**k-means++**

1. 模型的建立

（1）模型概述

Kmeans++是基于Kmeans的改进算法，基本原理先是使用K个聚类中心，将所有样本划分为K个不同的簇。，流程图

该问题数据量比较大，Kmeans++能够快速收敛，Kmeans是个简单高效的聚类方法，因此采用Kmeans++

（2）聚类中心的选取

Step1：**随机选取一个样本作为第一个聚类中心 c1**

Step2：

**计算每个样本与最近一个聚类中心的距离，用 D(x)表示；**

**D（x）=欧氏距离**

**这个值越大，表示被选取作为聚类中心的概率p较大；**

**P=**

**最后，用轮盘法选出下一个聚类中心；**

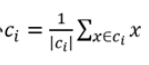
Step3：**重复步骤二，知道选出 k 个聚类中心**。

（3）Kmeans聚类

Step1: 对数据集中的每个Xi，计算它到K个聚类中心欧氏距离，将其归到距离最小的聚类中心所对应的类中

Dis=欧氏距离

Step2：针对每个类别Ci，更新其聚类中心为该类的重心



Step3:重复Step1和Step2，直到聚类中心的位置收敛或者达到规定的迭代次数，输出聚类结果

2.模型的求解

根据数据特征，具体问题具体分析，将样本数据聚类为K类，将K作为输入，根据上述流程进行聚类，最终聚类结果如下表所示

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

如果维度不超过三维可以打个散点图

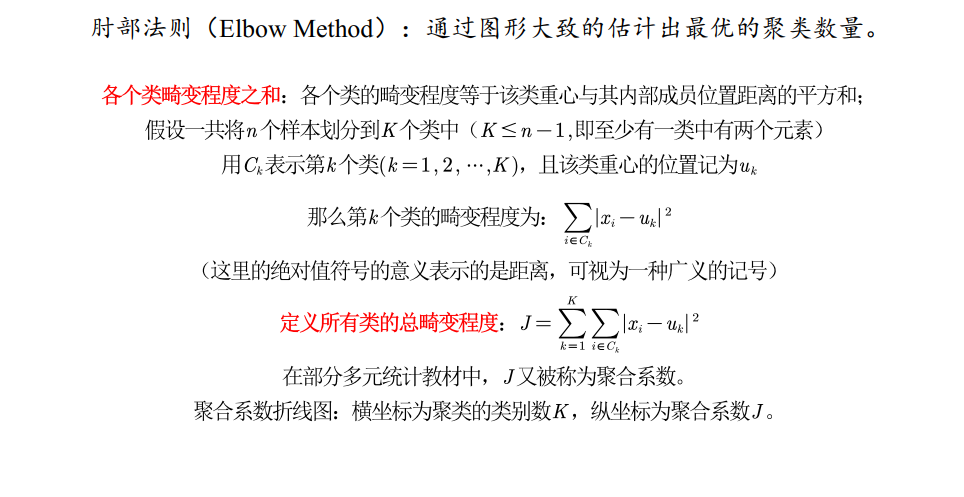
3.模型的评价

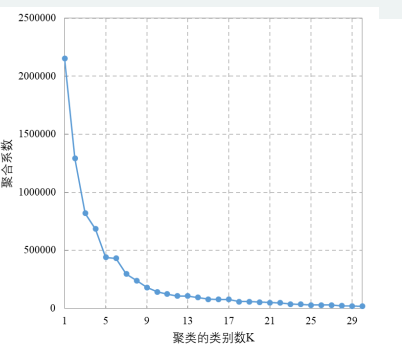
使用Kmeans++对与该数据集进行了聚类，改进了Kmeans对初值敏感的缺陷，使聚类结果更加稳定可靠。

但是仍然需要人工指定K的大小，主观性过强，未结合数据具体分析。

4.模型的改进

使用肘部法则判断K的取值，一定程度上降低了主观性，使得聚类更加客观公正





**层次聚类**

模型的建立与求解

为了研究世界各国森林、草原资源的分布规律，我们依据森林覆盖率，林木蓄积量和草原面积三个指标对21个国家进行聚类分析，由于难以确定类别数目，本文采取系统聚类进行聚类分析。系统聚类的合并算法通过合并算法通过计算两类数据点的欧式距离，对最接近的两类数据点进行组合，反复迭代，直到合成一类，并生成聚类谱系图。其步骤如下：

（1）数据预处理

原始数据的各指标的量纲是不同的，若直接计算数据点之间的距离，显然是不合理的，应先对原始数据标准化消去量纲的影响。标准化结果如下：

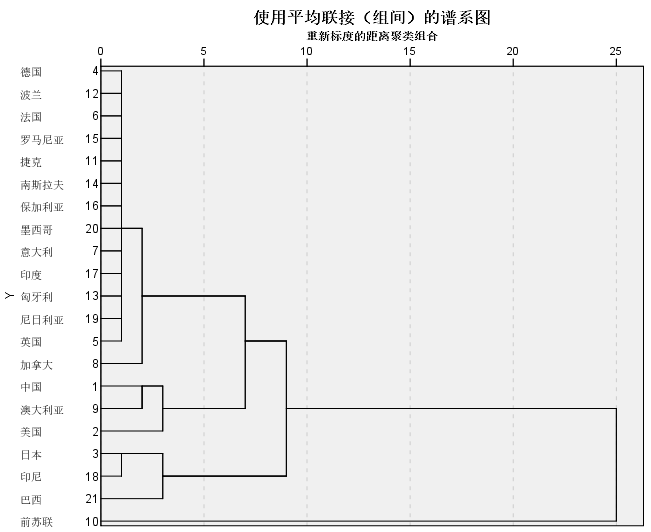
|  |  |  |  |
| --- | --- | --- | --- |
| 国别 | 森林覆盖率 | 林木蓄积量 | 草原面积 |
| 中国 | -1.01551 | 0.04596 | 1.69055 |
| 美国 | -0.08808 | 0.62389 | 1.10695 |
| 日本 | 1.8186 | -0.31997 | -0.58902 |
| 德国 | -0.1917 | -0.3775 | -0.5503 |
| 英国 | -1.21758 | -0.44408 | -0.51107 |
| 法国 | -0.27978 | -0.36684 | -0.50098 |
| 意大利 | -0.56993 | -0.43289 | -0.55638 |
| 加拿大 | 0.03109 | 0.57488 | -0.42247 |
| 澳大利亚 | -0.94298 | -0.39614 | 2.64116 |
| 前苏联 | 0.46631 | 4.03018 | 2.08147 |
| 捷克 | 0.1917 | -0.40466 | -0.58114 |
| 波兰 | -0.22279 | -0.39135 | -0.56418 |
| 匈牙利 | -0.76163 | -0.43875 | -0.58393 |
| 南斯拉夫 | 0.21761 | -0.39135 | -0.54736 |
| 罗马尼亚 | -0.27978 | -0.39188 | -0.56117 |
| 保加利亚 | 0.13471 | -0.43875 | -0.57885 |
| 印度 | -0.60102 | -0.2976 | -0.50728 |
| 印尼 | 2.68904 | -0.27256 | -0.50728 |
| 尼日利亚 | -0.82899 | -0.44781 | -0.44358 |
| 墨西哥 | -0.38859 | -0.27842 | -0.05996 |
| 巴西 | 1.83932 | 0.81564 | 0.54483 |
|  |  |  |  |

（2）利用spss进行系统聚类，生成聚类谱系图，结果如下：

计算类别间距离的方法为欧式距离

初始化每个样本点为一类

然后更新新类反复迭代，直到合为一类。



（3）绘制聚合系数折线图，判断类别数目

聚合系数（所有类的总畸变函数）：

：第k类样本点集 uk:第k类重心

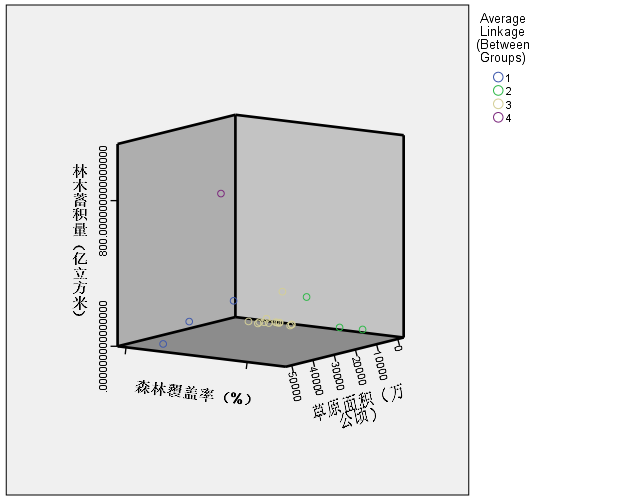
绘制聚合系数折线图如下：

根据聚合系数折线图可知，当类别数为4时，折线的下降趋势趋缓，故可将类别数设定为4。

（4）以聚类数目为4重新聚类并画图。

聚类结果如下：

|  |  |
| --- | --- |
| **聚类成员** | |
| 个案 | 4 个聚类 |
| 1:中国 | 1 |
| 2:美国 | 1 |
| 3:日本 | 2 |
| 4:德国 | 3 |
| 5:英国 | 3 |
| 6:法国 | 3 |
| 7:意大利 | 3 |
| 8:加拿大 | 3 |
| 9:澳大利亚 | 1 |
| 10:前苏联 | 4 |
| 11:捷克 | 3 |
| 12:波兰 | 3 |
| 13:匈牙利 | 3 |
| 14:南斯拉夫 | 3 |
| 15:罗马尼亚 | 3 |
| 16:保加利亚 | 3 |
| 17:印度 | 3 |
| 18:印尼 | 2 |
| 19:尼日利亚 | 3 |
| 20:墨西哥 | 3 |
| 21:巴西 | 2 |



分析：

**DBSCAN模型建立**

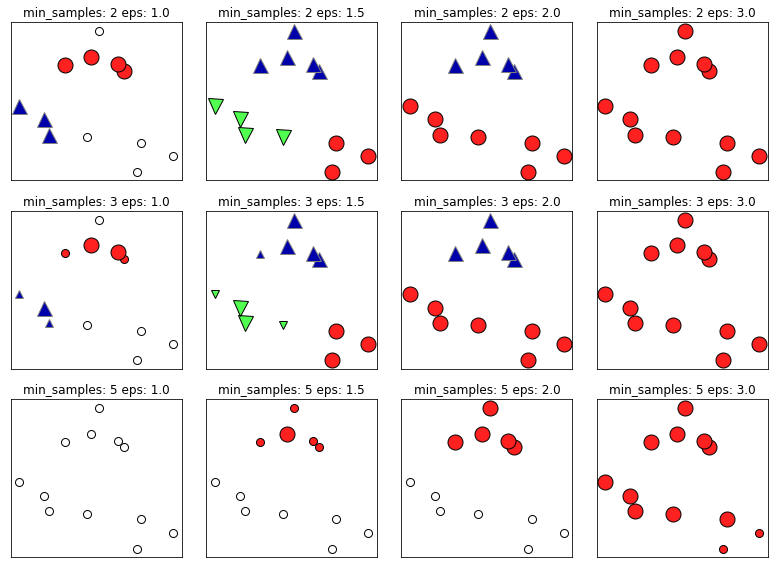
DBSCAN 的原理是识别特征空间的“拥挤”区域中的点，在这些区域中许多数据点靠近在一起。这些区域被称为特征空间中的密集区域。 DBSCAN 背后的思想是，簇形成数据的密集区域，并由相对较空的区域分隔开。它的优点是不需要先验地设置簇的个数，可以划分具有复杂形状的簇，还可以找出不属于任何簇的点。

DBSCAN算法主要有两个参数：eps与min\_samples. 如果在距一个给定数据点 eps 的距离内至少有 min\_samples 个数据点，那么这个数据点就是核心样本。 DBSCAN 将彼此距离小于 eps 的核心样本放到同一个簇中。DBSCAN将数据点分为三类：核心点、边界点与噪音点。

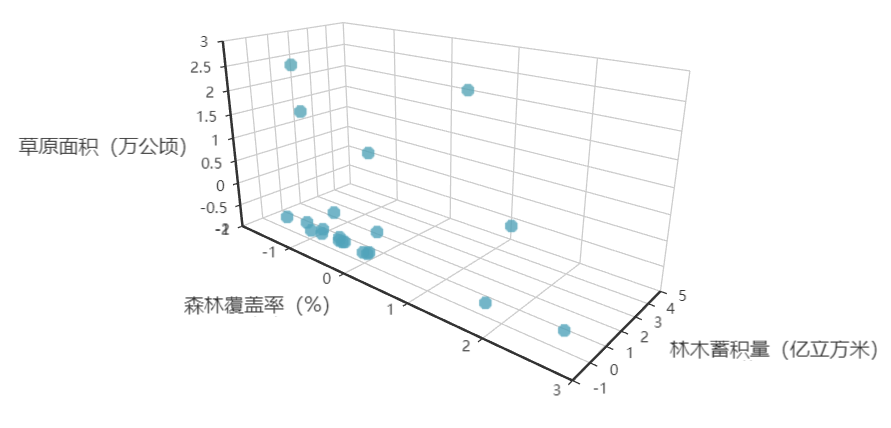
流程图更好

算法首先任意选取一个点，然后找到到这个点的距离小于等于 eps 的所有的点。如果距起始点的距离在 eps 之内的数据点个数小于 min\_samples，那么这个点被标记为噪声（noise），也就是说它不属于任何簇。如果距离在 eps 之内的数据点个数大于 min\_samples，则这个点被标记为核心样本，并被分配一个新的簇标签。然后访问该点的所有邻居（在距离 eps 以内）。如果它们还没有被分配一个簇，那么就将刚刚创建的新的簇标签分配给它们。如果它们是核心样本，那么就依次访问其邻居，以此类推。簇逐渐增大，直到在簇的 eps 距离内没有更多的核心样本为止。然后选取另一个尚未被访问过的点，并重复相同的过程。

设定不同的eps与min\_samples会得到不同的聚类结果，如图



模型的求解



通过观察可以大致将数据集分为四类：底覆盖率底储蓄量底草原面积，低储蓄量草原面积高覆盖率，低覆盖率中储蓄量高面积与其他。

利用DBSCAN算法对各国森林草原资源数据进行聚类，通过调整eps与min\_samples将类别确定在4，eps=1.8, min\_samples=1，聚类结果如下图：

