**9.4 试编程实现k均值算法，设置两组不同的k值、三组不同初始中心点，在西瓜数据集4.0上进行实验比较，并讨论什么样的初始中心有利于取得好结果。**

**①问题描述：**

本题利用西瓜数据集4.0对k均值算法中不同的k值和不同初始中心点的分类结果进行测试，从而分析什么样的初始中心有利于取得较好的结果。西瓜数据集4.0仅提供西瓜的密度和含糖率属性值，可作为无标签数据使用，因此该k均值算法为无监督学习，改变分类簇的个数和初始均值向量的相对位置，能够使分类的结果产生较大差别。

**②实现过程：**

a.将数据集中的密度及含糖率数据写死在代码内的函数中供后续调用。

b.设定好目标簇的个数和初始均值向量的位置。

c.进入主循环，遍历数据中的每个样本，计算每个样本与每个均值向量的距离，并将该样本分进距离该样本最近的均值向量所属的簇。

d.将初步分类后的每个簇分别计算其均值向量，若与上一代的均值向量不同则更新均值向量，若所有均值向量都未被更新，则结束主循环，返回已分类好的簇和均值向量。

e.打印出簇的内容及均值向量的坐标。

f.绘制密度和含糖率数据样本分类簇的效果图。

**③实现结果：**

***初始均值向量分布适中***

***k=3, u1=(0.403;0.237), u2=(0.343;0.099),u3=(0.532;0.472)***

Cluster result:

{

2: [0, 1, 2, 3, 20, 21, 23, 24, 25, 26, 27, 28, 29],

0: [4, 5, 6, 7, 8, 9, 12, 13, 14, 16, 17, 18, 19, 22],

1: [10, 11, 15]

}

Mean Vector result:

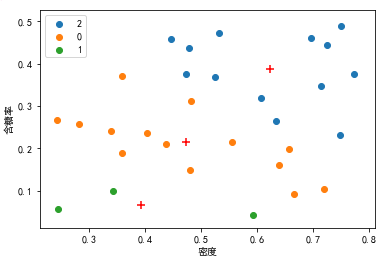
[

[0.47314286 0.21428571]

[0.39366667 0.066 ]

[0.62346154 0.38792308]

]



***初始均值向量分布分散***

***k=3,u1=(0.243;0.267), u2=(0.719;0.103),u3=(0.725;0.445)***

Cluster result:

{

2: [0, 1, 3, 21, 23, 24, 25, 26, 27, 28, 29],

1: [2, 4, 8, 12, 13, 15, 16, 20],

0: [5, 6, 7, 9, 10, 11, 14, 17, 18, 19, 22]

}

Mean Vector result:

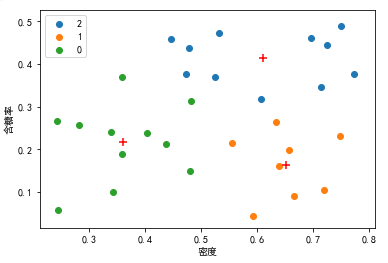
[

[0.36136364 0.21709091]

[0.6515 0.16325 ]

[0.61118182 0.41336364]

]



***初始均值向量分布密集***

***k=3,u1=(0.639;0.161), u2=(0.657;0.198),u3=(0.634;0.264)***

Cluster result:

{

1: [0, 1, 2, 3, 4, 8, 12, 13, 15, 16, 20, 21, 25, 28],

0: [5, 6, 7, 9, 10, 11, 17, 18, 19],

2: [14, 22, 23, 24, 26, 27, 29]

}

Mean Vector result:

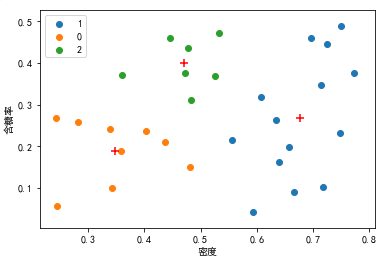
[

[0.348 0.18955556]

[0.67721429 0.26714286]

[0.471 0.39928571]

]



***初始均值向量分布适中***

***k=5,u1=(0.403;0.237), u2=(0.343;0.099),u3=(0.532;0.472),u4=(0.608;0.318),u5=(0.748;0.232)***

Cluster result:

{

3: [0, 1, 3, 21, 25, 28],

4: [2, 4, 8, 12, 13, 15, 16, 20],

0: [5, 6, 7, 9, 14, 17, 18, 19],

1: [10, 11],

2: [22, 23, 24, 26, 27, 29]

}

Mean Vector result:

[

[0.363 0.24 ]

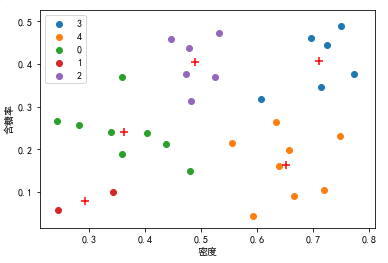
[0.294 0.078 ]

[0.4895 0.40416667]

[0.7115 0.40566667]

[0.6515 0.16325 ]

]



***初始均值向量分布分散***

***k=5,u1=(0.243;0.267), u2=(0.719;0.103),u3=(0.725;0.445),u4=(0.360;0.370),u5=(0.525;0.369)***

Cluster result:

{

2: [0, 1, 21, 25, 28],

1: [2, 4, 8, 12, 13, 15, 16, 20],

4: [3, 22, 23, 24, 26, 27, 29],

3: [5, 6, 7, 14],

0: [9, 10, 11, 17, 18, 19]

}

Mean Vector result:

[

[0.30183333 0.18483333]

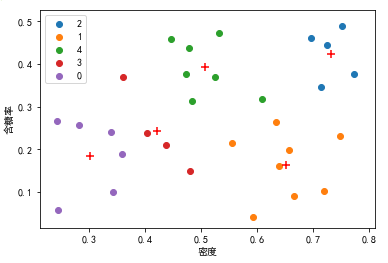
[0.6515 0.16325 ]

[0.7322 0.4232 ]

[0.42025 0.24175 ]

[0.50642857 0.39185714]

]



***初始均值向量分布密集***

***k=5,u1=(0.639;0.161), u2=(0.657;0.198),u3=(0.634;0.264),u4=(0.525;0.369),u5=(0.608;0.318)***

Cluster result:

{

4: [0, 1, 21, 25, 28],

2: [2, 3, 4],

0: [5, 6, 7, 9, 10, 11, 17, 18, 19],

1: [8, 12, 13, 15, 16, 20],

3: [14, 22, 23, 24, 26, 27, 29]

}

Mean Vector result:

[

[0.348 0.18955556]

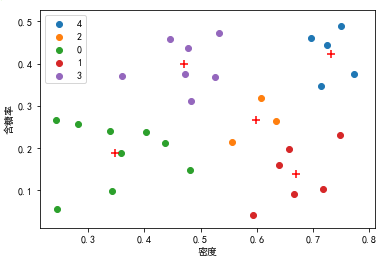
[0.67033333 0.13783333]

[0.59933333 0.26566667]

[0.471 0.39928571]

[0.7322 0.4232 ]

]



**④结果分析：**

分别将簇的个数设置为3和5，初始均值向量分别选取分布适中、分布分散、分布密集三种情况进行结果的分析。仅仅简单地增加分类簇的个数，在初始均值向量分布基本类似的情况下无法保证簇个数多的分类效果一定好于簇分类个数少的结果。当初始均值向量分布较密集时，分类结果相对较差；当初始均值向量分布较密集时，分类结果相对较好；当初始均值向量分布适中时，分类结果也适中。当初始均值向量分布较密集时，最终的均值向量往往距离样本数据的距离较远；当初始均值向量分布较分散时，最终的均值向量往往距离样本数据的距离较近。因此k-means算法的好坏与初始样本的选取有很大关系。