

MATLAB统计分析与应用

——聚类分析

主讲人：谢中华

SSstudy.com

科学软件学习网

主要内容

- 系统聚类法
- K均值聚类法
- 模糊C均值聚类法

第一节 系统聚类法

一、基本思想

开始时将 n 个样品或变量各自作为一类，并规定样品之间的距离和类与类之间的距离，然后将距离最近的两类合并成一个新类，计算新类与其它类之间的距离，重复进行两个最近类的合并，每次减少一类，直至所有的样品合并为一类。

二、系统聚类法的MATLAB函数

- pdist : 计算样品对之间的距离
- squareform : 将距离向量转为距离矩阵
- linkage : 创建系统聚类树
- dendrogram : 绘制聚类树形图
- cophenet : 计算cophenet相关系数
- inconsistent : 计算不一致系数
- cluster : 输出聚类结果
- clusterdata : 一步聚类，并输出聚类结果

三、样品聚类（Q型聚类）案例

【例13.1-1】 下表列出了2007年全国31个省、市和自治区的城镇居民家庭平均每人全年消费性支出的8个主要变量数据，这8个变量是

x_1 = 食品， x_2 = 衣着， x_3 = 居住， x_4 = 家庭设备用品及服务， x_5 = 医疗保健， x_6 = 交通和通信， x_7 = 教育文化娱乐服务， x_8 = 杂项商品和服务，

试利用系统聚类法对各地区进行聚类分析。

地 区	x1	x2	x3	x4	x5	x6	x7	x8
北 京	4934.05	1512.88	1246.19	981.13	1294.07	2328.51	2383.96	649.66
天 津	4249.31	1024.15	1417.45	760.56	1163.98	1309.94	1639.83	463.64
河 北	2789.85	975.94	917.19	546.75	833.51	1010.51	895.06	266.16
山 西	2600.37	1064.61	991.77	477.74	640.22	1027.99	1054.05	245.07
内蒙古	2824.89	1396.86	941.79	561.71	719.13	1123.82	1245.09	468.17
辽 宁	3560.21	1017.65	1047.04	439.28	879.08	1033.36	1052.94	400.16
吉 林	2842.68	1127.09	1062.46	407.35	854.8	873.88	997.75	394.29
黑龙江	2633.18	1021.45	784.51	355.67	729.55	746.03	938.21	310.67
上 海	6125.45	1330.05	1412.1	959.49	857.11	3153.72	2653.67	763.8
江 苏	3928.71	990.03	1020.09	707.31	689.37	1303.02	1699.26	377.37
浙 江	4892.58	1406.2	1168.08	666.02	859.06	2473.4	2158.32	467.52
安 徽	3384.38	906.47	850.24	465.68	554.44	891.38	1169.99	309.3
福 建	4296.22	940.72	1261.18	645.4	502.41	1606.9	1426.34	375.98
江 西	3192.61	915.09	728.76	587.4	385.91	732.97	973.38	294.6
山 东	3180.64	1238.34	1027.58	661.03	708.58	1333.63	1191.18	325.64
河 南	2707.44	1053.13	795.39	549.14	626.55	858.33	936.55	300.19
湖 北	3455.98	1046.62	856.97	550.16	525.32	903.02	1120.29	242.82
湖 南	3243.88	1017.59	869.59	603.18	668.53	986.89	1285.24	315.82
广 东	5056.68	814.57	1444.91	853.18	752.52	2966.08	1994.86	454.09
广 西	3398.09	656.69	803.04	491.03	542.07	932.87	1050.04	277.43
海 南	3546.67	452.85	819.02	519.99	503.78	1401.89	837.83	210.85
重 庆	3674.28	1171.15	968.45	706.77	749.51	1118.79	1237.35	264.01
四 川	3580.14	949.74	690.27	562.02	511.78	1074.91	1031.81	291.32
贵 州	3122.46	910.3	718.65	463.56	354.52	895.04	1035.96	258.21
云 南	3562.33	859.65	673.07	280.62	631.7	1034.71	705.51	174.23
西 藏	3836.51	880.1	628.35	271.29	272.81	866.33	441.02	335.66
陕 西	3063.69	910.29	831.27	513.08	678.38	866.76	1230.74	332.84
甘 肃	2824.42	939.89	768.28	505.16	564.25	861.47	1058.66	353.65
青 海	2803.45	898.54	641.93	484.71	613.24	785.27	953.87	331.38
宁 夏	2760.74	994.47	910.68	480.84	645.98	859.04	863.36	302.17
新 疆	2760.69	1183.69	736.99	475.23	598.78	890.3	896.79	331.8

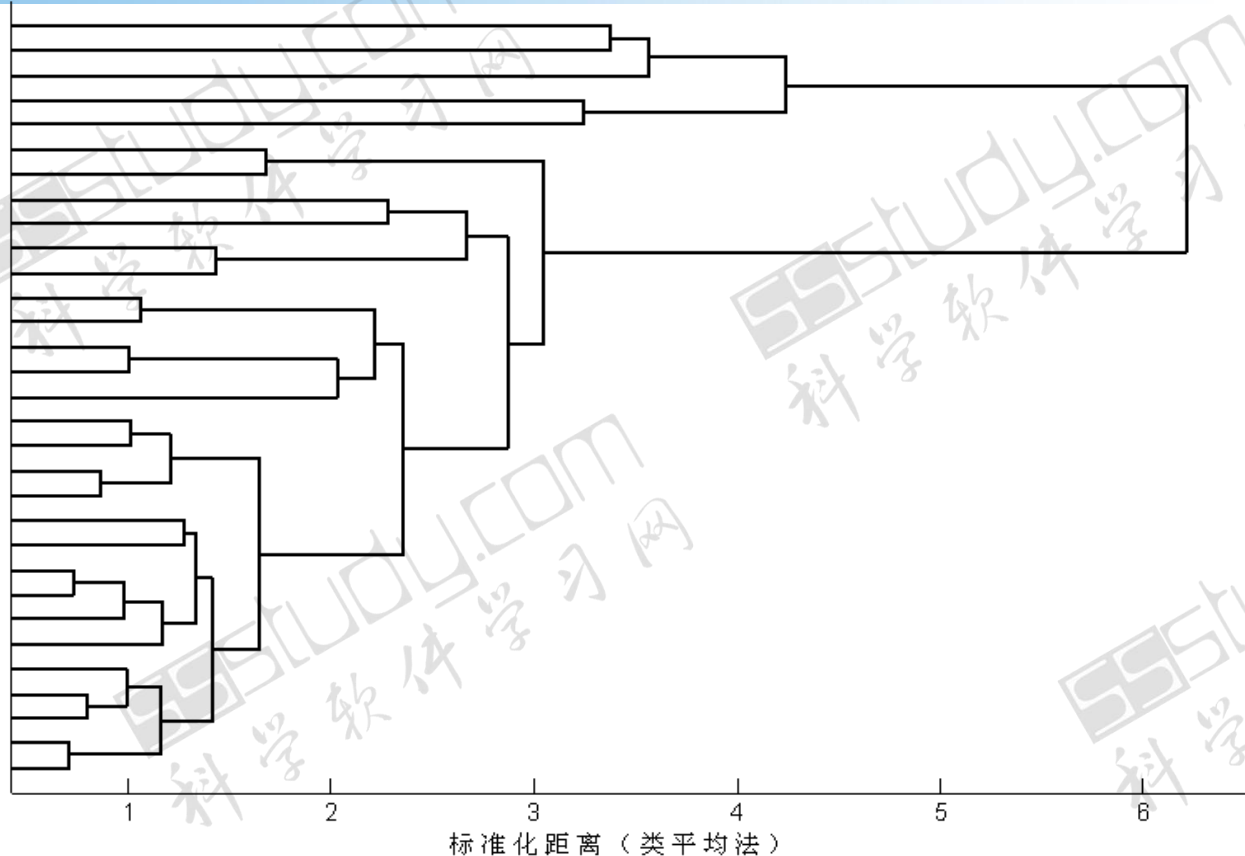
1. 一步聚类的MATLAB程序

```
>> [X,textdata] = xlsread('cluster1.xls');    % 读数据
>> X = zscore(X);    % 数据标准化
>> obslabel = textdata(2:end,1);    % 提取城市名称
% 利用类平均法将原始样品聚为3类
>> Taverage = clusterdata(X,'linkage','average','maxclust',3);
>> obslabel(Taverage == 1)    % 查看第1类所包含的城市
>> obslabel(Taverage == 2)    % 查看第2类所包含的城市
>> obslabel(Taverage == 3)    % 查看第3类所包含的城市
```

2. 分步聚类的MATLAB程序

```
>> y = pdist(X); % 计算样品间欧氏距离
>> Z = linkage(y,'average') % 创建系统聚类树
>> oblabel = textdata(2:end,1); % 提取城市名称
% 作出聚类树形图
>> H = dendrogram(Z,0,'orientation','right',...
'labels',oblabel);
>> set(H,'LineWidth',2,'Color','k');
>> xlabel('标准化距离 ( 类平均法 )');
```

东 江 津 海 京 建 苏 藏 南 西 林 宁 庆 东 古 川 北 州 西 疆 江 夏 南 西 北 南 西 徽 海 肃
 浙 天 上 北 福 江 西 云 海 广 吉 辽 重 山 蒙
 内 四 湖 贵 江 新 龙
 黑 宁 河 山 河 湖 陕 安 青 甘



3. 确定分类个数

- 阈值法
- 观察样品的散点图
- 使用统计量

(1) R^2 统计量

(2) 半偏 R^2 统计量

(3) 伪 F 统计量

(4) 伪 t^2 统计量

(5) 不一致系数

>> inconsistent0 = inconsistent(Z,40);

四、变量聚类（R型聚类）案例

【例13.1-2】在全国服装标准制定中，对某地区成年女子的14个部位尺寸（体型尺寸）进行了测量，根据测量数据计算得到14个部位尺寸之间的相关系数矩阵，如下表所列，试对14个变量进行聚类分析。

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
x_1 上体长	1												
x_2 手臂长	0.366	1											
x_3 胸围	0.242	0.233	1										
x_4 颈围	0.280	0.194	0.590	1									
x_5 总肩宽	0.360	0.324	0.476	0.435	1								
x_6 前胸宽	0.282	0.263	0.483	0.470	0.452	1							
x_7 后背宽	0.245	0.265	0.540	0.478	0.535	0.663	1						
x_8 前腰节高	0.448	0.345	0.452	0.404	0.431	0.322	0.266	1					
x_9 后腰节高	0.486	0.367	0.365	0.357	0.429	0.283	0.287	0.820	1				
x_{10} 总体长	0.648	0.662	0.216	0.316	0.429	0.283	0.263	0.527	0.547	1			
x_{11} 身高	0.679	0.681	0.243	0.313	0.430	0.302	0.294	0.520	0.558	0.957	1		
x_{12} 下体长	0.486	0.636	0.174	0.243	0.375	0.290	0.255	0.403	0.417	0.857	0.582	1	
x_{13} 腰围	0.133	0.153	0.732	0.477	0.339	0.392	0.446	0.266	0.241	0.054	0.099	0.055	1
x_{14} 臀围	0.376	0.252	0.676	0.581	0.441	0.447	0.440	0.424	0.372	0.363	0.376	0.321	0.627

MATLAB程序

```
[X,textdata] = xlsread('cluster2.xls'); % 读数据
```

```
y = 1 - X(tril(ones(size(X)),-1)>0)'
```

```
% 创建系统聚类树
```

```
Z = linkage(y,'average')
```

```
% 绘制聚类树形图
```

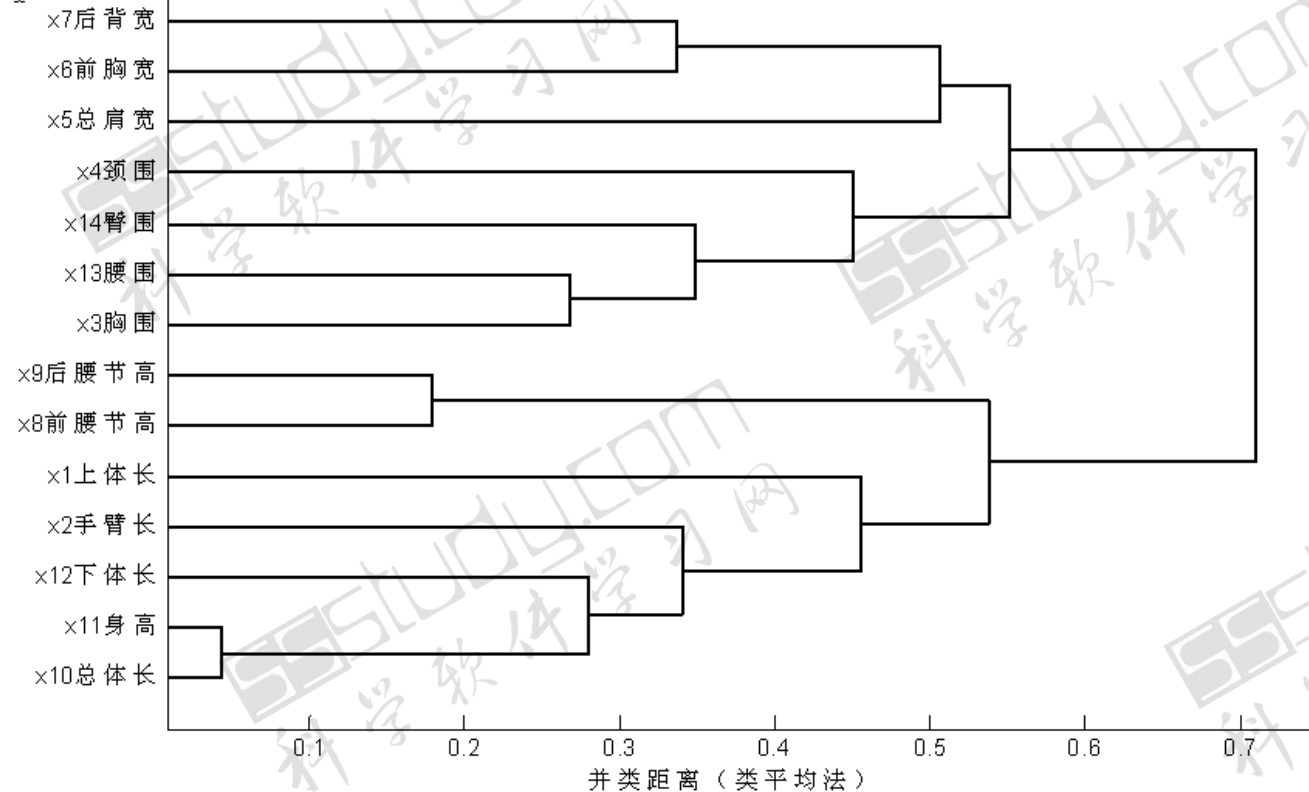
```
varlabel = textdata(2:end,1);
```

```
H = dendrogram(Z,0,'orientation','right','labels',varlabel);
```

```
set(H,'LineWidth',2,'Color','k');
```

```
xlabel('并类距离 ( 类平均法 ) ');
```

I



第二节 k均值聚类法

一、K均值聚类法的基本步骤

1. 选择 k 个样品作为初始凝聚点，或者将所有样品分成 k 个初始类，然后将这 k 个类的重心（均值）作为初始凝聚点。
2. 对除凝聚点之外的所有样品逐个归类，将每个样品归入凝聚点离它最近的那个类（通常采用欧式距离），该类的凝聚点更新为这一类目前的均值，直至所有样品都归了类。
3. 重复步骤2，直至所有的样品都不能再分配为止。

二、K均值聚类法的MATLAB函数

- kmeans : K均值聚类
- silhouette : 根据聚类结果绘制轮廓图

三、K均值聚类案例

【例13.2-1】下表列出了46个国家和地区3年（1990、2000和2006年）的婴儿死亡率和出生时预期寿命数据，数据来源：中华人民共和国国家统计局网站2008年国际统计数据。数据保存在文件kmeans.xls中，数据格式如表所示。本节将根据这些观测数据，利用K均值聚类法，对各国家和地区进行聚类分析。

国家和地区	婴儿死亡率(‰)	出生时平均预期寿命(岁)				
	1990年	2000年	2006年	1990年	2000年	2006年
中 国	36.3	29.9	20.1	68.9	70.3	72
中国香港				77.4	80.9	81.6
孟加拉国	100	66	51.6	54.8	61	63.7
文 莱	10	8	8	74.2	76.2	77.1
柬 埔 寨	84.5	78	64.8	54.9	56.5	58.9
印 度	80	68	57.4	59.1	62.9	64.5
印度尼西亚	60	36	26.4	61.7	65.8	68.2
伊 朗	54	36	30	64.8	68.9	70.7
以 色 列	10	5.6	4.2	76.6	79	80
日 本	4.6	3.2	2.6	78.8	81.1	82.3
哈萨克斯坦	50.5	37.1	25.8	68.3	65.5	66.2
朝 鲜	42	42	42	69.9	66.8	67
韩 国	8	5	4.5	71.3	75.9	78.5
老 挝	120	77	59	54.6	60.9	63.9
马来西亚	16	11	9.8	70.3	72.6	74
蒙 古	78.5	47.6	34.2	62.7	65.1	67.2
緬 甸	91	78	74.4	59	60.1	61.6
巴基斯坦	100	85	77.8	59.1	63	65.2
菲 律 宾	41	30	24	65.6	69.6	71.4
新 加 坡	6.7	2.9	2.3	74.3	78.1	79.9
斯里兰卡	25.6	16.1	11.2	71.2	73.6	75
泰 国	25.7	11.4	7.2	67	68.3	70.2
越 南	38	23	14.6	64.8	69.1	70.8
埃 及	66.7	40	28.9	62.2	68.8	71
尼日利亚	120	107	98.6	47.2	46.9	46.8
南 非	45	50	56	61.9	48.5	50.7
加 拿 大	6.8		4.9	77.4	79.2	80.4
墨 西 哥	41.5	31.6	29.1	70.9	74	74.5
美 国	9.4	6.9	6.5	75.2	77	77.8
阿 根 廷	24.7	16.8	14.1	71.7	73.8	75
巴 西	48.1	26.9	18.6	66.6	70.4	72.1
委内瑞拉	26.9	20.7	17.7	71.2	73.3	74.4
白俄罗斯	20.1	15	11.8	70.8		68.6
捷 克	10.9	4.1	3.2	71.4	75	76.5
法 国	7.4	4.4	3.6	76.7	78.9	80.6
德 国	7	4.4	3.7	75.2	77.9	79.1

.....

```
[X, textdata] = xlsread('kmeans.xls');
```

```
row = ~any(isnan(X), 2);
```

```
X = X(row, :); % 剔除缺失数据
```

```
countryname = textdata(3:end,1);
```

```
countryname = countryname(row);
```

```
X = zscore(X); % 数据标准化
```

```
% 选取第8、第27和第42个观测为初始凝聚点
```

```
startdata = X([8, 27, 42],:);
```

```
idx = kmeans(X,3,'Start',startdata);
```

```
[S, H] = silhouette(X,idx); % 绘制轮廓图
```

```
countryname(idx == 1) % 查看第1类所包含的国家或地区
```

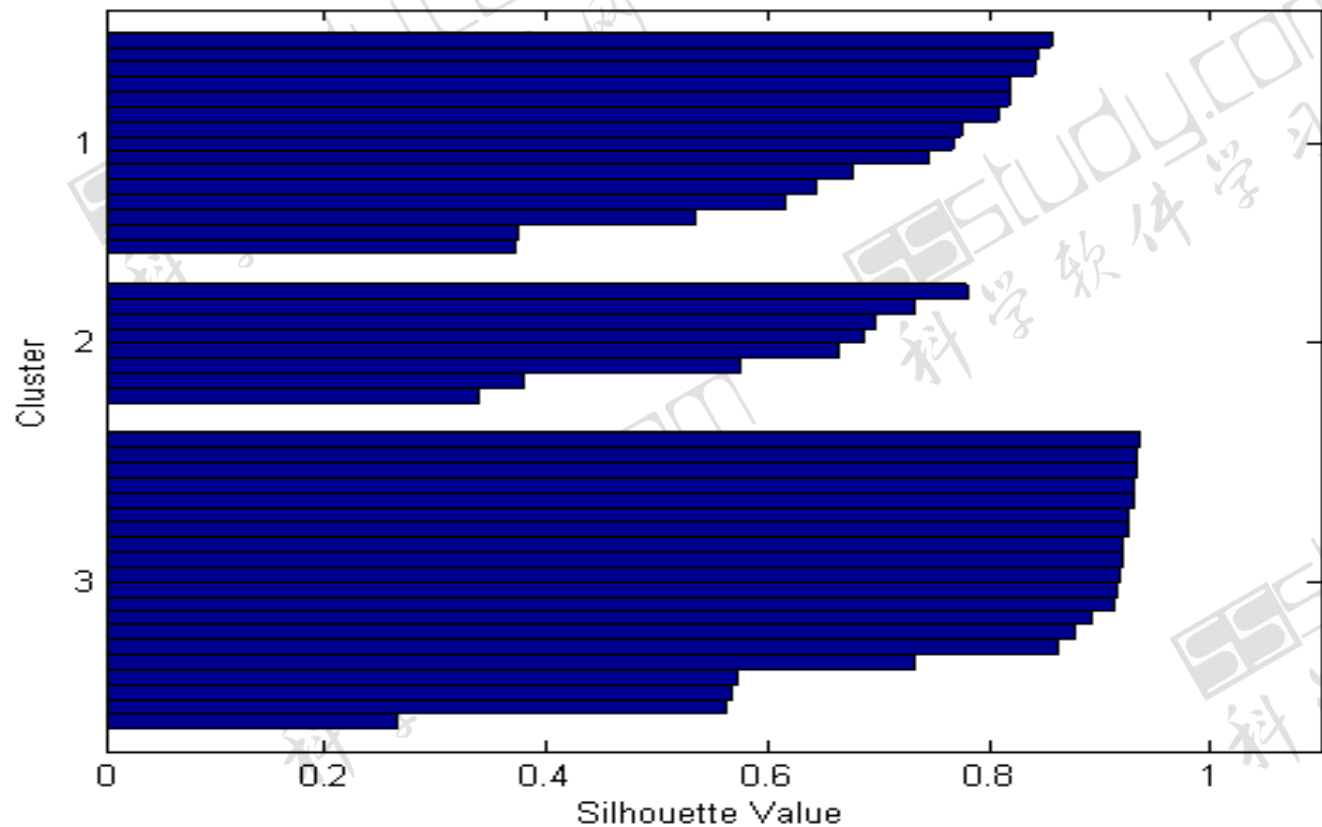
```
countryname(idx == 2) % 查看第2类所包含的国家或地区
```

```
countryname(idx == 3) % 查看第3类所包含的国家或地区
```

第三节 模糊C均值聚类法

一、基本思想

在很多分类问题中，分类对象之间没有明确的界限，往往具有亦此亦彼的表现，例如好与坏、高与矮之间没有明确的界限。在模糊划分中，每一个样品不是严格地划分为某一类，而是以一定的隶属度属于某一类。模糊C均值聚类法是通过求解一个优化问题，确定类中心坐标和隶属度矩阵。



二、模糊C均值聚类法的MATLAB函数

➤ fcm : 模糊C均值聚类

调用格式 :

```
[center,U,obj_fcn] = fcm(data,cluster_n)
```

```
[center,U,obj_fcn] = fcm(data,cluster_n,options)
```

三、模糊C均值聚类案例

【例13.3-1】下表列出了2006年我国31个省、市、自治区和直辖市的12个月的月平均气温数据，数据来源：中华人民共和国国家统计局网站，2007年《中国统计年鉴》。数据保存在文件fcm.xls中，数据格式如表所示。本节将根据这些观测数据，利用模糊C均值聚类法，对各地区进行聚类分析。

城	市	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
北	京	-1.9	-0.9	8.0	13.5	20.4	25.9	25.9	26.4	21.8	16.1	6.7	-1.0
天	津	-2.7	-1.4	7.5	13.2	20.3	26.4	25.9	26.4	21.3	16.2	6.5	-1.7
石	家 庄	-0.9	1.6	10.3	15.1	21.3	27.4	27.0	25.9	21.8	17.8	8.0	0.4
太	原	-3.6	-0.4	6.8	14.5	19.1	23.2	25.7	23.1	17.4	13.4	4.4	-2.5
呼	和 浩 特	-9.2	-7.0	2.2	10.3	17.4	21.8	24.5	22.0	16.3	11.5	1.3	-7.7
沈	阳	-12.7	-8.1	0.5	8.0	18.3	21.6	24.2	24.3	17.5	11.6	0.8	-6.7
长	春	-14.5	-10.6	-1.3	6.1	17.0	20.2	23.5	23.3	17.1	9.6	-2.3	-9.3
哈	尔 滨	-17.7	-12.6	-2.8	5.9	17.1	19.9	23.4	23.1	16.2	7.4	-4.5	-12.1
上	海	5.7	5.6	11.1	16.6	20.8	25.6	29.4	30.2	23.9	22.1	15.7	8.2
南	京	3.9	4.3	11.3	17.1	21.2	26.5	28.7	29.5	22.5	20.3	12.8	5.2
杭	州	5.8	6.1	12.4	18.3	21.5	25.9	30.1	30.6	23.3	21.9	15.1	7.7
合	肥	3.4	4.5	11.7	17.2	21.7	26.7	28.8	29.0	22.2	20.4	12.8	5.0
福	州	12.5	12.5	14.0	19.4	22.3	26.5	29.4	29.0	25.9	24.4	19.8	14.1
南	昌	6.6	6.5	12.7	19.3	22.7	26.0	30.0	30.0	24.3	22.1	15.0	8.1
济	南	0.0	2.1	10.2	16.5	21.5	26.9	27.4	26.0	21.4	19.5	10.0	1.6
郑	州	0.3	3.9	11.5	17.1	21.8	27.8	27.1	26.1	21.2	19.0	10.8	3.0
武	汉	4.2	5.8	12.8	19.0	23.9	28.4	30.2	29.7	24.0	21.0	14.0	6.8
长	沙	5.3	6.2	12.5	19.9	23.6	27.0	30.1	29.5	24.0	21.3	14.7	7.8
广	州	15.8	17.3	17.9	23.6	25.3	27.8	29.8	29.4	27.0	26.4	21.9	16.0
南	宁	14.3	14.3	17.5	23.9	25.2	27.6	28.0	27.2	25.7	25.6	20.4	14.0
海	口	18.5	20.5	21.8	26.7	28.3	29.4	30.0	28.5	27.4	27.1	25.3	20.8
重	庆	7.8	9.0	13.3	19.2	22.9	25.4	31.0	32.4	24.8	20.6	14.6	9.4
温	州	5.8	7.5	12.1	17.9	21.6	24.0	26.9	26.6	20.9	19.0	13.3	6.9
贵	阳	4.3	5.4	10.2	17.0	18.9	21.1	23.8	23.2	20.5	16.7	11.2	5.8
昆	明	10.8	13.2	15.9	18.0	18.0	20.4	21.3	20.6	18.3	16.9	13.2	9.8

.....

MATLAB程序

```
[xdata,textdata] = xlsread('fcm.xls');
```

```
city = textdata(4:end,1);
```

```
X = zscore(xdata);
```

```
options = [3, 200, 1e-6, 0];
```

```
% 调用fcm函数进行模糊C均值聚类
```

```
[center,U,obj_fcn] = fcm(X,3,options)
```

```
id1 = find(U(1,:) == max(U)); % 查找第1类中所有城市的序号
```

```
id2 = find(U(2,:) == max(U)); % 查找第2类中所有城市的序号
```

```
id3 = find(U(3,:) == max(U)); % 查找第3类中所有城市的序号
```

```
city(id1) % 查看第1类所包含的城市
```

```
city(id2) % 查看第2类所包含的城市
```

```
city(id3) % 查看第3类所包含的城市
```

Thank You