第 24 卷 第 3 期 吉首大学学报(自然科学版) Vol. 24 No. 3
2003 年 9 月 Journal of Jishou University (Natural Science Edition) Sept. 2003

文章编号: 1007- 2985(2003)03- 0037- 05

模糊相似矩阵的构造*

王新洲, 舒海翅
(武汉大学测绘学院, 湖北 武汉 430079)

摘 要: 对现有的 13 种模糊相似矩阵构造方法进行全面比较, 提出了 3 条选择模糊相似矩阵构造方法的原则, 即正确性原则、不变性原则和可区分性原则. 结果显示, 只有绝对值倒数法完全满足这 3 条原则, 最宜用来构造模糊相似矩阵, 而最大最小法和算术平均最小法可作为检验方法.

关键词: 模糊聚类分析; 模糊相似矩阵; 整体分辨率; 类间分辨率
中图分类号: O159 文献标识码: A

聚类分析在地球空间信息科学研究中经常采用. 例如, 在变形分析中, 需确定稳定与不稳定点;[1] 在土地管理中, 为了对不同地块确定合理的定价, 需要对地块进行分级;[2,3] 在遥感图像处理中, 常常要对影像进行分类.[4] 分级、分类等实际是一个聚类问题. 由于类与类之间存在模糊性, 所以人们将模糊集理论引入分类, 从而产生了模糊聚类分析.

模糊聚类分析的方法很多, 其中用得较多的有传递闭包法、最大树法和动态直接聚类法.[5] 这些聚类法有一个共同点, 就是聚类依据是由原始数据所构造的模糊相似矩阵. 聚类正确与否, 完全取决于模糊相似矩阵. 尽管模糊相似矩阵在模糊聚类分析中起决定性作用, 但遗憾的是, 模糊相似矩阵的构造方法不唯一. 据不完全统计, 构造模糊相似矩阵的方法有 13 种之多.[6] 笔者在深入研究模糊相似矩阵的各种构造方法的基础上, 解决了 2 个问题: (Ⅰ) 对于同一个聚类问题, 13 种方法是否得到相同的聚类结果? (Ⅱ) 如果得不到相同的结果, 那么哪种方法最好?

1 模糊相似矩阵的构造方法及聚类结果分析

庄恒扬等[6] 列举了 13 种模糊相似矩阵构造方法 (1) 海明距离法 $r_{ij} = 1 - c\sum_{k=1}^{m}|x_{ik} - x_{jk}|$; (2) 欧氏距离法 $r_{ij} = 1 - c\sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$; (3) 切比雪夫距离法 $r_{ij} = 1 - c\max_{k=1,\cdots,m}|x_{ik} - x_{jk}|$; (4) 绝对值倒数法 $r_{ij} = \begin{cases} 1 & i = j, \\ \dfrac{c}{\sum_{k=1}^{m}|x_{ik} - x_{jk}|} & i \neq j; \end{cases}$ (5) 绝对值指数法 $r_{ij} = \exp(-c\sum_{k=1}^{m}|x_{ik} - x_{jk}|)$; (6) 指数相似系数法 $r_{ij} = \frac{1}{m}\sum_{k=1}^{m}\exp[-\frac{3}{4}(\frac{x_{ik} - x_{jk}}{S_k})^2]$, 其中 $S_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik} - x_k)^2}$, $x_k = \frac{1}{n}\sum_{i=1}^{n}x_{ik}$; (7) 兰氏距离法 $r_{ij} = 1 - c\sum_{k=1}^{m}\frac{|x_{ik} - x_{jk}|}{|x_{ik} + x_{jk}|}$; (8) 数量积法 $r_{ij} = \begin{cases} C & i = j, \\ \sum_{k=1}^{m}x_{ik}x_{jk} & i \neq j; \end{cases}$ (9) 夹角余弦法 $r_{ij} = \frac{\sum_{k=1}^{m}x_{ik}x_{jk}}{\sqrt{(\sum_{k=1}^{m}x_{ik}^2)(\sum_{k=1}^{m}x_{jk}^2)}}$; (10) 相关系数法 $r_{ij} = \frac{\sum_{k=1}^{m}(x_{ik} - x_i)(x_{jk} - x_j)}{\sqrt{(\sum_{k=1}^{m}(x_{ik} - x_i)^2)(\sum_{k=1}^{m}(x_{jk} - x_j)^2)}}$, 其中 $x_i = \frac{1}{m}\sum_{k=1}^{m}x_{ik}$, x_j

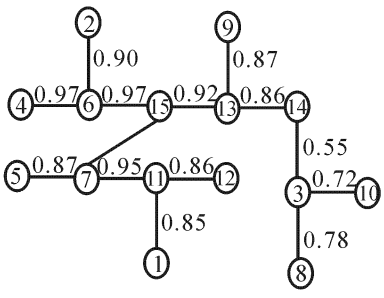
* 收稿日期: 2003- 06- 09
作者简介: 王新洲(1954-), 男, 湖北省黄陂人, 博士, 武汉大学测绘学院教授, 博士生导师, 主要从事测量数据处理理论与应用研究.

$= \frac{1}{m}\sum_{k=1}^{m} x_{jk}$; (11) 最大最小法 $r_{ij} = \frac{\sum_{k=1}^{m}\min(x_{ik}, x_{jk})}{\sum_{k=1}^{m}\max(x_{ik}, x_{jk})}$; (12) 算术平均最小法 $r_{ij} = \frac{\sum_{k=1}^{m}\min(x_{ik}, x_{jk})}{\frac{1}{2}\sum_{k=1}^{m}(x_{ik} + x_{jk})}$; (13) 几何平均最小法 r_{ij}

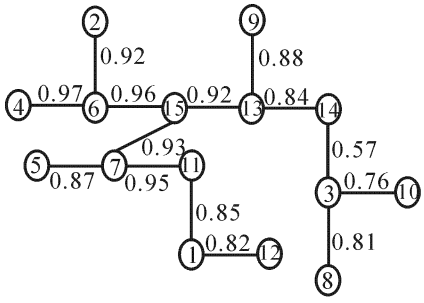
$= \frac{\sum_{k=1}^{m}\min(x_{ik}, x_{jk})}{\sum_{k=1}^{m}\sqrt{x_{ik}x_{jk}}}$. 为回答问题(Ⅰ), 笔者用这 13 种方法对同一聚类问题分别构造模糊相似矩阵, 进而用最大树法进行聚类. 设某聚类问题的标准化数据为

$$\boldsymbol{X}^{\mathrm{T}} = \begin{pmatrix} 0.27 & 0.65 & 2.37 & 0.45 & 0.30 & 0.50 & 0.20 & 2.72 & 1.02 & 3.58 & 0.25 & 0.72 & 0.77 & 0.72 & 0.47 \\ 0.78 & 0.70 & 2.35 & 1.03 & 0.37 & 0.87 & 0.21 & 2.43 & 1.32 & 3.01 & 0.08 & 0.25 & 0.45 & 0.58 & 0.58 \\ 0.40 & 0.02 & 3.21 & 0.08 & 0.48 & 0.29 & 0.02 & 4.32 & 0.77 & 3.20 & 0.06 & 0.08 & 0.96 & 0.76 & 0.36 \\ 1.07 & 0.36 & 3.48 & 0.09 & 0.36 & 0.09 & 0.09 & 4.64 & 0.13 & 2.01 & 0.36 & 2.05 & 0.11 & 0.09 & 0.09 \\ 2.17 & 0.99 & 1.43 & 0.53 & 0.92 & 0.68 & 0.77 & 1.73 & 0.64 & 0.79 & 1.14 & 1.01 & 1.12 & 0.42 & 0.68 \\ 0.95 & 1.55 & 0.48 & 1.17 & 0.10 & 1.21 & 1.05 & 1.33 & 0.51 & 1.24 & 1.11 & 1.27 & 1.08 & 0.76 & 1.21 \\ 0.83 & 0.82 & 0.47 & 1.25 & 0.26 & 1.35 & 0.83 & 1.20 & 0.73 & 1.41 & 0.94 & 1.09 & 0.99 & 0.72 & 1.09 \end{pmatrix}.$$

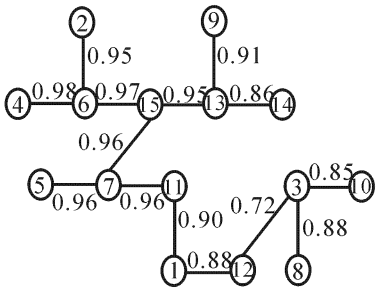
根据标准化矩阵 $\boldsymbol{X}$, 分别用 13 种方法构造模糊相似矩阵, 并用最大树法进行聚类, 结果用图形表示如图 1 所示, 用表格表示见表 1.



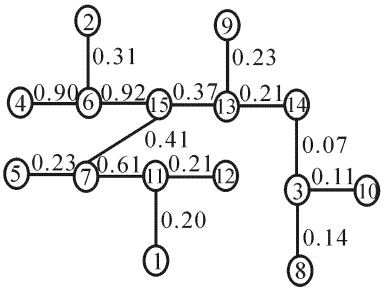
a 海明距离法



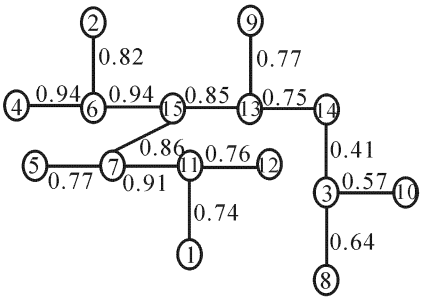
b 欧氏距离法



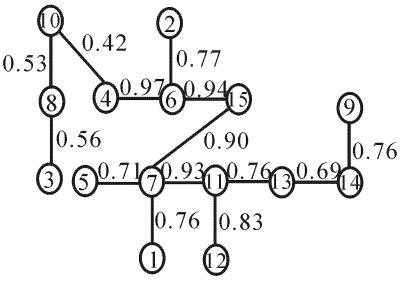
c 切比雪夫距离法



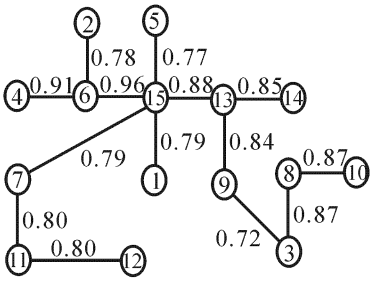
d 绝对值倒数法



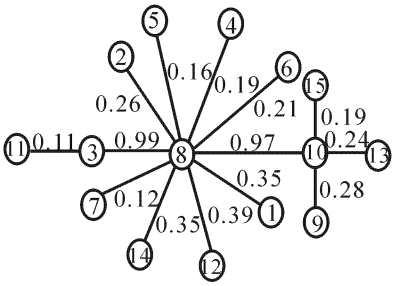
e 绝对值指数法



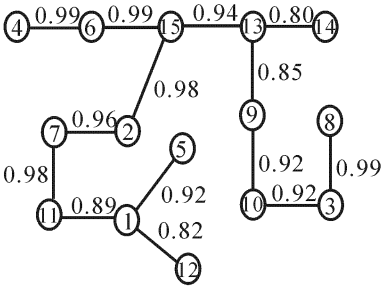
f 指数相似系数法



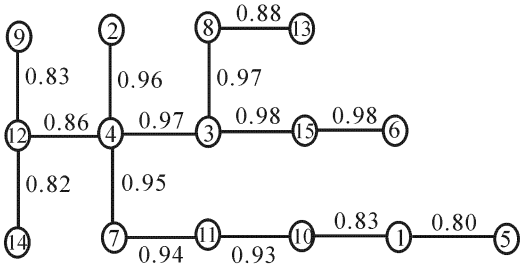
g 兰氏距离法



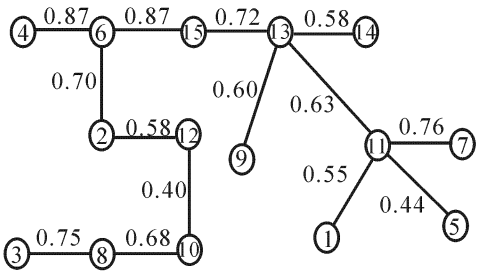
h 数量积法



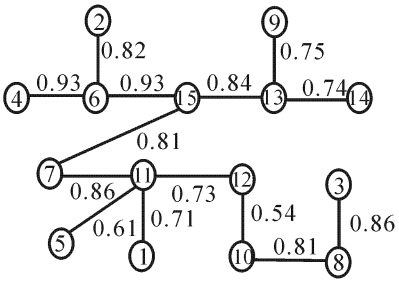
i 夹角余弦法



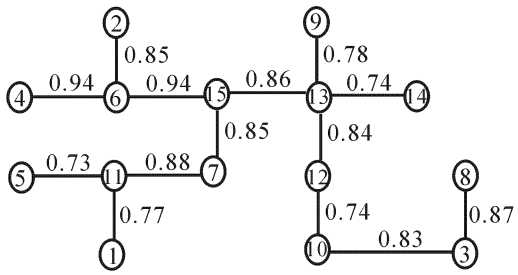
j　相关系数法



k　最大最小法



l　算术平均最小法



m　几何平均最小法

图 1　不同模糊相似矩阵构造方法的最大树法聚类结果

表 1　不同模糊相似矩阵构造方法的最大树法聚类结果

序号	方法	聚类结果			备注
		第 1 类	第 2 类	第 3 类	
1	海明距离法	3, 8, 10	4, 6, 7, 11, 15	1, 2, 5, 9, 12, 13, 14	先断 $r_{ij} \leqslant 0.6$, 再断 $r_{ij} \leqslant 0.92$
2	欧式距离法	3, 8, 10	4, 6, 7, 11, 15	1, 2, 5, 9, 12, 13, 14	先断 $r_{ij} \leqslant 0.6$, 再断 $r_{ij} \leqslant 0.92$
3	切比雪夫距离法	3, 8, 10	4, 5, 6, 7, 11, 15	1, 2, 9, 12, 13, 14	先断 $r_{ij} \leqslant 0.8$, 再断 $r_{ij} \leqslant 0.95$
4	绝对值倒数法	3, 8, 10	4, 6, 7, 11, 15	1, 2, 5, 9, 12, 13, 14	先断 $r_{ij} \leqslant 0.1$, 再断 $r_{ij} \leqslant 0.4$
5	绝对值指数法	3, 8, 10	4, 6, 7, 11, 15	1, 2, 5, 9, 12, 13, 14	先断 $r_{ij} \leqslant 0.5$, 再断 $r_{ij} \leqslant 0.86$
6	指数相似系数法	3, 8, 10	4, 6, 7, 11, 13, 15	1, 2, 5, 9, 12, 14	先断 $r_{ij} \leqslant 0.5$, 再断 $r_{ij} \leqslant 0.9$
7	兰氏距离法	3, 8, 10	4, 6, 9, 13, 14, 15	1, 2, 5, 7, 11, 12	砍断 $r_{ij} \leqslant 0.8$
8	数量积法	3, 8, 10	其余		砍断 $r_{ij} \leqslant 0.9$
9	夹角余弦法	3, 8, 10, 9	2, 4, 6, 7, 11, 13, 15	1, 5, 12, 14	砍断 $r_{ij} \leqslant 0.9$
10	相关系数法		2, 3, 4, 6, 7, 8, 10, 11, 15	1, 5, 9, 12, 13, 14	砍断 $r_{ij} \leqslant 0.9$
11	最大最小法	3, 8, 10	2, 4, 6, 7, 9, 11, 13, 15	1, 5, 12, 14	先断 $r_{ij} \leqslant 0.5$, 再断 $r_{ij} \leqslant 0.6$
12	算术平均最小法	3, 8, 10	2, 4, 6, 7, 11, 13, 15	1, 5, 9, 12, 14	先断 $r_{ij} \leqslant 0.6$, 再断 $r_{ij} \leqslant 0.8$
13	几何平均最小法	3, 8, 10	2, 4, 6, 7, 11, 12, 13, 14, 15	1, 5, 9	砍断 $r_{ij} \leqslant 0.8$

由表 1 可知, 方法 1 至方法 6 的聚类结果基本相同, 只是方法 3 将点 5 从第 3 类划归至第 2 类, 方法 6 将点 13 从第 3 类划归至第 2 类. 其余 7 种方法的分类结果差异较大, 尤其是方法 8 和方法 10, 几乎不能用于构造赖以分类的相似矩阵.

2　选择原则及聚类分辨率

2.1 选择构造方法的原则

问题(Ⅰ) 已通过实例分析得到答案. 要回答问题(Ⅱ), 就需要确定选择模糊相似矩阵构造方法的原则. 根据研究, 笔者

提出3条选择构造形似矩阵方法的原则:

(1) 正确性原则. 正确性原则是最一般的原则. 这条原则要求所选取的方法能客观地描述各样本之间的相似关系, 保证分类的正确性. (2) 不变性原则. 不变性原则主要是对各方法中常数 c 而言的. 在这13种方法中, 几乎50% 的方法都有1个任选常数 c, 以保证样本 i 与样本 j 的相似性在区间[0, 1] 上. 一个好的构造方法, 应能保证在选择不同的常数 c 时, 各样本之间相似关系的比例不变, 即不论 c 值如何变化, r_{ij} 与 r_{ik} 之比不变. (3) 可区分性原则. 可区分性原则是指不仅同一类样本能很好的聚在一起, 而且类与类之间界限明显, 能自然地分开, 即所选方法应具有很好的聚类分辨率.

2.2 聚类分辨率的定义

聚类分辨率越高, 类与类之间的界限就越明显, 这样就能保证1个样本只归于一类. 所以, 聚类分辨率越高, 可区分性就越好. 因此, 要判断相似矩阵构造方法的可区分性, 就要定义聚类分辨率. 笔者在最大树法聚类的基础上定义了2种聚类分辨率.

2.2.1 整体分辨率 对于具有 n 个样本的模糊聚类问题, 采用最大树法聚类时, 共有 $n-1$ 个连接边, 每个连接边上都有一个权值 r_{ij}, 这 $n-1$ 条边上权值 r_{ij} 的标准差为

$$\sigma = \pm \sqrt{\frac{(r_{ij} - r_{ij})^2}{n-2}},$$

其中 r_{ij} 为 $n-1$ 个权值 r_{ij} 的平均值. 标准差 σ 反映了 $n-1$ 个连接权的离散程度. 一般地, 离散程度越大, 类与类之间的分界就越明显. 因此, 将最大树的 $n-1$ 个连接权 r_j 的标准差 σ 定义为聚类的整体分辨率.

2.2.2 类间分辨率 整体分辨率只能从整体上粗略地判断可区分性. 由于聚类的结果是将 n 个样本聚为若干个类, 一种相似矩阵构造方法的可区分性如何, 主要是看类与类之间是否界限分明, 即类与类之间的分辨率如何. 为此, 定义类间分辨率如下:

在第 k 次分类时, 所分各类中最小的权与被砍断的所有边上最大的权之比, 定义为第 k 次分类时的类间分辨率, 即

$$\alpha_k = \frac{[\min(r_{ij}^S)]_k}{[\max(r_{ij}^D)]_k},$$

其中, $[\min(r_{ij}^S)]_k$ 表示在第 k 次分类时所分各类中最小的权, $[\max(r_{ij}^D)]_k$ 表示在第 k 次分类时被砍断的所有边上最大的权.

例如, 在图1-d中, 第1次分类时砍断权值 $r_{ij} \leqslant 0.1$ 的边得到2类. 第1类为(3, 8, 10), 第2类为其余各点. 在这2类中, 最小的权为3-10这条边上的权0.11, 即 $[\min(r_{ij}^S)]_1 = 0.11$. 第1次分类只砍断了1条边3-14, 而3-14边的权为0.07, 即 $[\max(r_{ij}^D)]_1 = 0.07$. 于是

$$\alpha_1 = \frac{[\min(r_{ij}^S)]_1}{[\max(r_j^D)]_1} = \frac{0.11}{0.07} = 1.57.$$

第2次分类时, 是在第2类中再砍断权值 $r_{ij} \leqslant 0.4$ 的边, 此时被砍断的边有7条, 即1-11, 2-6, 5-7, 9-13, 12-11, 13-15, 13-14. 在这7条边中, 最大的权值为 $r_{13,15} = 0.37$, 即 $[\max(r_{ij}^D)]_2 = r_{13,15} = 0.37$. 砍断这7条边后, 得7个点(1, 2, 5, 9, 12, 13, 14) 不相连, 而相连的5个点(4, 6, 7, 11, 15) 中连接权最小的为 $r_{7,15} = 0.41$, 即 $[\min(r_{ij}^S)]_2 = r_{7,15} = 0.41$. 于是

$$\alpha_2 = \frac{[\min(r_{ij}^S)]_2}{[\max(r_j^D)]_2} = \frac{0.41}{0.37} = 1.11.$$

其余各构造方法的整体分辨率和类间分辨率列于表2.

表2 不同模糊相似矩阵构造方法的整体分辨率及类间分辨率

序号	名 称	整体分辨率		类间分辨率	
		σ	r_{ij}	α_1	α_2
1	海明距离法	0.113	0.86	1.309	1.010
2	欧氏距离法	0.104	0.86	1.333	1.010
3	切比雪夫距离法	0.070	0.91	1.181	1.011
4	绝对值倒数法	0.273	0.35	1.571	1.108
5	绝对值指数法	0.147	0.77	1.390	1.012
6	指数相似系数法	0.168	0.76	1.333	1.096
7	兰氏距离法	0.060	0.83	1.013	

续表

序号	名　　称	整体分辨率		类间分辨率	
		σ	r_{ij}	α_1	α_2
8	数量积法				
9	夹角余弦法	0.064	0.93	1.034	
10	相关系数法	0.067	0.91	1.056	
11	最大最小法	0.14	0.65	1.545	1.034
12	算术平均最小法	0.111	0.78	1.593	1.110
13	几何平均最小法	0.064	0.84	1.064	

3　值得推荐的相似矩阵构造方法

根据选择原则及实例分析, 笔者认为方法 1 至方法 6 属于同一类方法, 均满足正确性原则; 而方法 7 至方法 10 以及方法 13 往往难以满足正确性原则, 分辨率又很低, 不宜作为相似矩阵的构造方法. 尤其是方法 8, 其分类结果受指标向量范数最大的那个样本的影响, 实际是以那个样本为核心, 将其余样本都与它进行比较, 根本不能很好地反映各样本之间的相似关系, 所以方法 8 应被排除. 在方法 1 至方法 6 中, 由于只有方法 4 完全满足 3 条选择原则, 且分辨率最高, 所以推荐选择方法 4 作为构造相似矩阵的方法. 至于方法 11 和方法 12, 虽然分辨率高且有不变性, 但由于与方法 4 不属于同类方法[6], 其分类结果有时与方法 4 的分类结果存在差异. 为了相互应证, 建议将方法 11 或方法 12 作为方法 4 的检验方法. 如果分类结果无显著差异, 表明分类结果完全正确, 如果分类结果有差异, 则以方法 4 的结果为准.

参考文献:

[1] 王新洲. 形变控制网稳定性的模糊聚类分析 [J]. 工程测量, 1986, (2): 10- 12.
[2] 王新洲, 王树良. 模糊综合法在土地定级中的应用 [J]. 武汉测绘科技大学学报, 1997, (1): 42- 46.
[3] 王新洲, 王树良. 土地评价中的模糊聚类分析 [J]. 武汉大学学报(自然科学版), 1997, 增刊(Ⅲ): 103- 107.
[4] 张景雄. 遥感影像的全模糊监督分类 [J]. 武汉测绘科技大学学报, 1998, (3): 211- 214.
[5] 李相镐, 李洪兴, 陈世权, 等. 模糊聚类分析及其应用 [M]. 贵阳: 贵州科技出版社, 1994.
[6] 庄恒扬, 沈新平, 陆建飞, 等. 模糊聚类计算方法的理论分析 [J]. 江苏农学院学报, 1998, (3): 37- 41.

Construction of Fuzzy Similar Matrix

WANG Xin-zhou, SHU Hai-chi
(College of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China)

Abstract: The construction of fuzzy similar matrix is not only the fundamental but also the most important work in fuzzy cluster's analysis. There are thirteen kinds of methods in the construction of the fuzzy similar matrix, which were put forward by many scholars for a long time. Without any standard for method choosing, people always choose their favorite methods, so there is some problems, such as whether those methods have the same cluster's result, which method is the best if the results are different. In order to answer those questions, three principles about how to choose the method of the construction of the fuzzy similar matrix are put forward in this paper, and then a complete comparison among all the construction methods is made, at last the best method used in the construction of the fuzzy similar matrix meeting the three principles is recommended.

Key words: fuzzy cluster's analysis; fuzzy similar matrix; whole resolution; cluster resolution