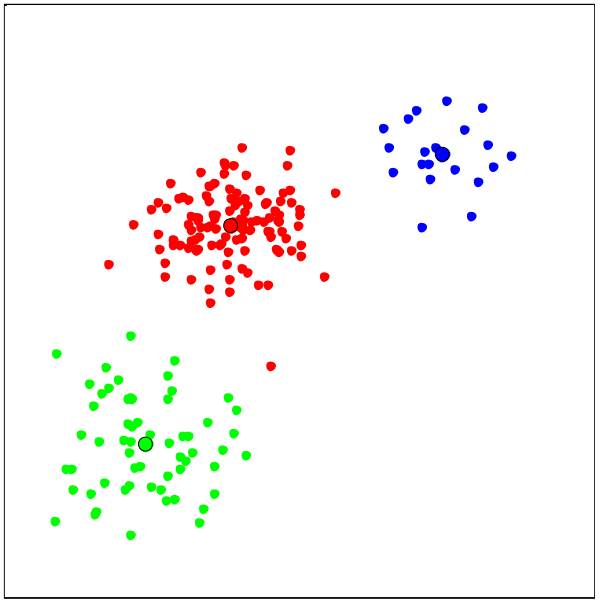
$$J(C,\boldsymbol{\mu}_k)=\sum_{k=1}^{r}\sum_{\mathbf{v}_i\in C_k}\left\|\mathbf{v}_i-\boldsymbol{\mu}_k\right\|^2$$

其中 C_k 为第 k 类所有样本的集合，$\boldsymbol{\mu}_k$ 为第 k 类的均值向量。

$$C=\{C_1,C_2,\cdots,C_r\}$$

K-means的模型即为

$$\min_{C,\boldsymbol{\mu}_k} J(C,\boldsymbol{\mu}_k)$$



K-means

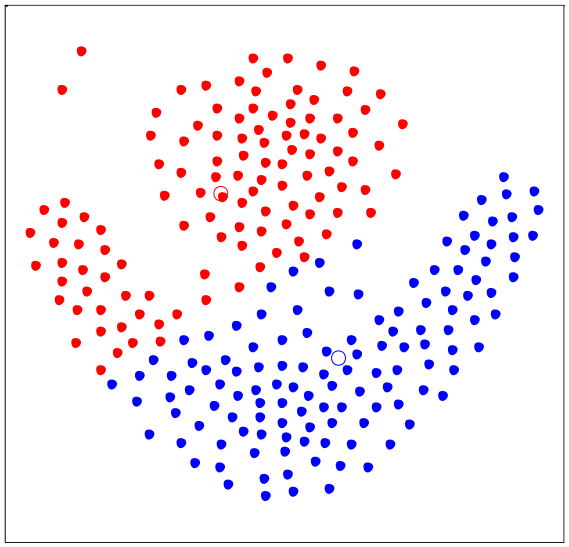
K-means问题的求解过程：

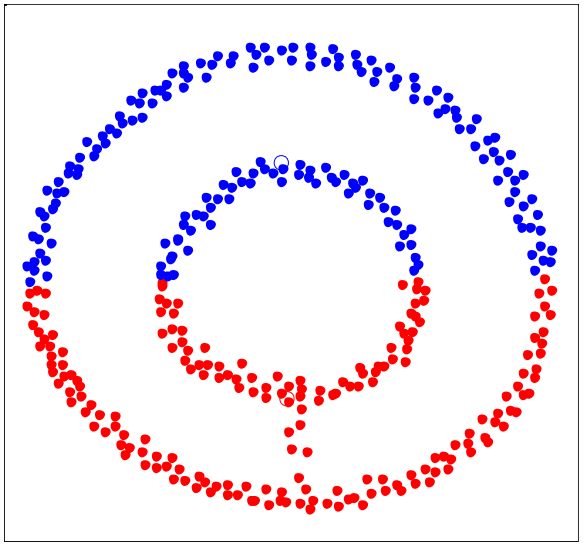
- 初始化选择r个类或r个类中心
- 重新分配每个点到距离其最近的类中心
- 重新计算r个类的均值，作为新的类中心，重复步骤2和3，直至类中心或类划分不变为止

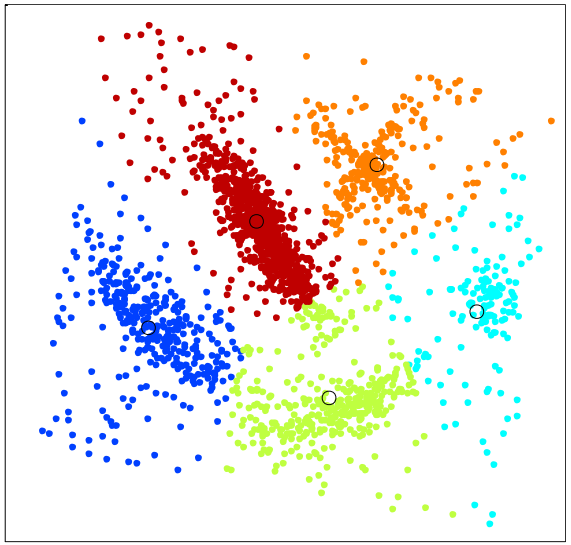
缺点：

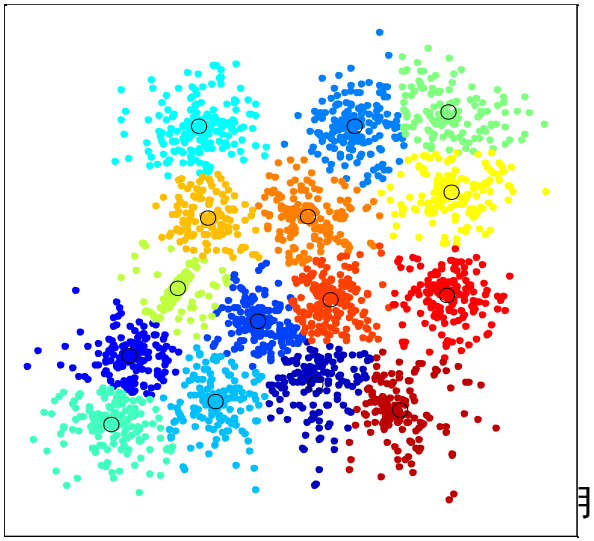
1. 需要给定类别数
2. 局部极值问题
3. 受散点分布形状影响
4. 受不同类的类内方差影响

K-means

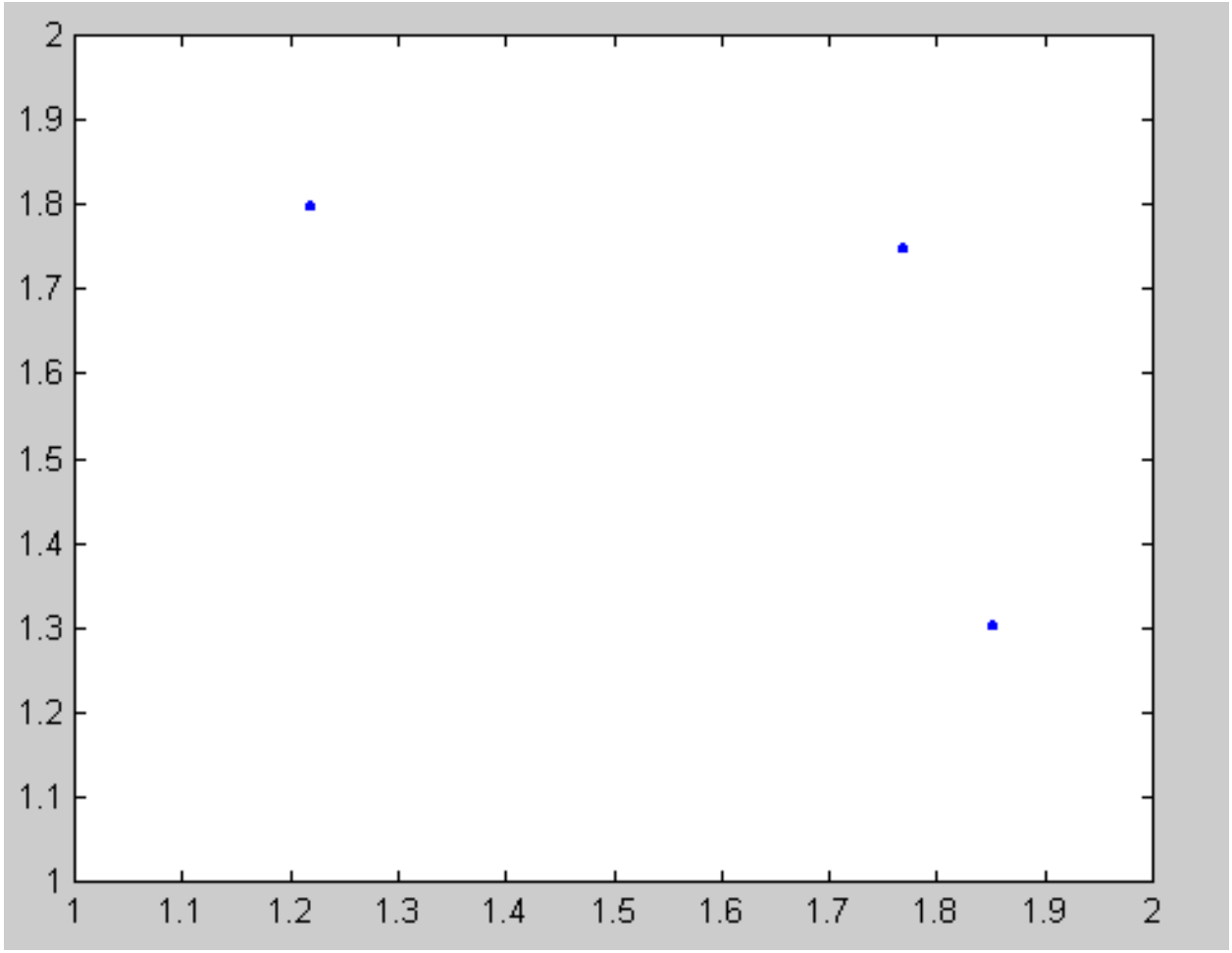








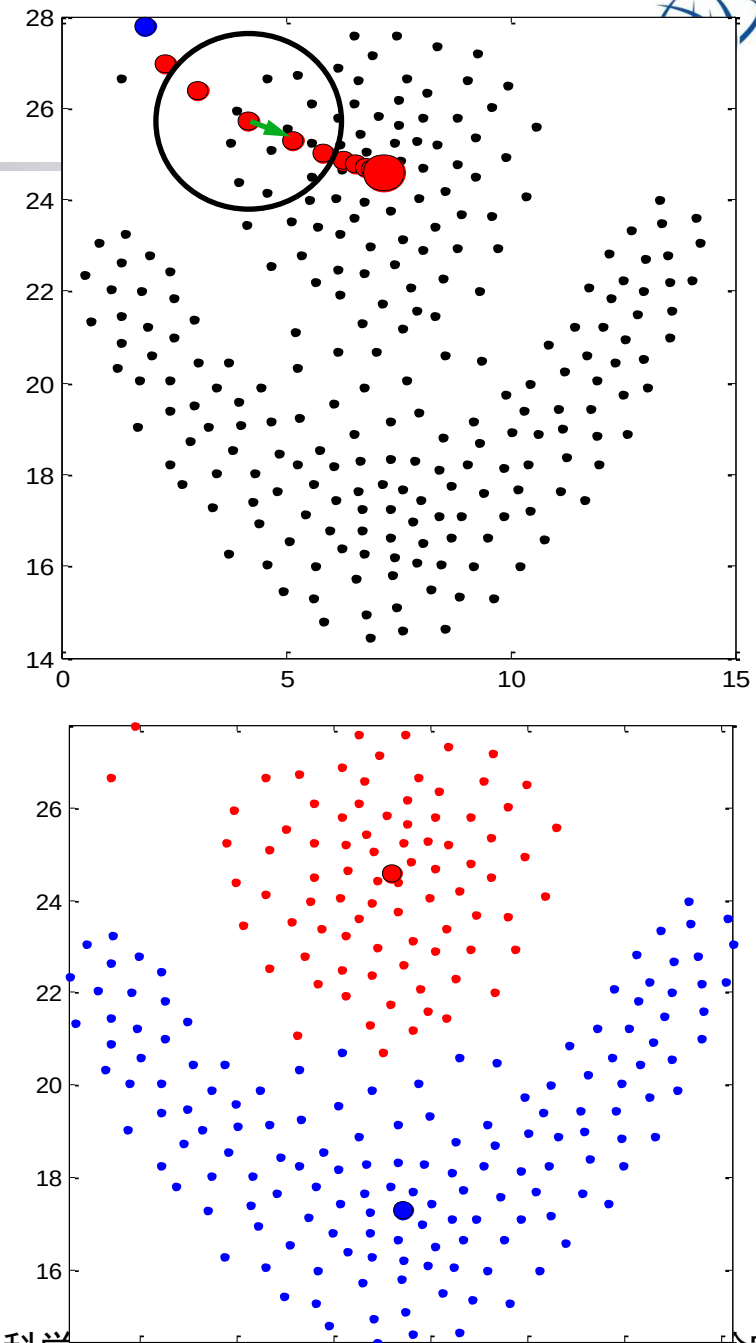
万有引力



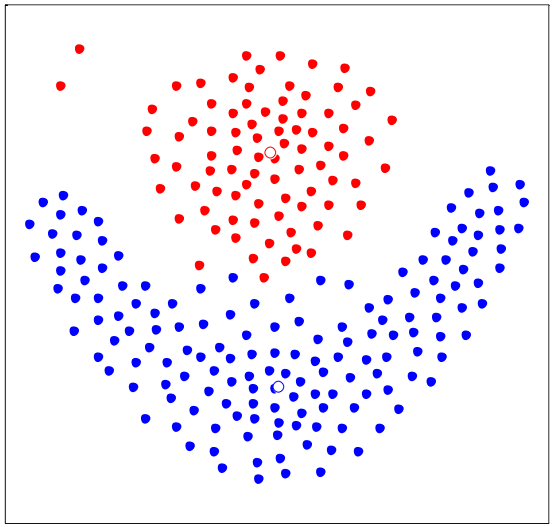
mean-shift

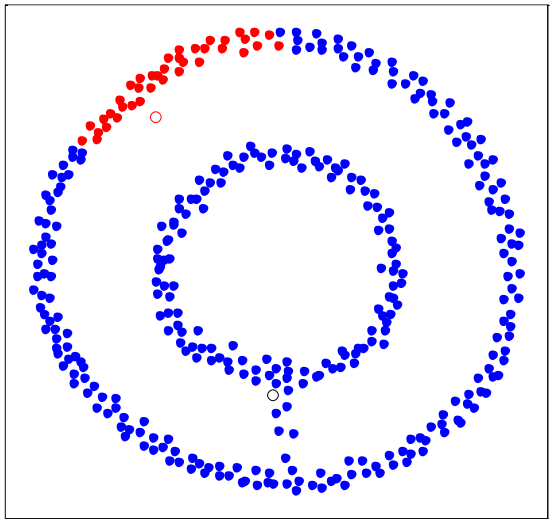
- 均值移动：以每个点为中心，一定半径范围内的所有点的密度中心（均值向量端点）作为该点的下一个位置，然后再以新的位置为中心，不停计算新的密度中心，直到密度中心不变为止。每个点都会对应到它们最终的密度中心。（右图，当半径设为2时，最终分为两类）
- 密度中心计算：核

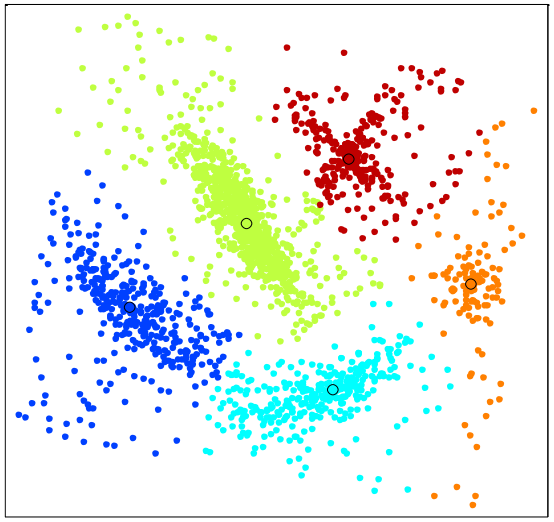
讨论优缺点？

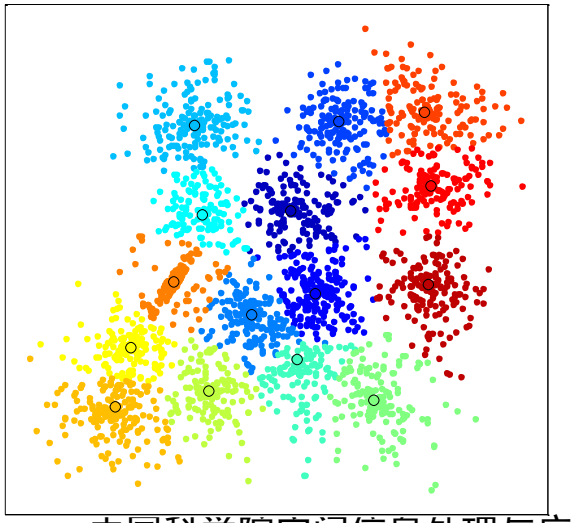


mean-shift









谱聚类

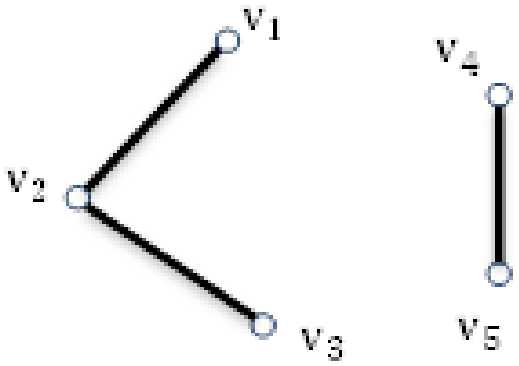
- 图：有向图，无向图

  由顶点(V)和边(E)组成，以G(V,E)表达。

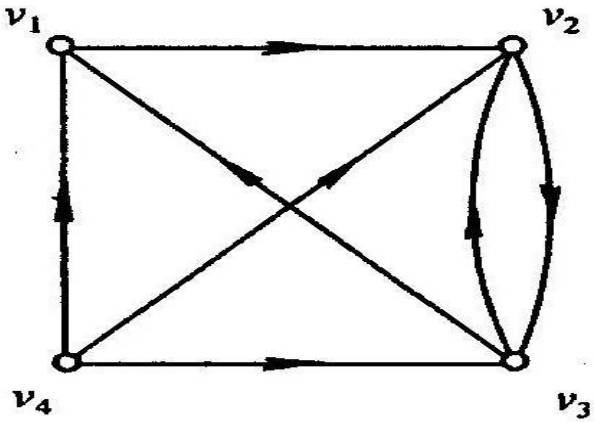
- 邻接矩阵A

  其中的任意元素a_{ij}表示两个结点$(\mathbf{v}_i, \mathbf{v}_j)$的连通情况。

无向图



有向图



对称

$$\mathbf{A}=\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&1&0&0&0\\0&0&0&0&1\\0&0&0&1&0\end{bmatrix}$$

0代表不连通
1代表连通

$$\mathbf{A}=\begin{bmatrix}0&1&0&0\\0&0&1&0\\1&1&0&0\\1&1&1&0\end{bmatrix}$$

不对称

谱聚类

相似度矩阵 W

其中的任意元素w_{ij}表示两个结点（$\mathbf{v}_i, \mathbf{v}_j$）的相似程度。相似度矩阵也叫加权邻接矩阵。典型的加权方式：

$$\begin{cases} w_{ij} = \exp\left(-\left\|\mathbf{v}_i - \mathbf{v}_j\right\|^2 \Big/ \sigma^2\right), i \neq j \\ w_{ij} = 0, \qquad\qquad\qquad\qquad i = j \end{cases}$$

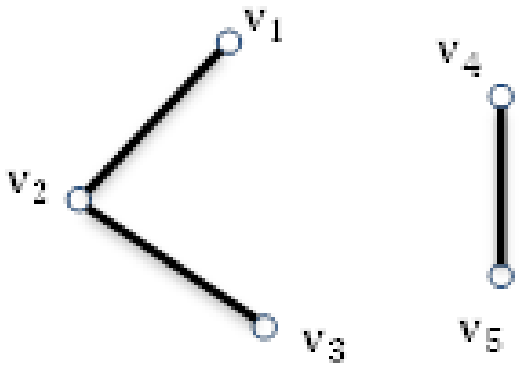
度矩阵 D

$$\mathbf{D} = diag\left(d_1, d_2, \cdots, d_m\right) \quad d_i = \sum_{j=1}^{m} w_{ij}$$

拉普拉斯矩阵 L

$$\mathbf{L} = \mathbf{D} - \mathbf{W} \text{或者} \mathbf{L} = \mathbf{D} - \mathbf{A}$$

谱聚类



邻接矩阵：$\mathbf{A}=\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&1&0&0&0\\0&0&0&0&1\\0&0&0&1&0\end{bmatrix}$

度矩阵：$\mathbf{D}=\begin{bmatrix}1&0&0&0&0\\0&2&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix}$

拉普拉斯矩阵：$\mathbf{L}=\mathbf{D}-\mathbf{A}=\begin{bmatrix}1&-1&0&0&0\\-1&2&-1&0&0\\0&-1&1&0&0\\0&0&0&1&-1\\0&0&0&-1&1\end{bmatrix}$

讨论相似度矩阵如何构建？

谱聚类

拉普拉斯矩阵的性质

- $\mathbf{L}$ 是对称半正定矩阵？
- $\mathbf{L}$ 的最小特征值是0，相应的特征向量是元素全为1的向量。其他的特征值也都非负

$$0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_m$$

$$\mathbf{L1} = \left(\mathbf{D} - \mathbf{W}\right)\mathbf{1} = \mathbf{D1} - \mathbf{W1} = 0 \times \mathbf{1}$$

谱聚类

- 经典谱聚类方法
  - 比例割谱聚类
  - 归一化割谱聚类，Shi and Malik (2000)
  - 归一化谱聚类，Ng et al. (2002)
  - 随机游走谱聚类

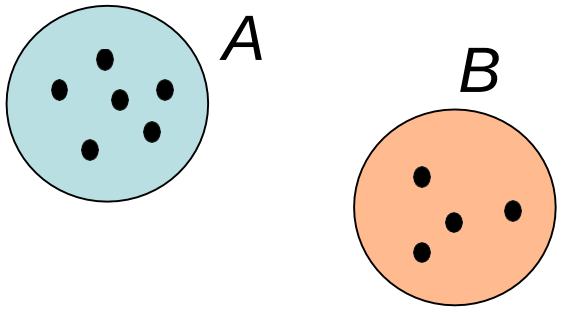
谱聚类

- 比例割谱聚类（两类问题）

对有m个节点的无向图，对应的相似度矩阵（加权邻接矩阵）为$\mathbf{W} \in R^{m\times m}$，拉普拉斯矩阵为$\mathbf{L} \in R^{m\times m}$：

假定所有的点可以分为2类，分别记为A,B，设

$$cut\left(A,B\right)=\sum_{i\in A,j\in B} w_{ij}$$



比例割函数为

$$Rcut\left(A,B\right)=\frac{cut\left(A,B\right)}{\left|A\right|}+\frac{cut\left(A,B\right)}{\left|B\right|}$$

穷举法？

其中$\left|A\right|$为A类节点数目，$\left|B\right|$为B类节点数目。显然最小化比例割就对应着一个最佳的2类分类问题。

谱聚类

- 拉普拉斯矩阵的一个重要公式

$$\mathbf{f}^T\mathbf{L}\mathbf{f} = \frac{1}{2}\sum_{i,j=1}^{m} w_{ij}\left(f_i - f_j\right)^2$$

$$\mathbf{f}^T\mathbf{L}\mathbf{f} = \mathbf{f}^T\mathbf{D}\mathbf{f} - \mathbf{f}^T\mathbf{W}\mathbf{f} = \sum_{i=1}^{m} d_i f_i^2 - \sum_{i,j=1}^{m} w_{ij} f_i f_j$$

$$= \frac{1}{2}\left(\sum_{i=1}^{m} d_i f_i^2 - 2\sum_{i,j=1}^{m} w_{ij} f_i f_j + \sum_{j=1}^{m} d_j f_j^2\right) = \frac{1}{2}\sum_{i,j=1}^{m} w_{ij}\left(f_i - f_j\right)^2$$

讨论此公式的意义，是否可以为自动类别归属带来启示！

谱聚类

- 比例割谱聚类（两类问题）

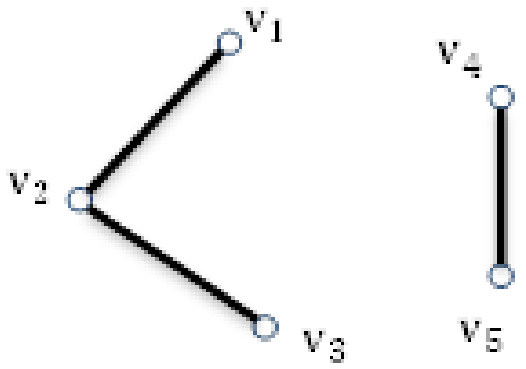
于是我们可以尝试将2分类问题转化为如下优化问题

$$\min_{\mathbf{f}} \mathbf{f}^T \mathbf{L}\mathbf{f}$$
$$s.t\ \mathbf{f}^T \mathbf{1} = 0, \quad \mathbf{f}^T \mathbf{f} = 1$$

该模型的解是否对应着一个自动的类别分配结果呢？

谱聚类

简单例子1



$$\mathbf{A}=\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&1&0&0&0\\0&0&0&0&1\\0&0&0&1&0\end{bmatrix}\qquad \mathbf{D}=\begin{bmatrix}1&0&0&0&0\\0&2&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix}$$

$$\mathbf{L}=\mathbf{D}-\mathbf{A}=\begin{bmatrix}1&-1&0&0&0\\-1&2&-1&0&0\\0&-1&1&0&0\\0&0&0&1&-1\\0&0&0&-1&1\end{bmatrix}$$

$$\mathbf{V}=\begin{bmatrix}0.4472&0.3651&-0.7071&0&0.4082\\0.4472&0.3651&0&0&-0.8165\\0.4472&0.3651&0.7071&0&0.4082\\0.4472&-0.5477&0&-0.7071&0\\0.4472&-0.5477&0&0.7071&0\end{bmatrix}$$

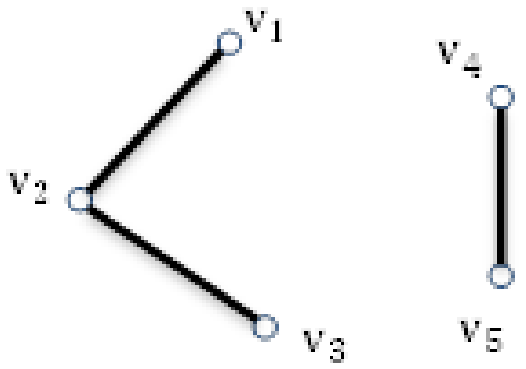
$$\mathbf{\Lambda}=\begin{bmatrix}0&0&0&0&0\\0&0.0000&0&0&0\\0&0&1.0000&0&0\\0&0&0&2.0000&0\\0&0&0&0&3.0000\end{bmatrix}$$

$$\mathbf{f}=\begin{bmatrix}\sqrt{2/3}&\sqrt{2/3}&\sqrt{2/3}&-\sqrt{3/2}&-\sqrt{3/2}\end{bmatrix}^T$$

$$\mathbf{f}/\|\mathbf{f}\|=\begin{bmatrix}0.3651&0.3651&0.3651&-0.5477&-0.5477\end{bmatrix}^T$$

谱聚类

简单例子1



$$\mathbf{L}=\begin{bmatrix}1&-1&0&0&0\\-1&2&-1&0&0\\0&-1&1&0&0\\0&0&0&1&-1\\0&0&0&-1&1\end{bmatrix}\qquad \mathbf{\Lambda}=\begin{bmatrix}0&0&0&0&0\\0&0.0000&0&0&0\\0&0&1.0000&0&0\\0&0&0&2.0000&0\\0&0&0&0&3.0000\end{bmatrix}$$

$$\mathbf{V}=\begin{bmatrix}0.4472&0.3651&-0.7071&0&0.4082\\0.4472&0.3651&0&0&-0.8165\\0.4472&0.3651&0.7071&0&0.4082\\0.4472&-0.5477&0&-0.7071&0\\0.4472&-0.5477&0&0.7071&0\end{bmatrix}$$

$$\mathbf{f}_1^T\mathbf{L}\mathbf{f}_1=\frac{1}{2}\sum_{i,j=1}^{m}w_{ij}\left(f_{i1}-f_{j1}\right)^2=\lambda_2=0$$

$$\mathbf{f}_2^T\mathbf{L}\mathbf{f}_2=\frac{1}{2}\sum_{i,j=1}^{m}w_{ij}\left(f_{i2}-f_{j2}\right)^2=\lambda_3=1$$

$$\mathbf{f}=\begin{bmatrix}\sqrt{2/3}&\sqrt{2/3}&\sqrt{2/3}&-\sqrt{3/2}&-\sqrt{3/2}\end{bmatrix}^T$$

$$\mathbf{f}_1=\mathbf{f}/\|\mathbf{f}\|=\begin{bmatrix}0.3651&0.3651&0.3651&-0.5477&-0.5477\end{bmatrix}^T$$

谱聚类

比例割谱聚类（两类问题）

假定有m维向量$\mathbf{f}$，它的每个元素代表这个每个节点的类别归属，且有：

$$f_i = \begin{cases} \sqrt{|B|}/\sqrt{|A|} & \mathbf{v}_i \in A \\ -\sqrt{|A|}/\sqrt{|B|} & \mathbf{v}_i \in B \end{cases}$$

则

$$\mathbf{f}^T\mathbf{L}\mathbf{f} = \frac{1}{2}\sum_{i,j=1}^{m} w_{ij}\left(f_i - f_j\right)^2$$

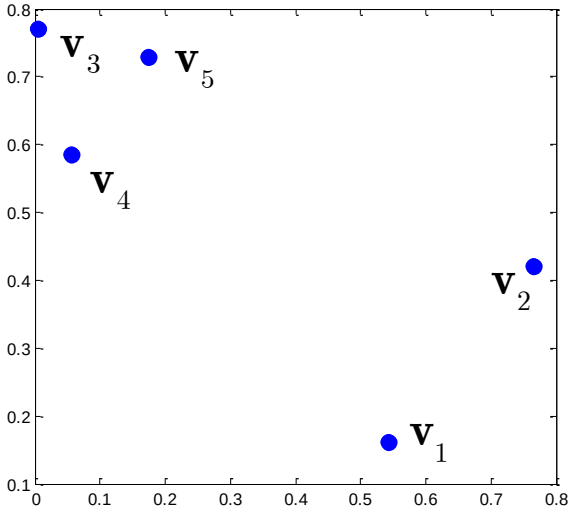
$$= \frac{1}{2}\sum_{i,j\in A}^{m} w_{ij}\left(f_i - f_j\right)^2 + \frac{1}{2}\sum_{i,j\in B}^{m} w_{ij}\left(f_i - f_j\right)^2 + \frac{1}{2}\sum_{i\in A,j\in B}^{m} w_{ij}\left(f_i - f_j\right)^2 + \frac{1}{2}\sum_{i\in B,j\in A}^{m} w_{ij}\left(f_i - f_j\right)^2$$

$$= 0 + 0 + \frac{1}{2}\sum_{i\in A,j\in B}^{m} w_{ij}\left(\frac{\sqrt{|B|}}{\sqrt{|A|}} + \frac{\sqrt{|A|}}{\sqrt{|B|}}\right)^2 + \frac{1}{2}\sum_{i\in B,j\in A}^{m} w_{ij}\left(\frac{\sqrt{|A|}}{\sqrt{|B|}} + \frac{\sqrt{|B|}}{\sqrt{|A|}}\right)^2$$

$$= \sum_{i\in A,j\in B}^{m} w_{ij}\left(\frac{|B|}{|A|} + \frac{|A|}{|B|} + 2\right) = \sum_{i\in A,j\in B}^{m} w_{ij}\left(\frac{|V|}{|A|} + \frac{|V|}{|B|}\right) = |V|\,Rcut\left(A,B\right)$$

谱聚类

简单例子2



$$V = \begin{bmatrix} 0.5430 & 0.1623 \\ 0.7648 & 0.4211 \\ 0.0057 & 0.7715 \\ 0.0568 & 0.5857 \\ 0.1742 & 0.7286 \end{bmatrix}$$

$$w_{ij} = \exp\left(-\left\|\mathbf{v}_i - \mathbf{v}_j\right\|^2 \Big/ \sigma^2\right), \sigma = 0.2$$

$$\mathbf{W} = \begin{bmatrix} 0 & 0.0549 & 0 & 0 & 0 \\ 0.0549 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.3954 & 0.4696 \\ 0 & 0 & 0.3954 & 0 & 0.4255 \\ 0 & 0 & 0.4696 & 0.4255 & 0 \end{bmatrix}$$

$$\mathbf{L} = \begin{bmatrix} 0.0549 & -0.0549 & 0 & 0 & 0 \\ -0.0549 & 0.0549 & 0 & 0 & 0 \\ 0 & 0 & 0.8650 & -0.3954 & -0.4696 \\ 0 & 0 & -0.3954 & 0.8209 & -0.4255 \\ 0 & 0 & -0.4696 & -0.4255 & 0.8952 \end{bmatrix}$$

$$[\mathbf{H}\ \mathbf{\Lambda}] = eig(\mathbf{L})$$

$$\min_{\mathbf{H} \in R^{m \times k}} tr\left(\mathbf{H}^T \mathbf{L} \mathbf{H}\right)$$

$$s.t\ \mathbf{H}^T \mathbf{H} = \mathbf{I}$$

$$\mathbf{V} = \begin{bmatrix} -0.4472 & 0.5476 & 0.7072 & 0.0000 & -0.0000 \\ -0.4472 & 0.5478 & -0.7070 & -0.0000 & -0.0000 \\ -0.4472 & -0.3652 & -0.0001 & 0.5452 & -0.6078 \\ -0.4472 & -0.3651 & -0.0000 & -0.7990 & -0.1683 \\ -0.4472 & -0.3651 & -0.0001 & 0.2537 & 0.7761 \end{bmatrix}$$

$$\mathbf{\Lambda} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0000 & 0 & 0 & 0 \\ 0 & 0 & 0.1098 & 0 & 0 \\ 0 & 0 & 0 & 1.2259 & 0 \\ 0 & 0 & 0 & 0 & 1.3552 \end{bmatrix}$$

谱聚类

归一化割谱聚类（两类问题）

目标函数：

$$Ncut(A,B)=\frac{cut(A,B)}{\sum_{\mathbf{v}_i\in A,\mathbf{v}_j\in V} w_{ij}}+\frac{cut(A,B)}{\sum_{\mathbf{v}_i\in B,\mathbf{v}_j\in V} w_{ij}}$$

模型：

$$\min_{\mathbf{y}} \frac{\mathbf{y}^T\mathbf{L}\mathbf{y}}{\mathbf{y}^T\mathbf{D}\mathbf{y}}, s.t\ \mathbf{y}^T\mathbf{1}=0$$

等价于：

$$\min_{\mathbf{g}} \mathbf{g}^T\mathbf{D}^{-0.5}\mathbf{L}\mathbf{D}^{-0.5}\mathbf{g}$$

$$s.t\ \mathbf{g}^T\mathbf{D}^{0.5}\mathbf{1}=0,\quad \mathbf{g}^T\mathbf{g}=\sum w_{ij}$$

$$f_i=\begin{cases}\sqrt{vol(B)}/\sqrt{vol(A)} & \mathbf{v}_i\in A\\ -\sqrt{vol(A)}/\sqrt{vol(B)} & \mathbf{v}_i\in B\end{cases}$$

$$vol(A)=\sum_{\mathbf{v}_i\in A} w_{ij}$$

$$vol(B)=\sum_{\mathbf{v}_i\in B} w_{ij}$$

$$vol(V)=\sum_{i,j} w_{ij}$$

谱聚类

- 比例割谱聚类（k类问题）

$$\min_{\mathbf{H}\in R^{m\times k}} tr\left(\mathbf{H}^T\mathbf{L}\mathbf{H}\right)$$
$$s.t\ \mathbf{H}^T\mathbf{H}=\mathbf{I}$$

- 归一化割谱聚类（k类问题）

$$\min_{\mathbf{T}\in R^{m\times k}} tr\left(\mathbf{T}^T\mathbf{D}^{-0.5}\mathbf{L}\mathbf{D}^{-0.5}\mathbf{T}\right)$$
$$s.t\ \mathbf{T}^T\mathbf{T}=\mathbf{I}$$

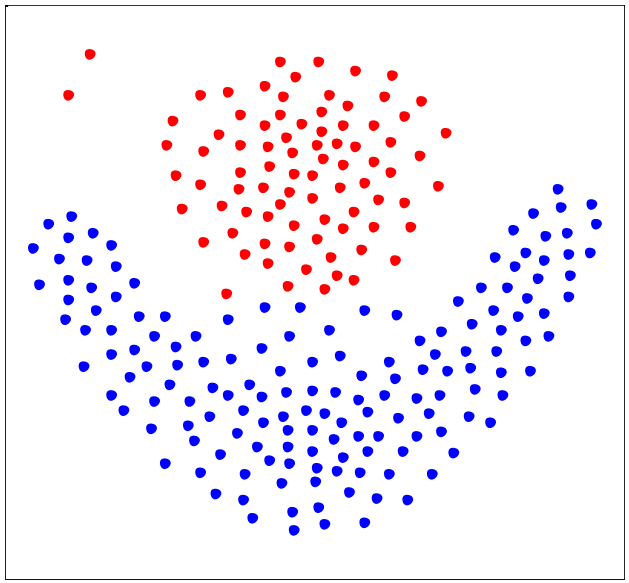
- 归一化谱聚类（k类问题）

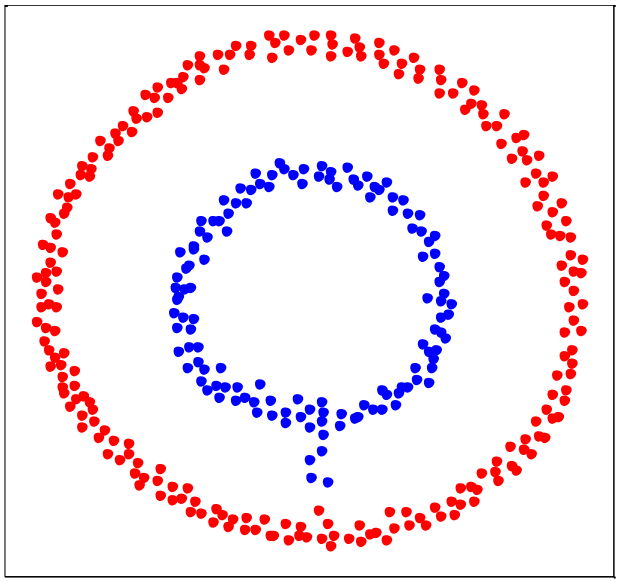
$$\max_{\mathbf{T}\in R^{m\times k}} tr\left(\mathbf{T}^T\mathbf{D}^{-0.5}\mathbf{W}\mathbf{D}^{-0.5}\mathbf{T}\right)$$
$$s.t\ \mathbf{T}^T\mathbf{T}=\mathbf{I}$$

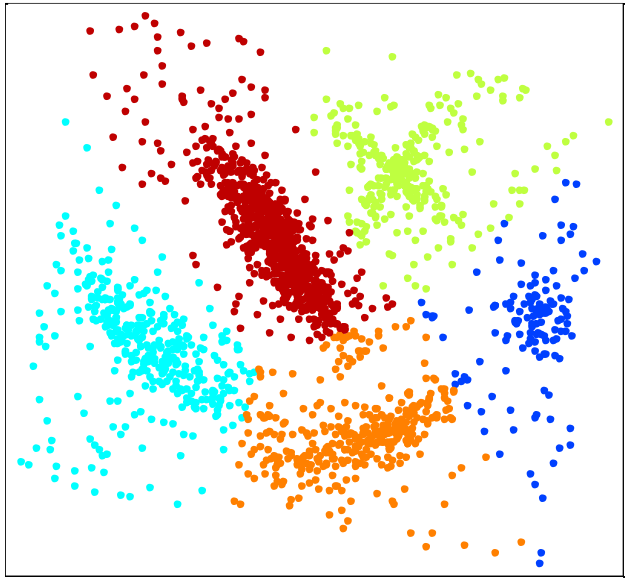
- 随机游走谱聚类（k类问题）

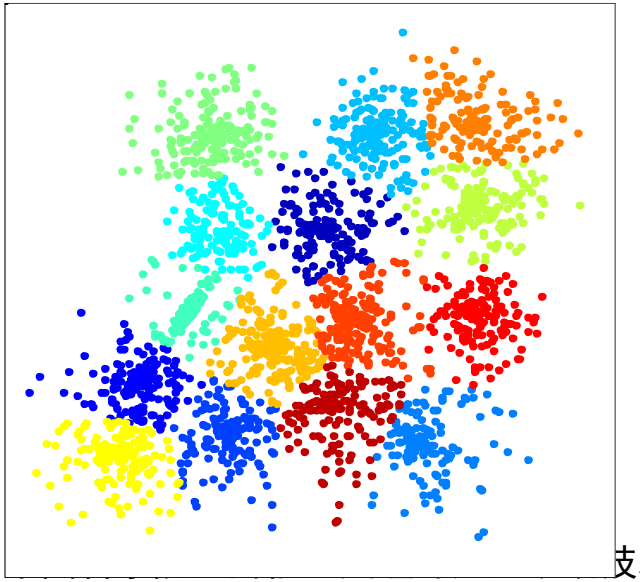
$$\max_{\mathbf{T}\in R^{m\times k}} tr\left(\mathbf{T}^T\mathbf{D}^{-1}\mathbf{W}\mathbf{T}\right)$$
$$s.t\ \mathbf{T}^T\mathbf{T}=\mathbf{I}$$

谱聚类-归一化割例子







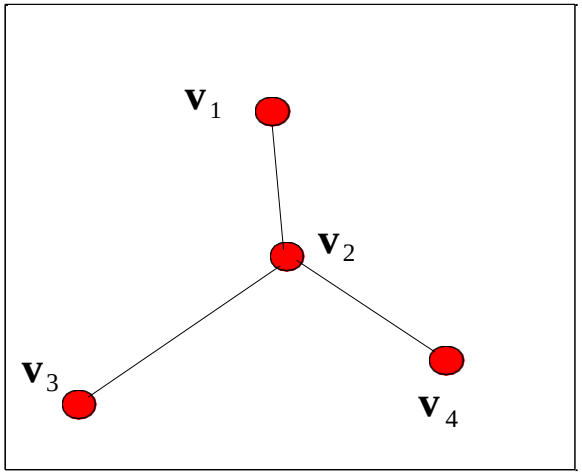


谱聚类方法的问题

- 需要给定类别数
- 计算复杂度问题
- 参数敏感性

连通中心演化（CCE）

自相邻邻接矩阵（相似度矩阵）



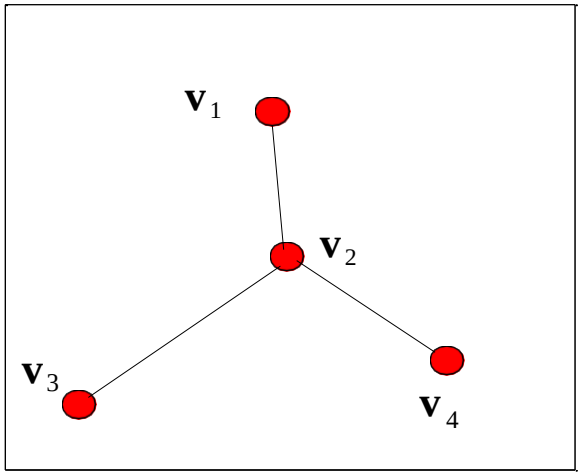
$$\mathbf{A}=\begin{bmatrix}1 & 1 & 0 & 0\\ 1 & 1 & 1 & 1\\ 0 & 1 & 1 & 0\\ 0 & 1 & 0 & 1\end{bmatrix}$$

邻接矩阵

与谱聚类不同的是，我们这里假定每个点都是自连通的，因此邻接矩阵的对角线元素都为1。

连通中心演化（CCE）

动机



	所有从v2到v2长度为3的路径
1	v2 → v2 → v2 → v2
2	v2 → v1 → v2 → v2
3	v2 → v3 → v2 → v2
4	v2 → v4 → v2 → v2
5	v2 → v1 → v1 → v2
6	v2 → v3 → v3 → v2
7	v2 → v4 → v4 → v2
8	v2 → v2 → v1 → v2
9	v2 → v2 → v3 → v2
10	v2 → v2 → v4 → v2

$$\mathbf{A}=\begin{bmatrix}1&1&0&0\\1&1&1&1\\0&1&1&0\\0&1&0&1\end{bmatrix}\quad \mathbf{A}^2=\begin{bmatrix}2&2&1&1\\2&4&2&2\\1&2&2&1\\1&2&1&2\end{bmatrix}$$

$$\mathbf{A}^3=\begin{bmatrix}4&6&3&3\\6&\boxed{10}&6&6\\3&6&4&3\\3&6&3&4\end{bmatrix}$$

邻接矩阵的k次方的任意元素的值即为对应顶点间长度为k的路径的条数

连通中心演化（CCE）

连通度的定义

给定一个相似度矩阵S，它的k次方S^k的元素 $s_{ij}^{(k)}$ 被定义为顶点 $\mathbf{v}_i,\mathbf{v}_j$ 的k阶连通度，记为 $con^{(k)}(\mathbf{v}_i,\mathbf{v}_j)$，特别的，对角元素 $s_{ii}^{(k)}$ 称之为顶点 $\mathbf{v}_i$ 的k阶连通度，记为 $con^{(k)}(\mathbf{v}_i,\mathbf{v}_i)$

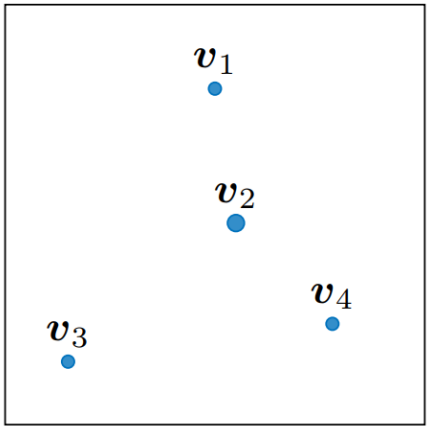
连通中心（聚类中心）的定义

如果一个顶点 $\mathbf{v}_i$ 满足如下不等式，它将是一个无向图的连通中心，并定义其为该数据的k阶聚类中心：

$$con^{(k)}(\mathbf{v}_i,\mathbf{v}_i) > con^{(k)}(\mathbf{v}_i,\mathbf{v}_j), j=1,\cdots,n(j\neq i)$$

连通中心演化（CCE）

连通中心的确定（相似度矩阵）



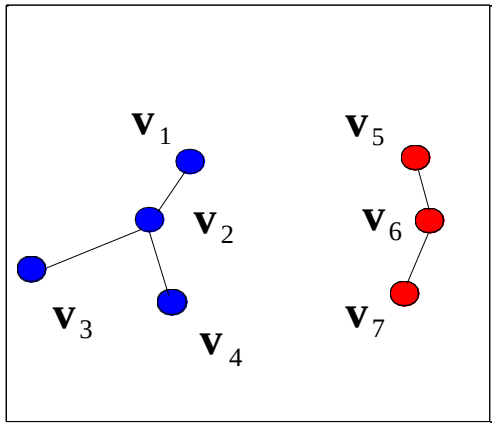
$$\mathbf{S}^2 = \begin{bmatrix} 3.16 & 3.34 & 3.02 & 3.16 \\ 3.34 & 3.56 & 3.24 & 3.39 \\ 3.02 & 3.24 & 3.01 & 3.09 \\ 3.16 & 3.39 & 3.09 & 3.24 \end{bmatrix}$$

$$\mathbf{S}^4 = \begin{bmatrix} 40.28 & 42.96 & 39.25 & 40.88 \\ 42.96 & 45.83 & 41.87 & 43.60 \\ 39.25 & 41.87 & 38.26 & 39.84 \\ 40.88 & 43.60 & 39.84 & 41.49 \end{bmatrix}$$

根据连通度和连通中心的定义，可以自动确定得到各个尺度下（k）的数据中心。

连通中心演化（CCE）

多类情形（邻接矩阵）



$$\mathbf{A}=\begin{bmatrix}1&1&0&0&0&0&0\\1&1&1&1&0&0&0\\0&1&1&0&0&0&0\\0&1&0&1&0&0&0\\0&0&0&0&1&1&0\\0&0&0&0&1&1&1\\0&0&0&0&0&1&1\end{bmatrix}\quad \mathbf{A}^3=\begin{bmatrix}4&6&3&3&0&0&0\\6&10&6&6&0&0&0\\3&6&4&3&0&0&0\\3&6&3&4&0&0&0\\0&0&0&0&4&5&3\\0&0&0&0&5&7&5\\0&0&0&0&3&5&4\end{bmatrix}$$

根据连通中心的定义，可以自动确定得到各个尺度下（k）的类别中心和类别数。那么对于那些非中心点，如何对它们进行分类呢？

连通中心演化（CCE）

相对连通度的定义

对于数据集中两个顶点$\mathbf{v}_i, \mathbf{v}_j$，$\mathbf{v}_j$相对于 $\mathbf{v}_i$ 的k阶互连通度定义为

$$rcon^{(k)}\left(\mathbf{v}_i, \mathbf{v}_j\right) = con^{(k)}\left(\mathbf{v}_i, \mathbf{v}_j\right) \Big/ con^{(k)}\left(\mathbf{v}_i, \mathbf{v}_i\right)$$

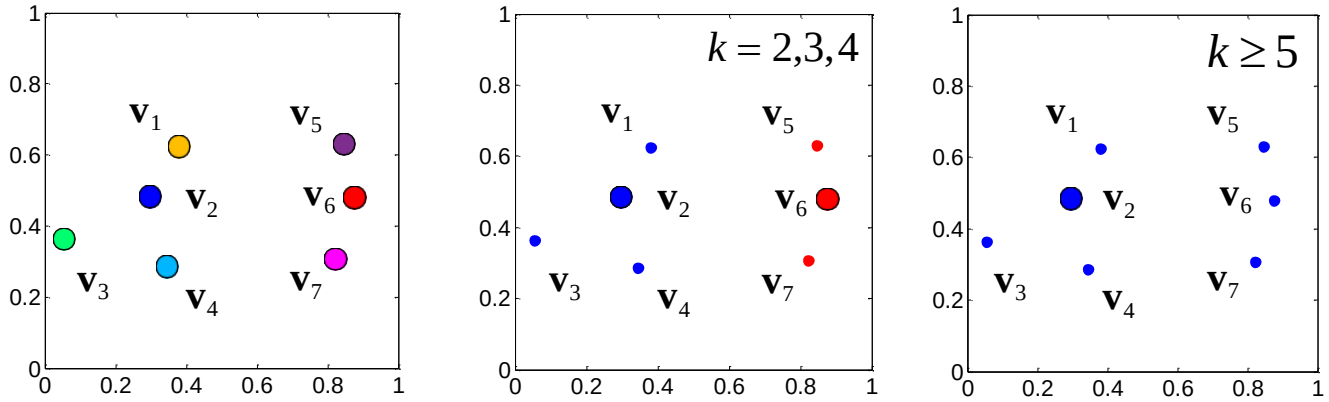
分类规则

如果我们有m个聚类中心，$\mathbf{v}_{c_i}\left(c_i \in \{1,2,\cdots,n\}, i = 1,2,\cdots,m\right)$，对于任意一个顶点 $\mathbf{v}_j$，它将按下列规则归为 $\mathbf{v}^*$

$$\mathbf{v}^* = \arg\max_{\mathbf{v}_{c_i}} \left(rcon^{(k)}\left(\mathbf{v}_{c_i}, \mathbf{v}_j\right)\right)$$

连通中心演化（CCE）

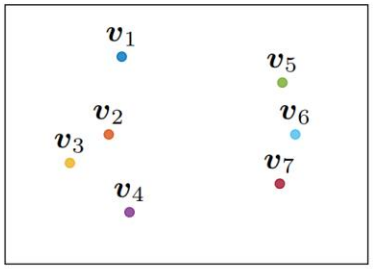
多类情形（相似度矩阵）



当k=1时，由于相似度矩阵一般都为对角最大矩阵，所以，此时每个点都可以当做聚类中心。因此有多少点，数据就包含多少类别数。随着k的增大，类别数一般逐渐减少，并最终所有的点都聚为一类。对每一个k，都可以得到相应的类别中心、类别数以及分类结果。所以，基于CCE，可以得到数据从微观到宏观的各个尺度的聚类结果。

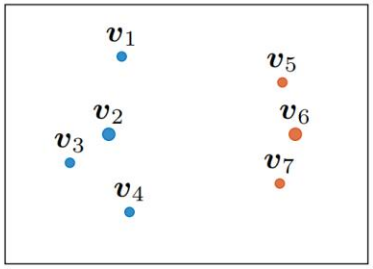
连通中心演化（CCE）

多类情形（相似度矩阵）



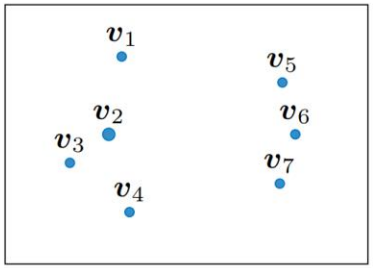
$$\mathbf{S} = \begin{bmatrix} 1 & 0.75 & 0.52 & 0.32 & 0.29 & 0.19 & 0.15 \\ 0.75 & 1 & 0.9 & 0.74 & 0.22 & 0.2 & 0.23 \\ 0.52 & 0.9 & 1 & 0.76 & 0.09 & 0.09 & 0.13 \\ 0.32 & 0.74 & 0.76 & 1 & 0.15 & 0.21 & 0.34 \\ 0.29 & 0.22 & 0.09 & 0.15 & 1 & 0.88 & 0.62 \\ 0.19 & 0.2 & 0.09 & 0.21 & 0.88 & 1 & 0.88 \\ 0.15 & 0.23 & 0.13 & 0.34 & 0.62 & 0.88 & 1 \end{bmatrix}$$

(a)



$$\mathbf{S}^4 = \begin{bmatrix} 20.1 & 25.8 & 23.02 & 22.16 & 15.57 & 16.26 & 16.26 \\ 25.8 & 33.47 & 30.07 & 28.88 & 19.22 & 20.09 & 20.27 \\ 23.02 & 30.07 & 27.22 & 25.92 & 15.96 & 16.62 & 16.96 \\ 22.16 & 28.88 & 25.92 & 25.11 & 17.16 & 18.04 & 18.18 \\ 15.57 & 19.22 & 15.96 & 17.16 & 20.28 & 21.8 & 20.76 \\ 16.26 & 20.09 & 16.62 & 18.04 & 21.8 & 23.48 & 22.35 \\ 16.26 & 20.27 & 16.96 & 18.18 & 20.76 & 22.35 & 21.39 \end{bmatrix}$$

(b)

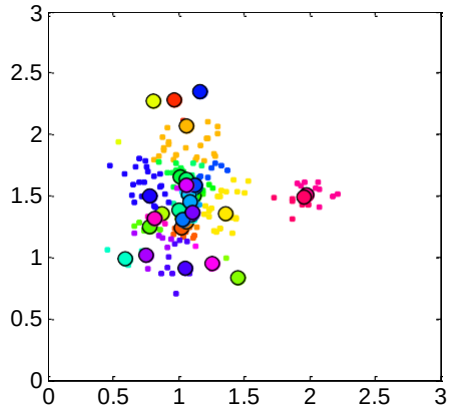


$$\mathbf{S}^5 = \begin{bmatrix} 68.58 & 88.24 & 78.39 & 76.48 & 56.88 & 59.83 & 59.54 \\ 88.24 & 113.95 & 101.56 & 98.77 & 71.37 & 75.02 & 74.95 \\ 78.39 & 101.56 & 90.94 & 87.90 & 60.76 & 63.70 & 64.00 \\ 76.48 & 98.77 & 87.90 & 85.83 & 63.20 & 66.56 & 66.37 \\ 56.88 & 71.37 & 60.76 & 63.20 & 65.08 & 69.63 & 67.11 \\ 59.83 & 75.02 & 63.70 & 66.56 & 69.63 & 74.59 & 71.81 \\ 59.54 & 74.95 & 64.00 & 66.37 & 67.11 & 71.81 & 69.34 \end{bmatrix}$$

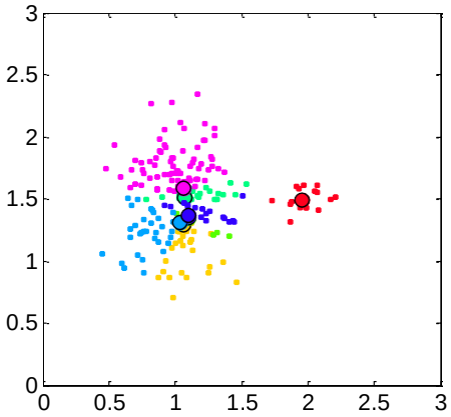
(c)

连通中心演化（CCE）

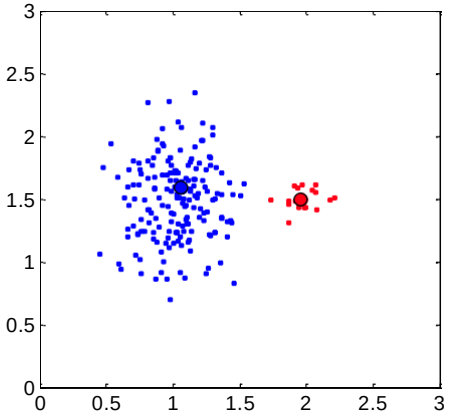
简单例子



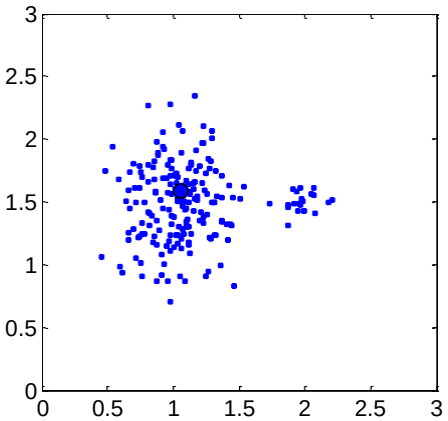
k=2



k=3



k=4~11



k>11

① 先计算数据的相似度矩阵S

② 计算S^2并根据前面的定义和规则确定聚类中心及分类

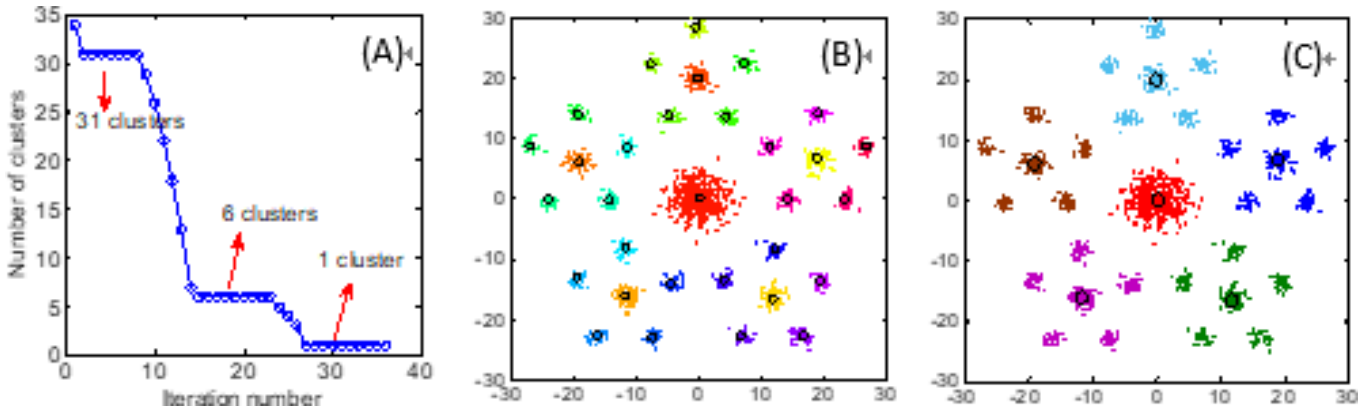
③ 计算S^3并根据前面的定义和规则确定聚类中心及分类

④ 计算S^3到S^{11}并给出聚类结果

⑤ K>11时，找到全局唯一的聚类中心

连通中心演化（CCE）

多个结果的例子



对于这块数据，利用CCE可以自动得到两个聚类结果

连通中心演化（CCE）

总结：

1. 介绍了一种新的数据中心确定机制连通中心演化（CCE）
2. 一定程度而言，CCE 是描述数据一阶统计信息的天然工具
3. CCE 将图论中顶点间途径条数的概念推广到实数情形，建立了连通度的概念
4. 基于相对连通度的概念，CCE 可以自动确定局部中心的位置和个数
5. CCE 能够提供数据从微观到宏观、从局部到整体的各尺度中心确定和类别分配结果
6. 连通度提供了一种衡量数据相似性的新的度量方式

思考：

1. 自适应参数？
2. 尺度自适应？
3. 分数阶次幂？
4. 计算复杂度？
5. 相似性度量？
6. 全自动聚类？

谢　谢

耿修瑞

中国科学院空天信息创新研究院

gengxr@sina.com.cn