电子科技大学
University of Electronic Science and Technology of China

基于同步原理的多边聚类算法研究

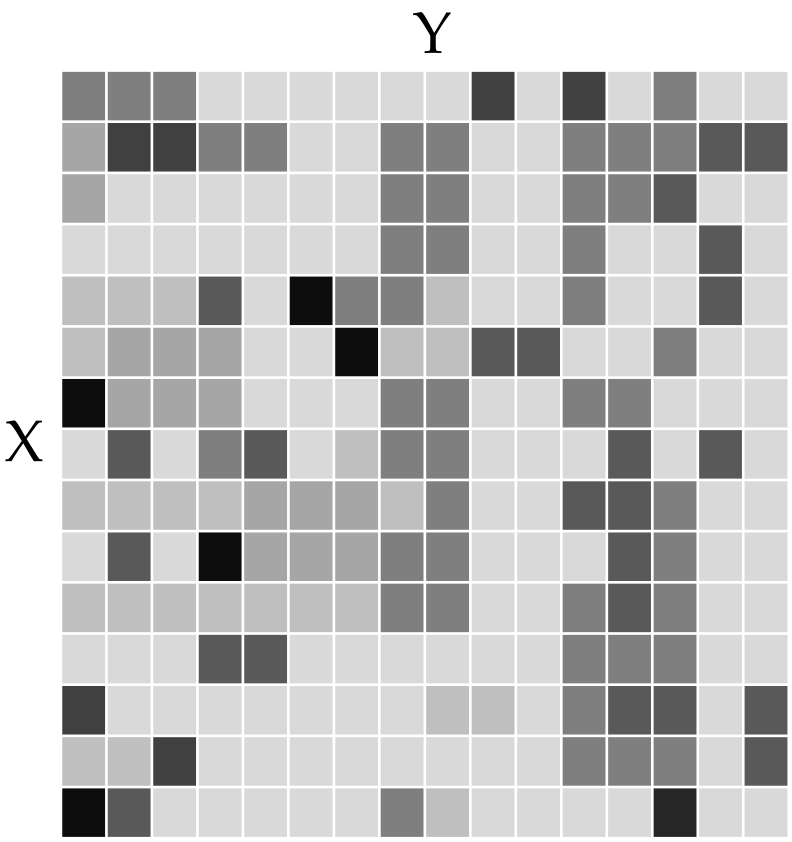
毕业答辩

英才实验学院
高崇铭（2012001010016）
指导老师：邵俊明

目录

1. 双边聚类背景与相关工作

背景与相关工作



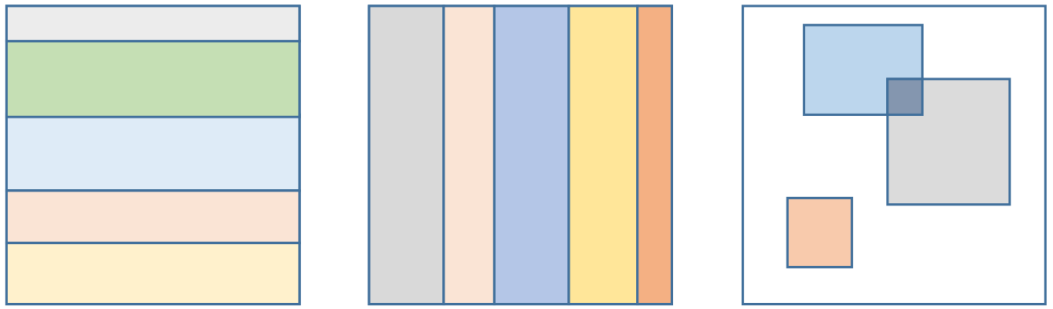
关联矩阵中X和Y含义:

推荐系统：商品与顾客

文本分析：文档与词汇

基因表达：基因与样本（主要）

Co-clustering思路：抓住X空间与Y空间的依赖性，同时对X空间与Y空间进行聚类。对应于关联矩阵，找寻行、列都相似的子矩阵，也称为联合簇(Bi-Cluster)



a. Row clustering　b. Column clustering　c. Co-clustering

背景与相关工作

- **国际上的主流算法：**

1. **基于Residue的方法：** Cheng and Church's Algorithm(2000) plaid(2002) , MSSRCC(2008) …
2. **基于图分割的方法：**二分图划分、谱聚类及其变种, (H. Zha, 2001), (Y. Kluger, 2003), (I. S. Dhillon, 2004), (B. Gao, 2005), Qubic(G. Li, 2009) …
3. **基于信息论的方法：** (R. El-Yaniv, 2001), ITCC (I. S. Dhillon, 2003), (A. Banerjee,2004), (Y. Song,2013) …

…

背景与相关工作

- **各自存在的主要缺陷：**

1. 多数算法是基于在行、列空间上分别划分而进行的，不能解决联合簇在原矩阵中随机分布的问题。
2. 需要人为指定联合簇的数目。
3. 只能处理二维矩阵的问题，对高维张量下的问题无能为力。
4. 参数的选取很敏感，参数对算法结果的影响很大。

…

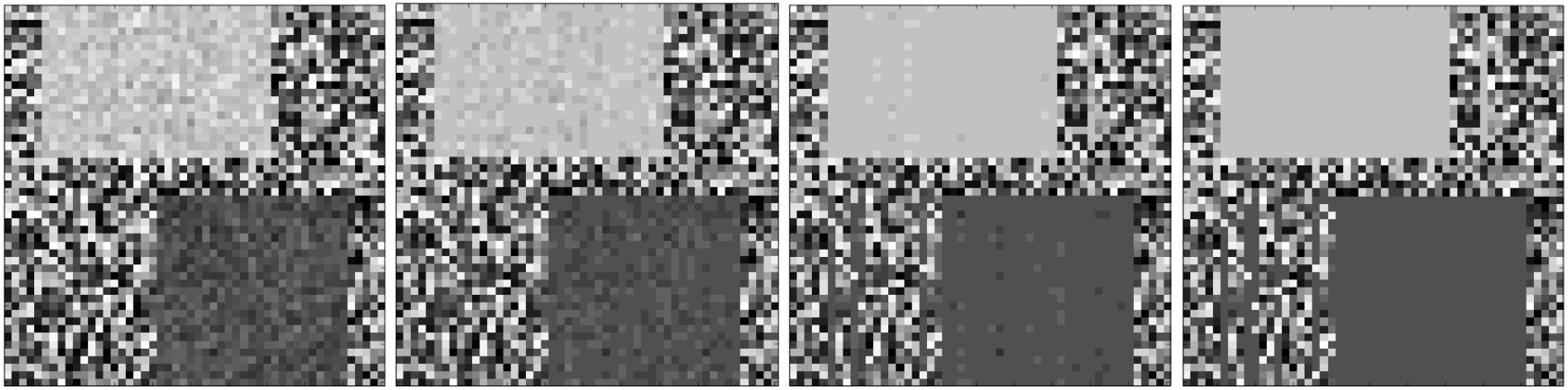
2. 基于同步思想的CoSync算法原理

(1) 动态双边交互模型

(2) 同值最大子矩阵找寻算法

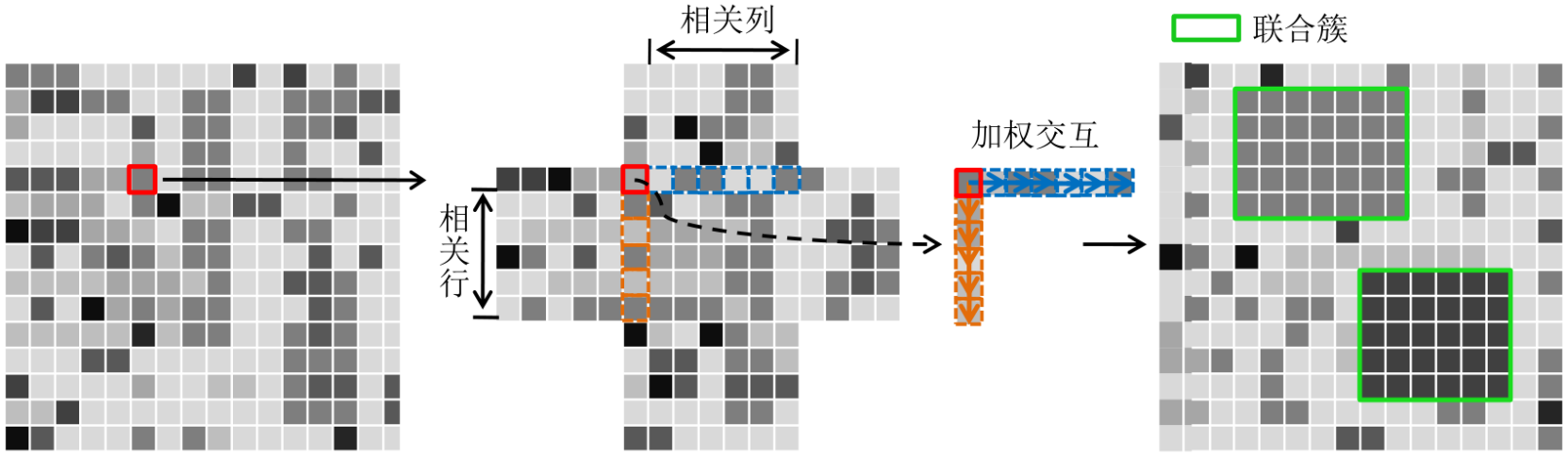
(3) 非负矩阵分解(NMF)

(1) 动态双边加权交互模型:



(a) 原始矩阵　(b) T = 1　(c) T = 5　(d) T = 30

算法运行演示图



(a) 原始矩阵　(b) 动态交互　(c) 结果矩阵

动态交互原理图

(1) 动态双边加权交互模型:

定义1(行或列的ϵ邻域邻居)：与该行或列距离小于阈值ϵ的行或列集合。

$$N_\epsilon^r(a_{i\cdot}) = \{a_{p\cdot} \big| dist(a_{p\cdot}, a_{i\cdot}) \le \epsilon, p \in I\}$$

$$N_\epsilon^c(a_{\cdot j}) = \{a_{\cdot q} \big| dist(a_{\cdot q}, a_{\cdot j}) \le \epsilon, q \in J\}$$

定义2(动态双边加权交互模型)：矩阵中每一个元素的当前值为上一时刻的值加上行与列的ϵ邻域邻居对其的加权影响值。

$$a_{ij}(t+1) = a_{ij}(t) + \frac{w^r(j)}{2|N_\epsilon^r(a_{i\cdot}(t))|} \cdot \sum_{a_{p\cdot} \in N_\epsilon^c(a_{i\cdot}(t))} \sin(a_{pj}(t) - a_{ij}(t))$$

$$+ \frac{w^c(i)}{2|N_\epsilon^c(a_{\cdot j}(t))|} \cdot \sum_{a_{\cdot q} \in N_\epsilon^r(a_{\cdot j}(t))} \sin(a_{iq}(t) - a_{ij}(t))$$

定义3(同步因子)：在每一次迭代中计算，若该值收敛，则停止模型迭代。

$$r = \frac{1}{2|I|} \sum_{i=1}^{|I|} \frac{1}{|N_\epsilon^r(a_{i\cdot})|} \sum_{a_{p\cdot} \in N_\epsilon^r(a_{i\cdot})} \mathbf{e}^{-||a_{p\cdot} - a_{i\cdot}||}$$

$$+ \frac{1}{2|J|} \sum_{i=1}^{|J|} \frac{1}{|N_\epsilon^c(a_{\cdot j})|} \sum_{a_{\cdot q} \in N_\epsilon^c(a_{\cdot j})} \mathbf{e}^{-||a_{\cdot q} - a_{\cdot j}||}$$

(2) 同值最大子矩阵搜索算法

问题定义：

$$A=\begin{pmatrix} 0.3 & 0.3 & 1.1 & 1.1 \\ 0.3 & 0.3 & 1.1 & 1.1 \\ 0.3 & 0.3 & 0.3 & 0.3 \\ 0.5 & 0.5 & 0.5 & 0.5 \\ 0.3 & 0.5 & 0.3 & 0.3 \\ 0.5 & 0.5 & 0.5 & 0.5 \end{pmatrix}$$

交互结果矩阵

交互结果矩阵中同值子块的搜寻问题可以写为以下形式：

$$\max_{I_S,J_S} \quad Size(B)=|I_S|\cdot|J_S|$$

$$s.t. \quad \begin{cases} A(I_S,J_S)=\pi \\ I_S\in I, J_S\in J \end{cases}$$

第一步：**(建立指示矩阵)** 对于交互结果矩阵中的每一个值π，建立一个指示矩阵$A^{(\pi)}$

最大联合簇

	a	b	c	d
1	1	1	0	0
2	1	1	0	0
3	1	1	1	1
4	0	0	0	0
5	1	0	1	1
6	0	0	0	0

(a) 指示矩阵$A^{(0.3)}$

	a	b	c	d
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	1	1	1	1
5	0	1	0	0
6	1	1	1	1

(b) 指示矩阵$A^{(0.5)}$

	a	b	c	d
1	0	0	1	1
2	0	0	1	1
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0

(c) 指示矩阵$A^{(1.1)}$

(2) 同值最大子矩阵搜索算法

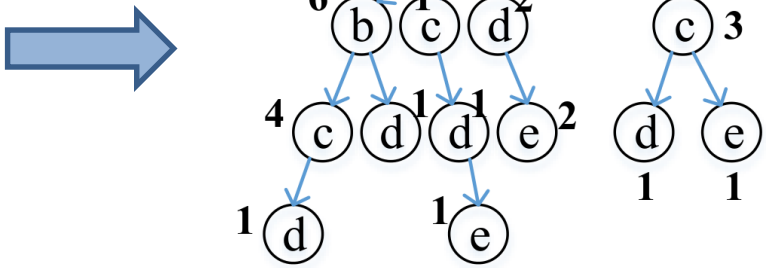
第二步：(建立FP-Tree) 对任一指示矩阵，首先统计每一列包含数字1的频数，之后将矩阵的列按照频数递减的顺序重新排序。根据排序后的指示矩阵，建立FP-tree。

	a	b	c	d	e
1	1	1	1	0	0
2	0	1	1	1	0
3	1	0	1	1	1
4	1	0	0	1	1
5	1	1	1	0	0
6	1	1	1	1	0
7	1	0	0	0	0
8	1	1	1	0	0
9	1	1	0	1	0
10	0	1	1	0	1
11	1	1	0	0	0
12	1	0	0	1	1
13	0	1	1	0	0
14	1	0	0	0	0

指示矩阵

列号	频数
a	11
b	9
c	8
d	6
e	4

列频数统计表

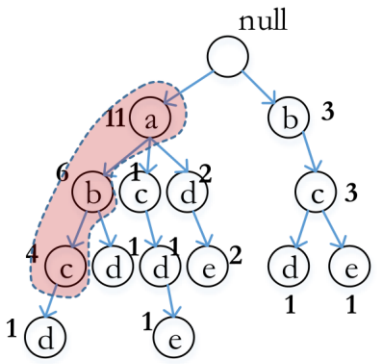


FP-tree

(2) 同值最大子矩阵搜索算法

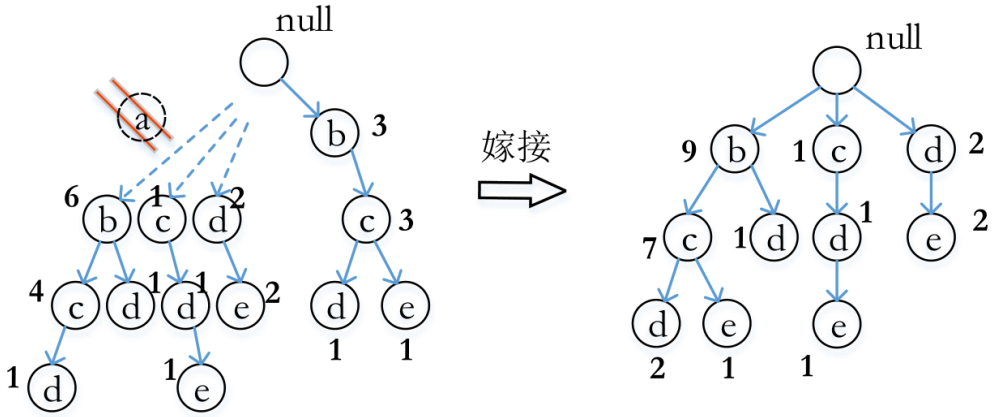
第三步: (遍历FP-tree) 用深度优先的方式遍历搜索FP-tree的一个分支，结束后进行嫁接步骤。迭代这个过程，直到FP-tree只剩下null节点。

- 搜索：



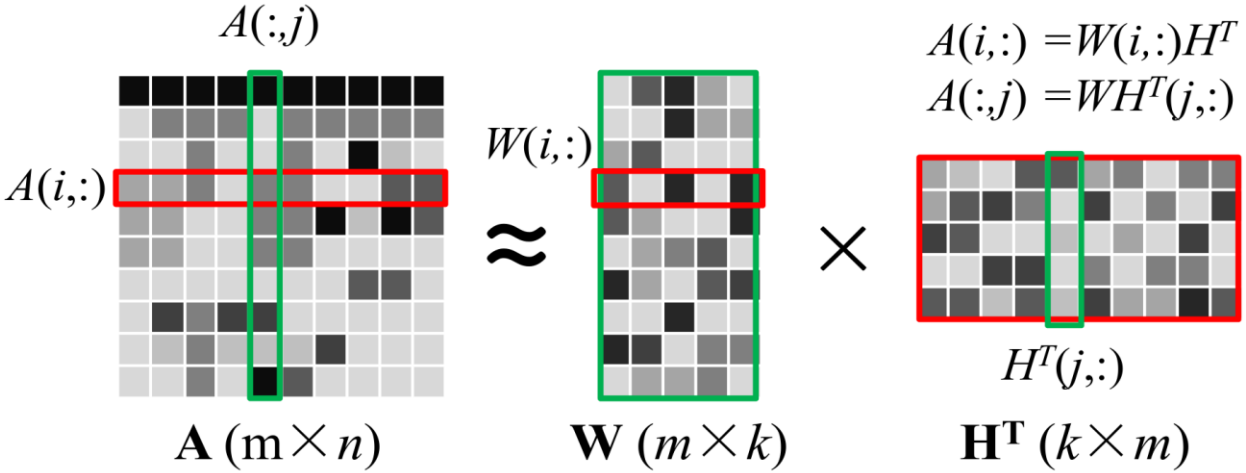
	a	b	c	d	e
1	1	1	1	0	0
2	0	1	1	1	0
3	1	0	1	1	1
4	1	0	0	1	1
5	1	1	1	0	0
6	1	1	1	1	0
7	1	0	0	0	0
8	1	1	1	0	0
9	1	1	0	1	0
10	0	1	1	0	1
11	1	1	0	0	0
12	1	0	0	1	1
13	0	1	1	0	0
14	1	0	0	0	0

- 嫁接：



(3) 高维数据下的非负矩阵分解(NMF)

目的：将高维数据进行降维，消除“高维诅咒”带来的影响。

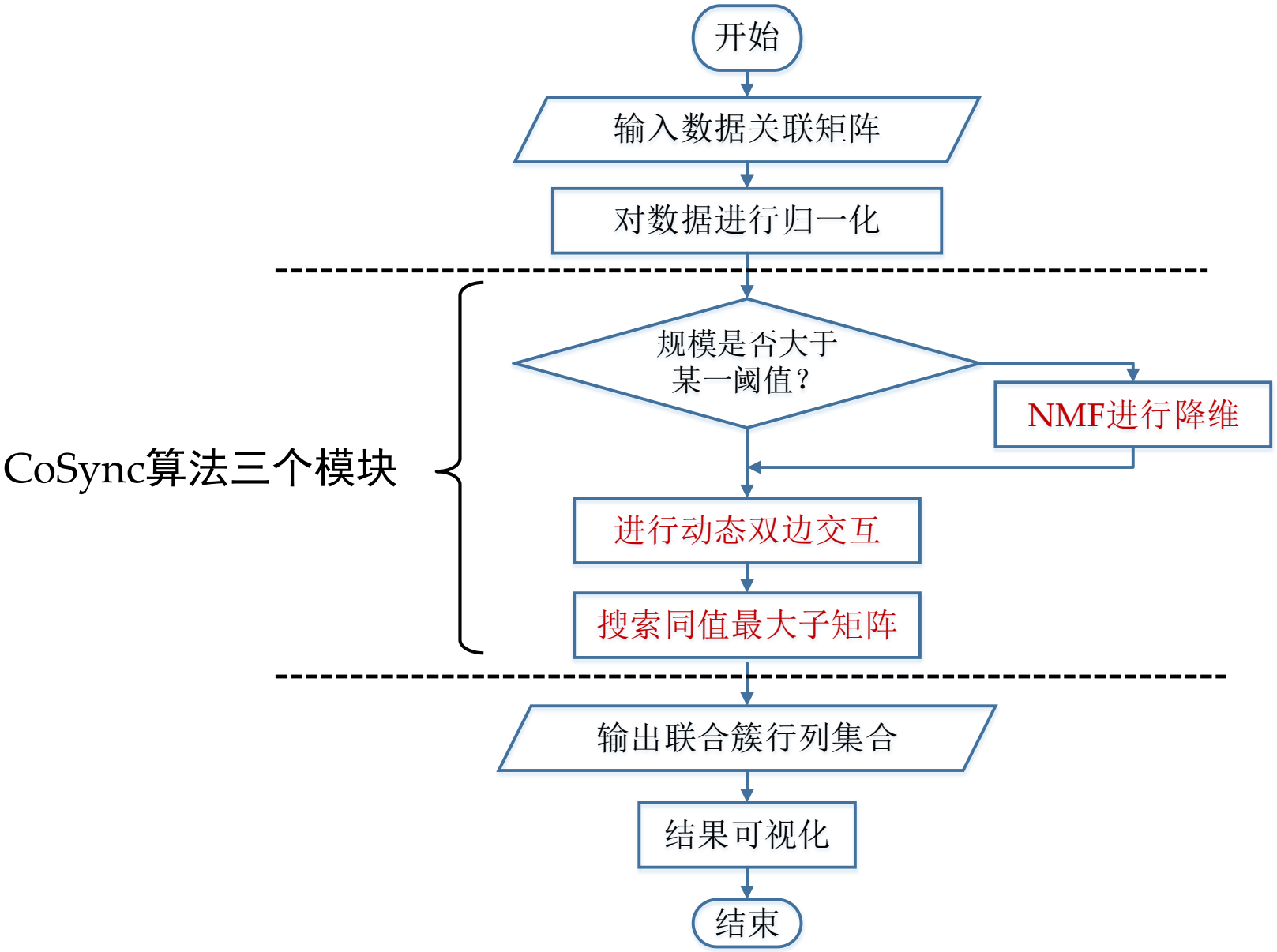


非负矩阵矩阵分解示意图

非负矩阵分解求解数学表达式：

$$\min_{\mathbf{W},\mathbf{H}} \quad f(\mathbf{W},\mathbf{H}) = ||\mathbf{A} - \mathbf{W}\mathbf{H}^T||_F^2$$
$$s.t. \quad \mathbf{W} \geq 0, \mathbf{H} \geq 0$$

CoSync 算法流程


开始
输入数据关联矩阵
对数据进行归一化
规模是否大于
某一阈值？
NMF进行降维
CoSync算法三个模块
进行动态双边交互
搜索同值最大子矩阵
输出联合簇行列集合
结果可视化
结束

3. CoSync 算法原理

Algorithm 1 CoSync

输入: 数据集矩阵A
$A = norm(A)$; //行或列归一化
if 矩阵规模超过一阈值 **then**
    [W, H] = NMF(A); //非负矩阵分解
end if
while 循环变量为1 **do**
    // 行交互
    for 每一个行向量 $a_{i\cdot} \in A$ **do**
        if 矩阵规模超过一阈值 **then**
            在矩阵 W 上寻找ϵ邻域邻居 $N_\epsilon^r(a_{i\cdot})$;
        else
            在矩阵 A 上寻找ϵ邻域邻居 $N_\epsilon^r(a_{i\cdot})$;
        end if
        for $N_\epsilon^r(a_{i\cdot})$中每一列 $j \in J$ **do**
            用(3-4)式来计算列权重因子$w(j)$;
        end for
        用(3-6)式来计算每一个ϵ邻域邻居 $a_{p\cdot} \in N_\epsilon^r(a_{i\cdot})$对$a_{i\cdot}$的交互值;
    end for
    // 列交互
    for 对每一个列向量 $a_{\cdot j} \in A$ **do**
        if 矩阵规模超过一阈值 **then**
            在矩阵 H 上寻找ϵ邻域邻居 $N_\epsilon^c(a_{\cdot j})$;
        else
            在矩阵 A 上寻找ϵ邻域邻居 $N_\epsilon^c(a_{\cdot j})$;
        end if
        for $N_\epsilon^c(a_{\cdot j})$中每一行 $i \in I$ **do**
            用(3-5)式来计算行权重因子$w(i)$;
        end for
        用(3-7)式来计算每一个ϵ邻域邻居 $a_{\cdot q} \in N_\epsilon^c(a_{\cdot j})$对$a_{\cdot j}$的交互值;
    end for
    用式(3-8)来更新矩阵 A;
    用式(3-9)来计算同步因子 r;
    if 同步因子 r 收敛 **then**
        循环变量设为0;
    end if
end while
//寻找联合簇
for 对于每一个离散值 π_k **do**
    用(3.3)节提出的算法搜寻其对应的最大子矩阵$B^{(\pi_k)}$;
end for
找到所有的联合簇 B;
输出: 联合簇矩阵B

3. 实验验证与评估

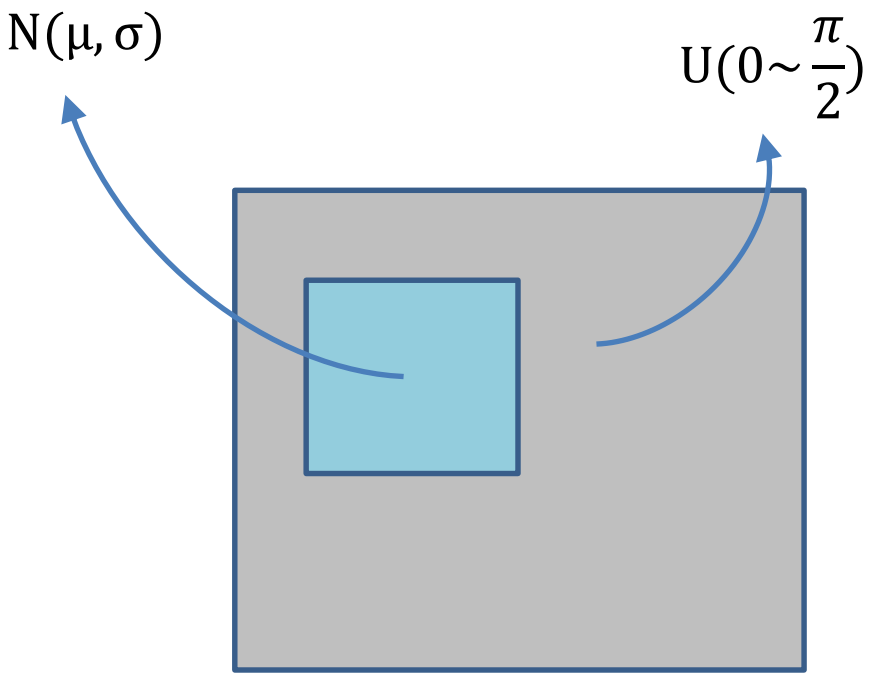
(1) 人工数据集

(2) 基因数据集

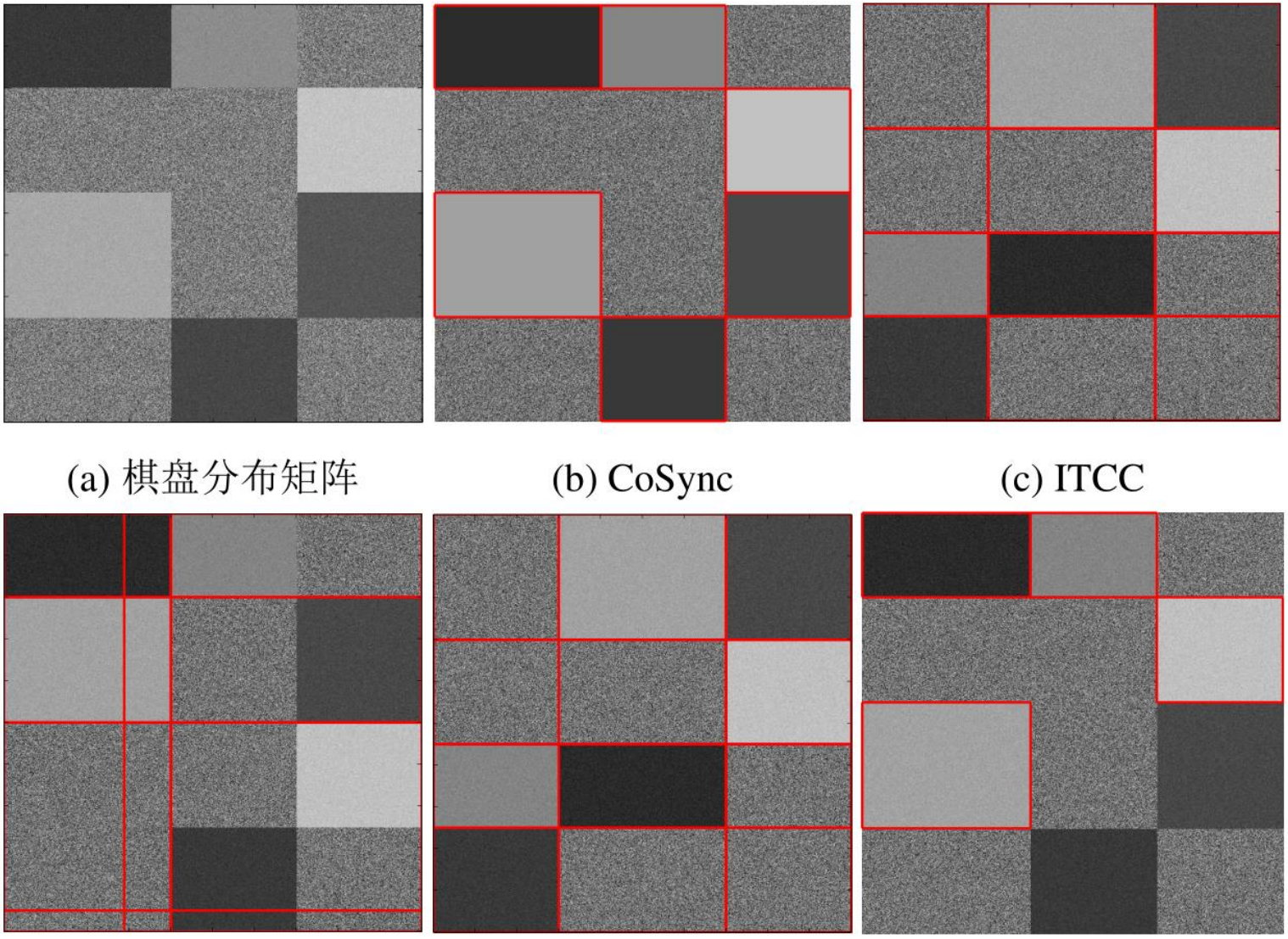
混合高斯人工数据集

数据集构造方案：

- 联合簇用高斯分布 $N(\mu, \sigma)$ 来刻画。
- 不相关部分用均匀分布$U(0\sim\frac{\pi}{2})$来刻画。



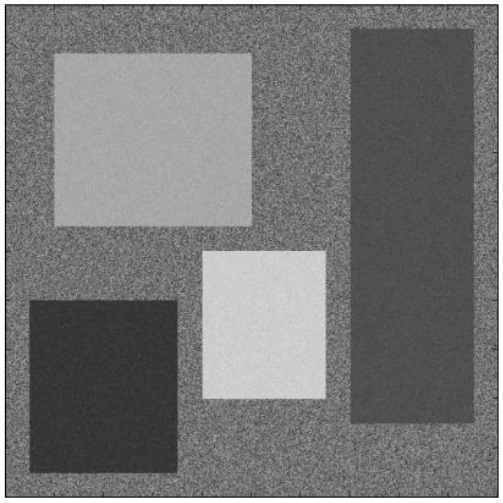
人工数据集：联合簇棋盘分布



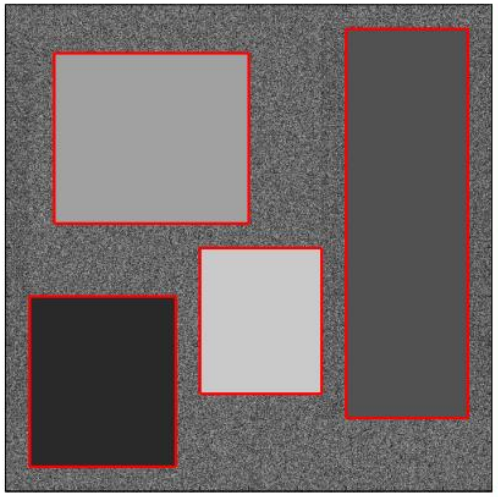
(a) 棋盘分布矩阵　(b) CoSync　(c) ITCC

(d) MSSRCC　(e) Spectral　(f) Plaid

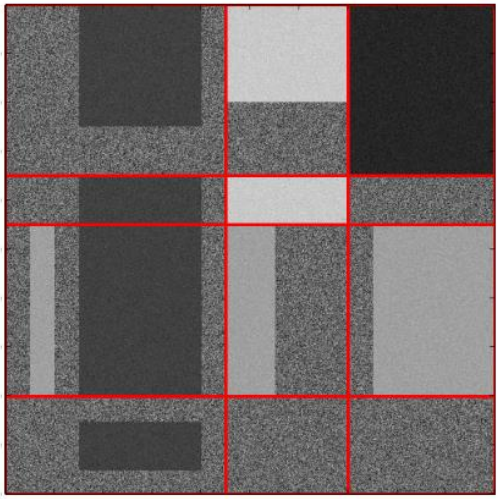
人工数据集：联合簇随机分布



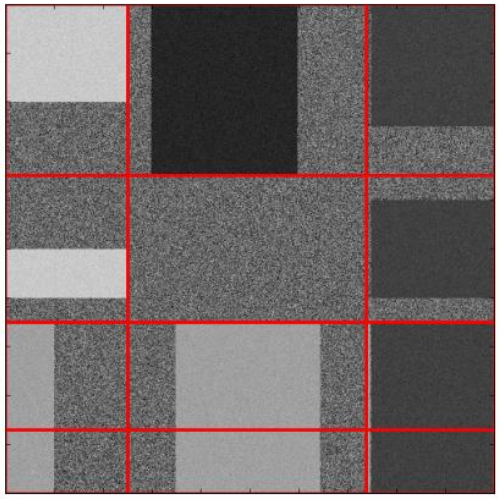
(g) 随机分布矩阵



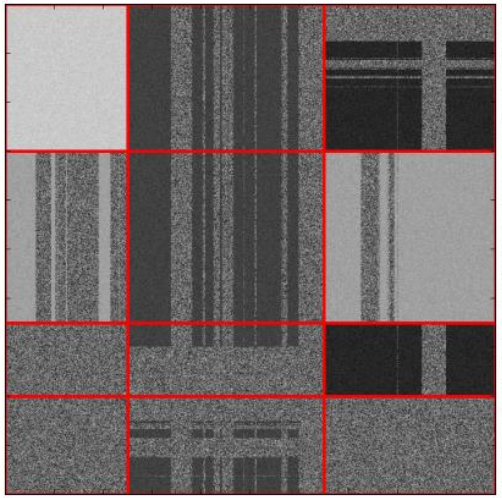
(h) CoSync



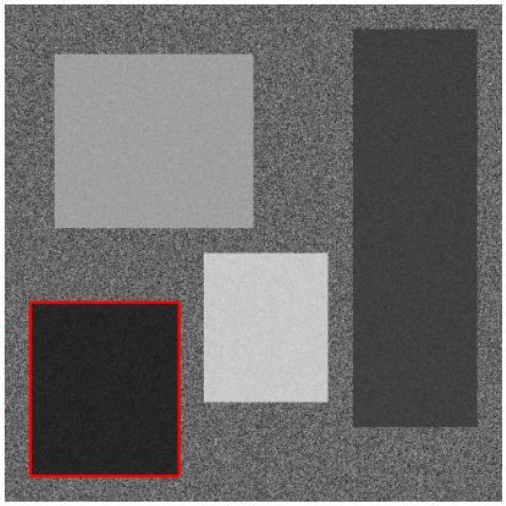
(i) ITCC



(j) MSSRCC



(k) Spectral



(l) Plaid

基因数据集上的实验评估

- **基因数据集说明：**

$$\text{基因}\begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1|J|} \\ b_{21} & b_{22} & \cdots & b_{2|J|} \\ \vdots & \vdots & \ddots & \vdots \\ b_{|I|1} & b_{|I|2} & \cdots & b_{|I||J|} \end{pmatrix}\text{样本}$$

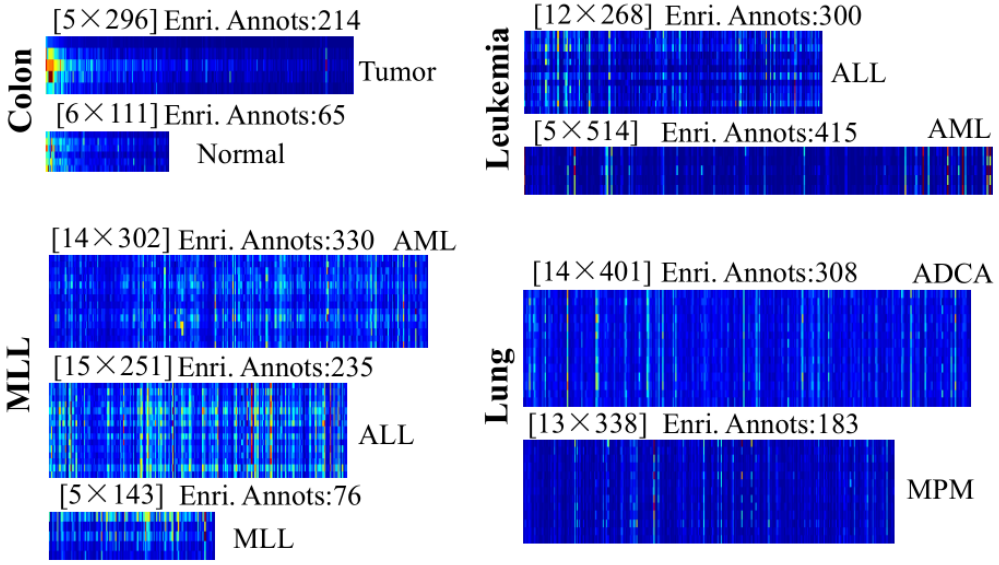
基因表达矩阵：矩阵中每个值代表某特定基因在某特定样本中的表达程度。

- **4个基因数据集信息：**

数据集	#基因数目	#样本数目	样本类别名称	样本分布
Colon	1096	62	Normal/Tumor	20/42
Leukemia	3571	72	ALL/AML	47/25
Lung	2401	181	ADCA/MPM	150/31
MLL	2474	72	ALL/AML/MLL	24/28/20

基因数据集上的实验评估

- CoSync在四个基因数据上找出的最大联合簇：



- 五种算法Precision和Recall对比结果：

	CoSync			Plaid			ITCC		MSSRCC		Spectral	
	#C	Pre.	Rec.	#C	Pre.	Rec.	#C	Avg.	#C	Avg.	#C	Avg.
Colon	11	0.95	0.66	4	0.71	0.66	2	0.82	2	0.86	2	0.73
Leukemia	28	0.96	0.71	2	0.81	0.43	2	0.95	2	0.93	2	0.74
Lung	23	1.00	0.96	3	1.00	0.50	2	0.85	2	0.99	2	1.00
MLL	23	0.99	0.67	4	0.88	0.88	3	0.83	3	0.93	3	0.64

4. 工作总结、讨论与未来展望

总结与讨论

➢ 本工作贡献与创新点

1. **（新视角）**提出了一种全新的双边聚类算法。
2. **（抓住数据本质）**基于自然界同步现象进行算法设计。
3. **（高性能）**对数据集中联合簇的分布没有限制，性能好于对比算法。
4. **（鲁班稳定）**对参数依赖程度较低。

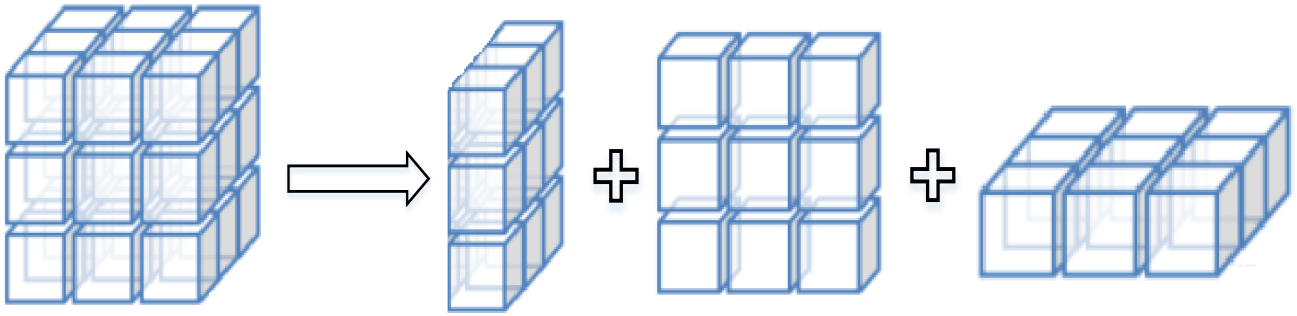
➢ 存在缺陷

1. 对比算法较少。
2. 较难挖掘数据集中嵌入的小而多联合簇。

未来工作展望：Multi-Sync

设想思路：

- **对于3维数据立方体：**分解为三个数据矩阵求解。



- **更高为维度的数据张量（tensor）：**分解为 C_d^2 个数据矩阵求解。

Thanks

英才实验学院

高崇铭

2012001010016

指导教师:邵俊明