清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

第十讲:聚类模型

“物以类聚，人以群分”，所谓的聚类，就是将样本划分为由类似的对象组成的多个类的过程。聚类后，我们可以更加准确的在每个类中单独使用统计模型进行估计、分析或预测;也可以探究不同类之间的相关性和主要差异。

聚类和上一讲分类的区别：分类是已知类别的，聚类未知。

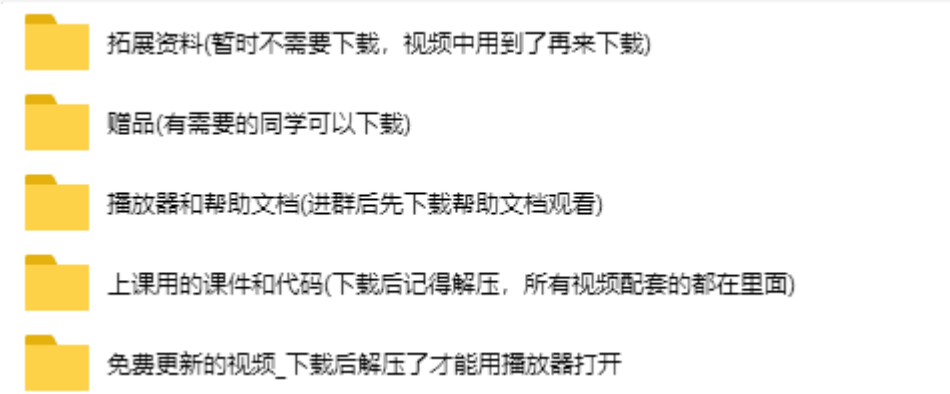
数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

温馨提示

（1）视频中提到的附件可在**售后群的群文件**中下载。

包括**讲义、代码、我视频中推荐的资料**等。



（2）关注我的**微信公众号《数学建模学习交流》**，后台发送**“软件”**两个字，可获得常见的建模软件下载方法；发送**“数据”**两个字，可获得建模数据的获取方法；发送**“画图”**两个字，可获得数学建模中常见的画图方法。另外，也可以看看公众号的历史文章，里面发布的都是对大家有帮助的技巧。

（3）**购买更多优质精选的数学建模资料**，可关注我的微信公众号《数学建模学习交流》，在后台发送**“买”**这个字即可进入店铺进行购买。

（4）视频价格不贵，但价值很高。单人购买观看只需要**58元**，和另外两名队友一起购买人均仅需**46元**，视频本身也是下载到本地观看的，所以请大家**不要侵犯知识产权**，对视频或者资料进行二次销售。

数学建模学习交流

K-means聚类算法

K-means聚类的算法流程：

一、指定需要划分的簇[cù]的个数K值（类的个数）；

二、随机地选择K个数据对象作为初始的聚类中心（不一定要是我们的样本点）；

三、计算其余的各个数据对象到这K个初始聚类中心的距离，把数据对象划归到距离它最近的那个中心所处在的簇类中；

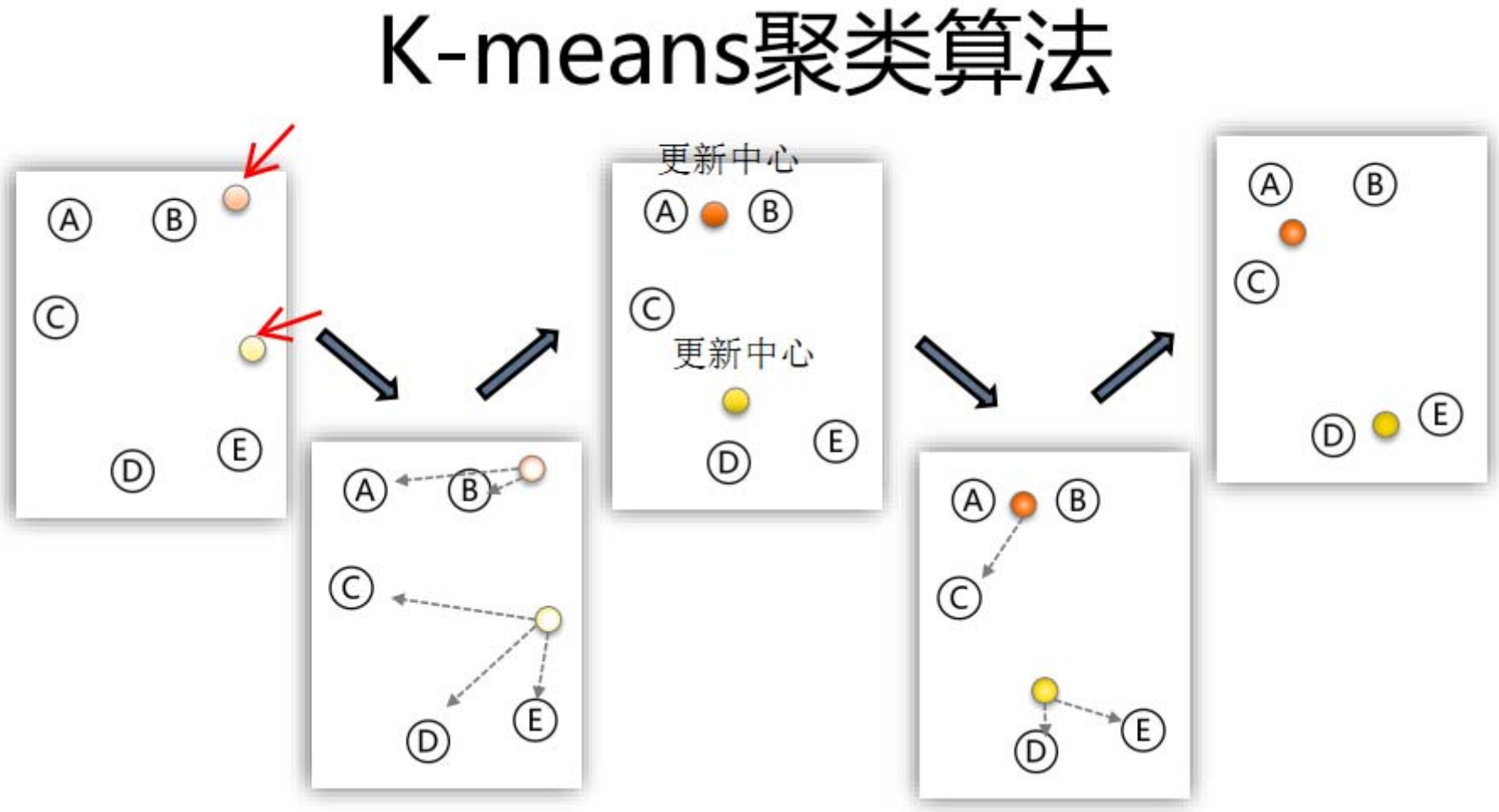
四、调整新类并且重新计算出新类的中心；

五、循环步骤三和四，看中心是否收敛（不变），如果收敛或达到迭代次数则停止循环；

六、结束。

数学建模学习交流

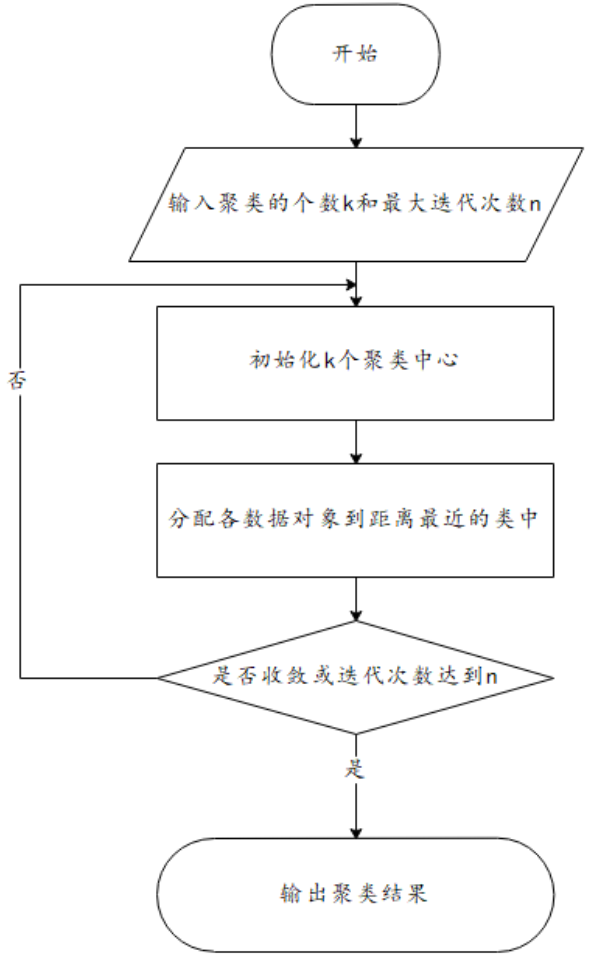
图解K-means算法



K-均值聚类可视化：https://www.naftaliharris.com/blog/visualizing-k-means-clustering/
（如果网页失效的话也没关系，仅供演示用，不影响后面的学习）

数学建模学习交流

算法流程图



亿图、PPT、Visio等软件都可以画
（上图是我使用亿图绘制的）

数学建模学习交流

K-means算法的评价

优点：

（1）算法简单、快速。

（2）对处理大数据集，该算法是相对高效率的。

缺点：

（1）要求用户必须事先给出要生成的簇的数目K。

（2）对初值敏感。

（3）对于孤立点数据敏感。

K-means++算法可解决2和3这两个缺点。

数学建模学习交流

K-means++算法

k-means++算法选择初始聚类中心的基本原则是：**初始的聚类中心之间的相互距离要尽可能的远。**

算法描述如下：

（只对K-means算法“初始化K个聚类中心” 这一步进行了优化）

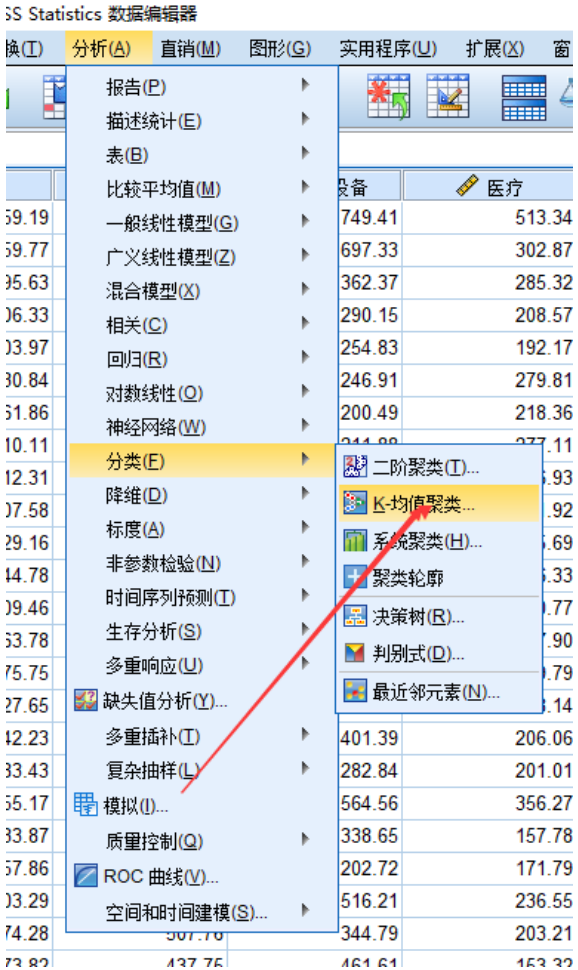
步骤一：随机选取一个样本作为第一个聚类中心；

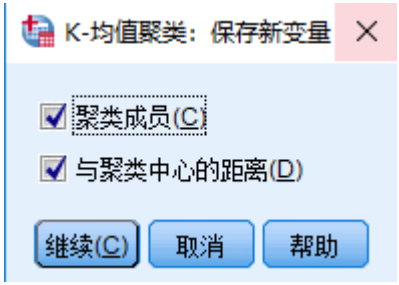
步骤二：计算每个样本与当前已有聚类中心的最短距离（即与最近一个聚类中心的距离），这个值越大，表示被选取作为聚类中心的概率较大；最后，用轮盘法（依据概率大小来进行抽选）选出下一个聚类中心；

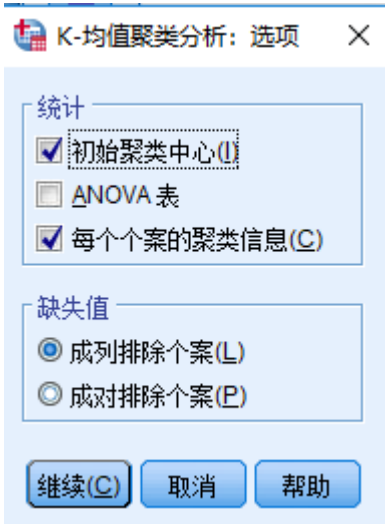
步骤三：重复步骤二，直到选出K个聚类中心。选出初始点后，就继续使用标准的K-means算法了。

数学建模学习交流

Spss软件操作







默认使用的就是K-means++算法

聚类成员

个案号	城市	聚类	距离
1	北京	1	25.188
2	天津	2	6.856
3	河北	3	7.437
4	山西	3	16.033
5	内蒙古	3	3.107
6	辽宁	3	32.207
7	吉林	3	48.307
8	黑龙江	3	43.217
9	上海	1	29.578
10	江苏	2	45.684
11	浙江	2	25.136
12	安徽	3	44.677
13	福建	3	36.223
14	江西	3	35.947
15	山东	3	15.773
16	河南	3	30.747
17	湖南	3	30.383
18	湖北	3	13.547
19	广东	1	11.643
20	广西	3	22.203
21	海南	2	19.034
22	重庆	3	29.733
23	四川	3	27.393
24	贵州	3	4.587
25	云南	2	32.726
26	西藏	1	43.123
27	陕西	3	4.727
28	甘肃	3	32.123
29	青海	3	40.443
30	宁夏	3	.137
31	新疆	3	18.333

数学建模学习交流

K-means算法的一些讨论

（1）聚类的个数K值怎么定？

答：分几类主要取决于个人的经验与感觉，通常的做法是多尝试几个K值，看分成几类的结果更好解释，更符合分析目的等。

（2）数据的量纲不一致怎么办？

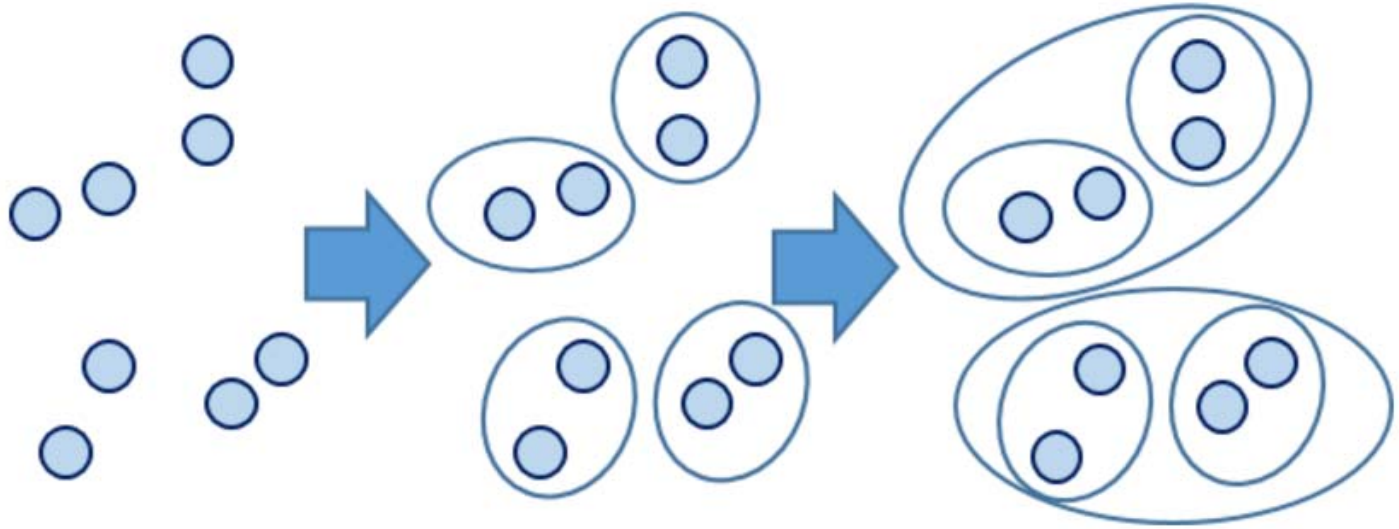
答：如果数据的量纲不一样，那么算距离时就没有意义。例如：如果X1单位是米，X2单位是吨，用距离公式计算就会出现“米的平方”加上“吨的平方”再开平方，最后算出的东西没有数学意义，这就有问题了。

$$z_i = \frac{x_i - \overline{x}}{\sigma_x}$$(先减去均值再除以标准差)

数学建模学习交流

系统（层次）聚类

系统聚类的合并算法通过计算两类数据点间的距离，对最为接近的两类数据点进行组合，并反复迭代这一过程，直到将所有数据点合成一类，并生成聚类谱系图。



下面的内容请查看于老师的PPT
于晶贤-辽宁石油化工大学-聚类分析之系统聚类法.pdf

数学建模学习交流

系统（层次）聚类算法流程

系统（层次）聚类的算法流程：

一、将每个对象看作一类，计算两两之间的最小距离；

二、将距离最小的两个类合并成一个新类；

三、重新计算新类与所有类之间的距离；

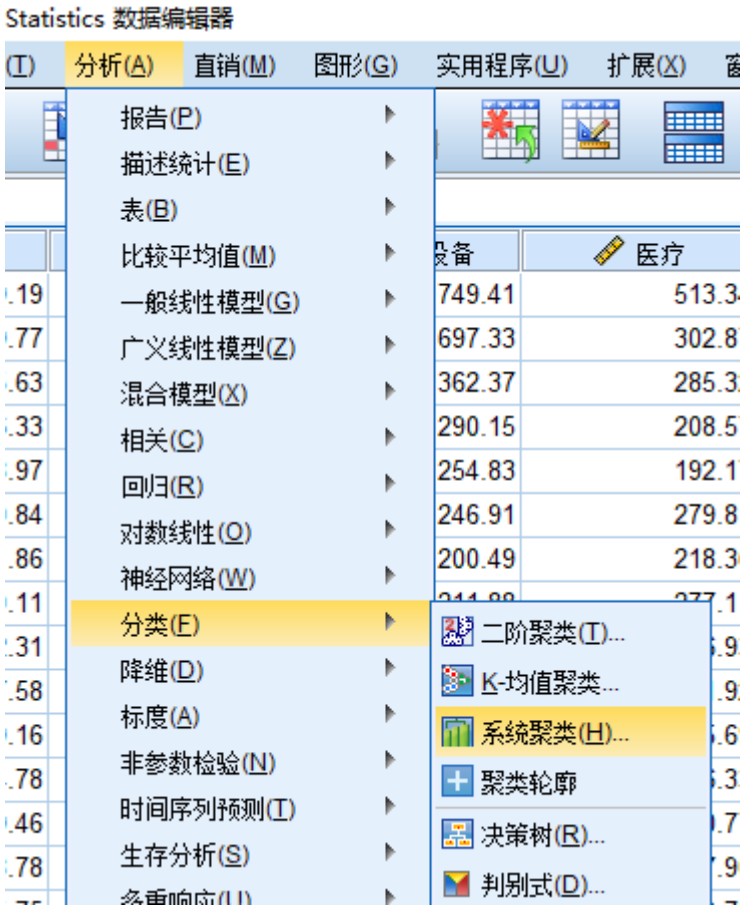
四、重复二三两步，直到所有类最后合并成一类；

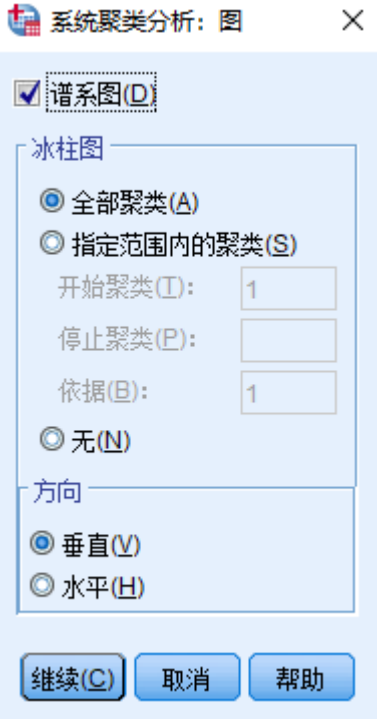
五、结束。

课后作业：将上述文字表述的流程绘制成一个流程图。
（避免被查重的最好方法就是自己动手总结）

数学建模学习交流

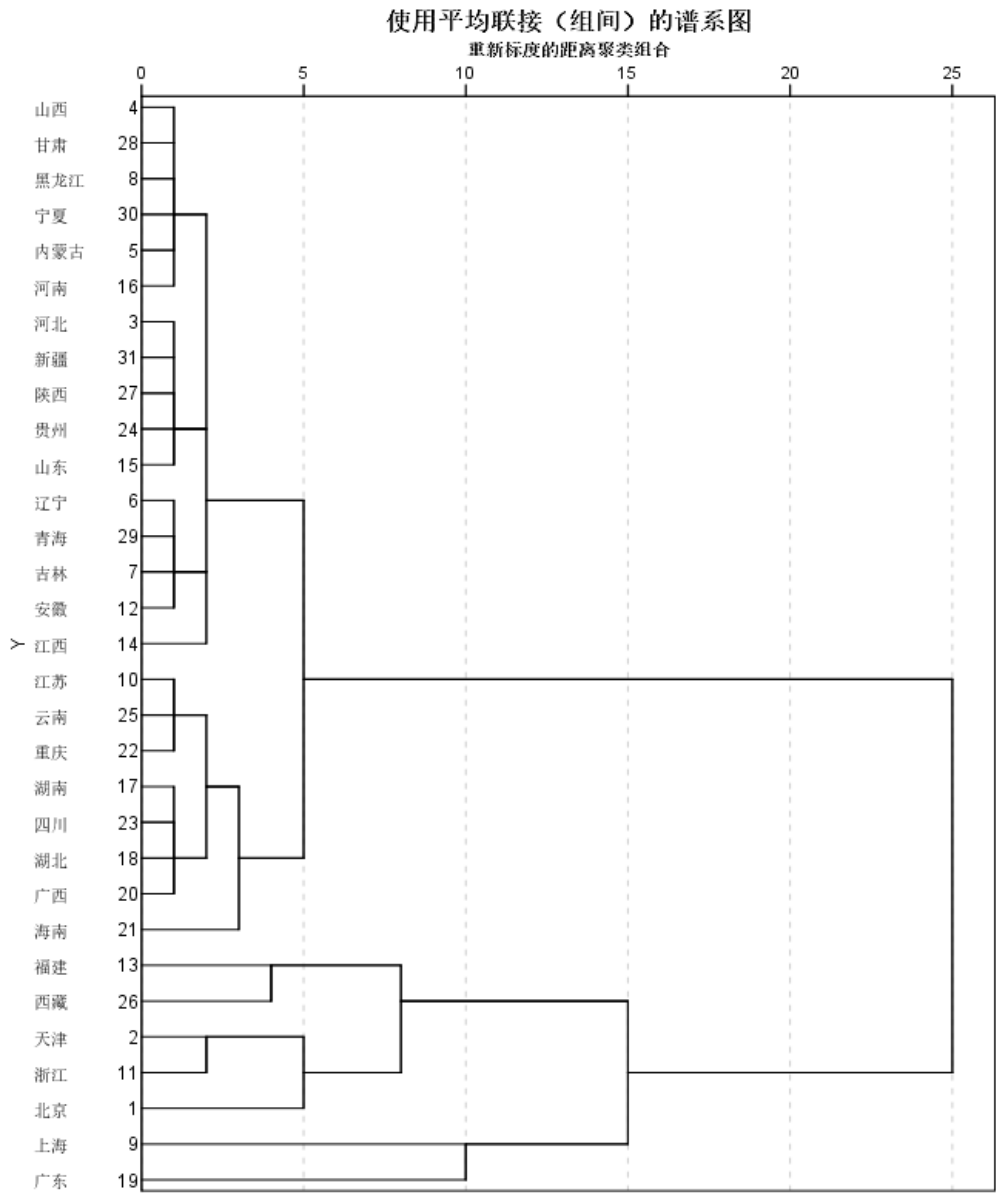
Spss软件操作





数学建模学习交流

聚类谱系图（树状图）

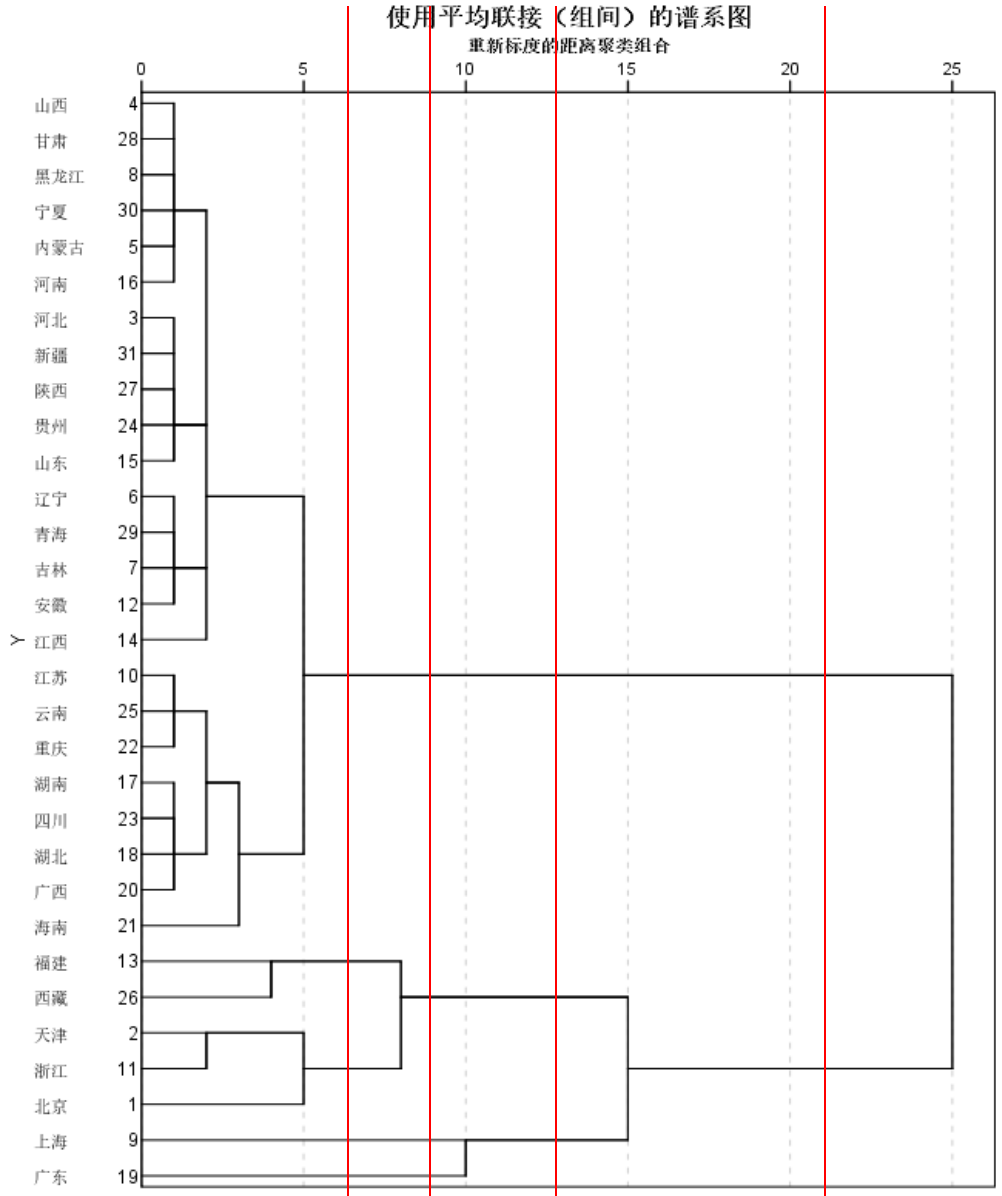


谱系图是较新的Spss版本添加的功能
横轴表示各类之间的距离
（该距离经过了重新标度）
聚类的个数可以自己从图中决定。

Spss结果中还有一种图，被称为冰柱图，
目前已经很少用了。

数学建模学习交流

聚类谱系图（树状图）



谱系图是较新的Spss版本添加的功能
横轴表示各类之间的距离
（该距离经过了重新标度）
聚类的个数可以自己从图中决定。

Spss结果中还有一种图，被称为冰柱图，
目前已经很少用了。

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

用图形估计聚类的数量

肘部法则（Elbow Method）：通过图形大致的估计出最优的聚类数量。

各个类畸变程度之和：各个类的畸变程度等于该类重心与其内部成员位置距离的平方和；

假设一共将n个样本划分到K个类中（$K \leq n-1$,即至少有一类中有两个元素）

用C_k表示第k个类($k=1,2,\cdots,K$)，且该类重心的位置记为u_k

那么第k个类的畸变程度为：$\sum_{i \in C_k} |x_i - u_k|^2$

（这里的绝对值符号的意义表示的是距离，可视为一种广义的记号）

定义所有类的总畸变程度：$J = \sum_{k=1}^{K} \sum_{i \in C_k} |x_i - u_k|^2$

在部分多元统计教材中，J又被称为聚合系数。

聚合系数折线图：横坐标为聚类的类别数K，纵坐标为聚合系数J。

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

画图前先对数据进行处理

集中计划

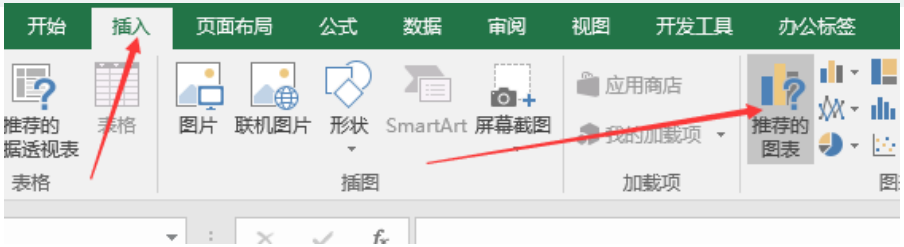
阶段	组合聚类 聚类 1	聚类 2	系数	首次出现聚类的阶段 聚类 1	聚类 2	下一个阶段
1	4	28	18414.012	0	0	7
2	8	30	20297.111	0	0	5
3	3	31	22161.800	0	0	10
4	6	29	28190.529	0	0	8
5	5	8	29537.326	0	2	7
6	17	23	29624.380	0	0	11
7	4	5	34367.611	1	5	9
8	6	7	38093.577	4	0	16
9	4	16	48752.992	7	0	20
10	3	27	49242.893	3	0	13
11	17	18	53983.761	6	0	17
12	10	25	57922.410	0	0	14
13	3	24	59288.905	10	0	15
14	10	22	77201.747	12	0	22
15	3	15	77370.143	13	0	19
16	6	12	79172.059	8	0	19
17	17	20	94389.183	11	0	22
18	2	11	107280.103	0	0	26
19	3	6	107957.125	15	16	20
20	3	4	124430.310	19	9	21
21	3	14	142727.072	20	0	25
22	10	17	180442.145	14	17	23
23	10	21	239702.036	22	0	25
24	13	26	296564.664	0	0	27
25	3	10	431624.250	21	23	30
26	1	2	440027.064	0	18	27
27	1	13	683977.679	26	24	29
28	9	19	819519.221	0	0	29
29	1	9	1291671.774	27	28	30
30	1	3	2152790.747	29	25	0

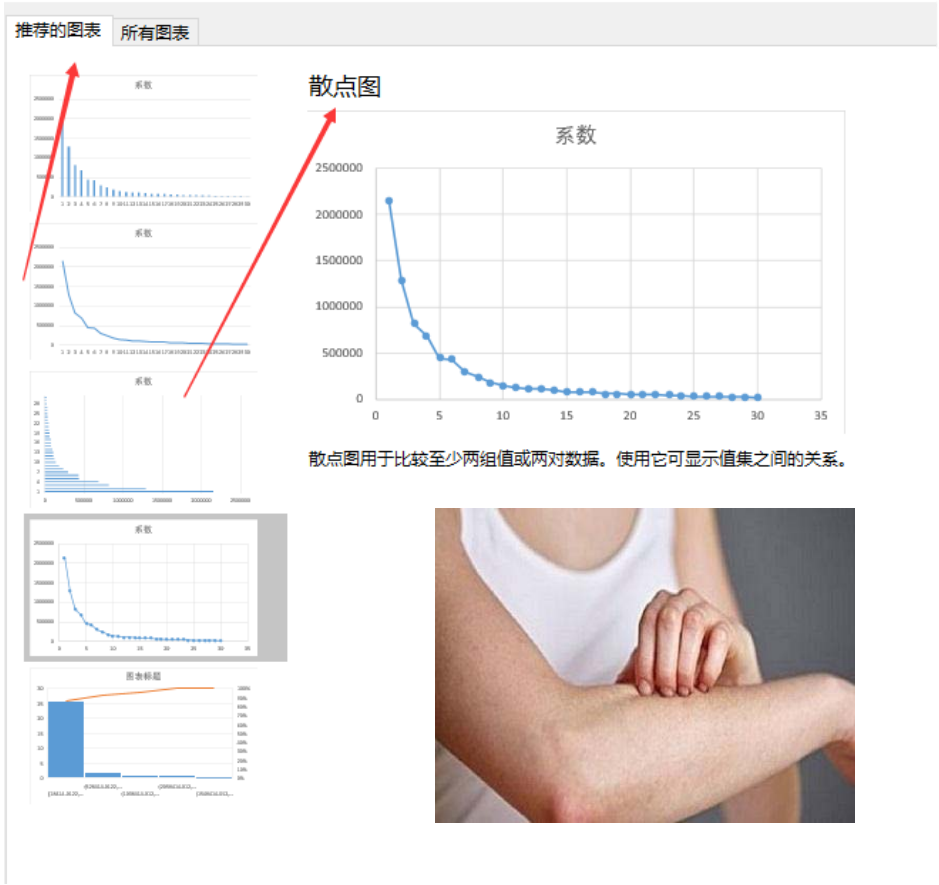
	A
1	系数
2	2152791
3	1291672
4	819519.2
5	683977.7
6	440027.1
7	431624.2
8	296564.7
9	239702
10	180442.1
11	142727.1
12	124430.3
13	107957.1
14	107280.1
15	94389.18
16	79172.06
17	77370.14
18	77201.75
19	59288.9
20	57922.41
21	53983.76
22	49242.89
23	48752.99
24	38093.58
25	34367.61
26	29624.38
27	29537.33
28	28190.53
29	22161.8
30	20297.11
31	18414.01

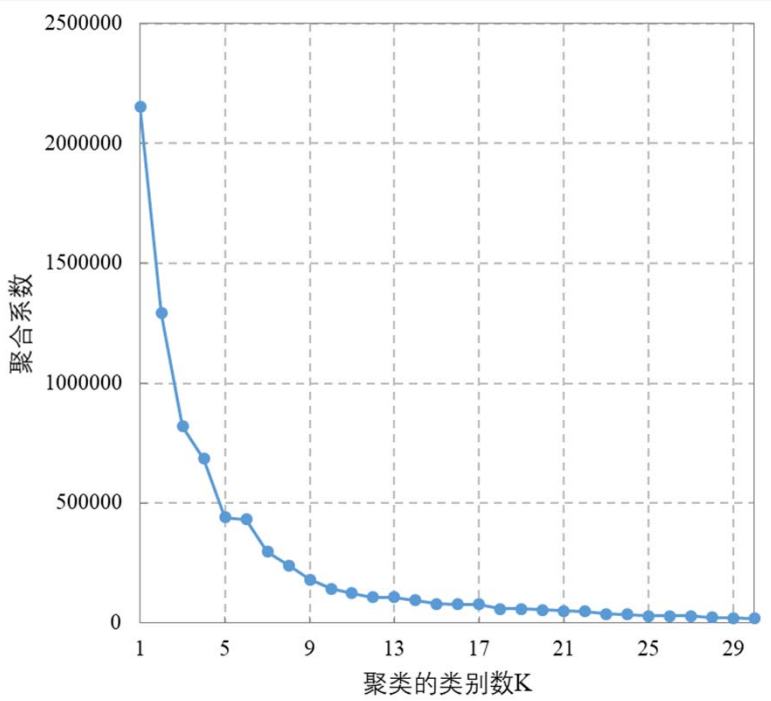
把数据粘贴到Excel表格中，
并按照降序排好。

数学建模学习交流

聚合系数折线图的画法







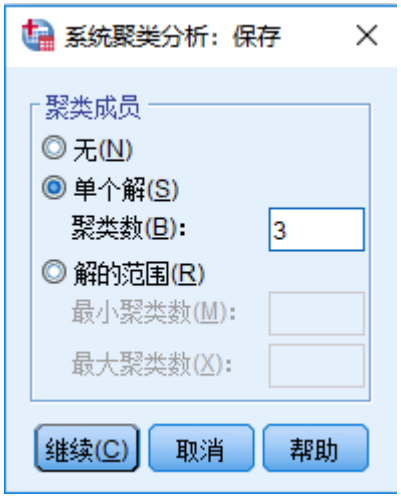
根据图来进行解释：

(1)根据聚合系数折线图可知，当类别数为5时，折线的下降趋势趋缓，故可将类别数设定为5.

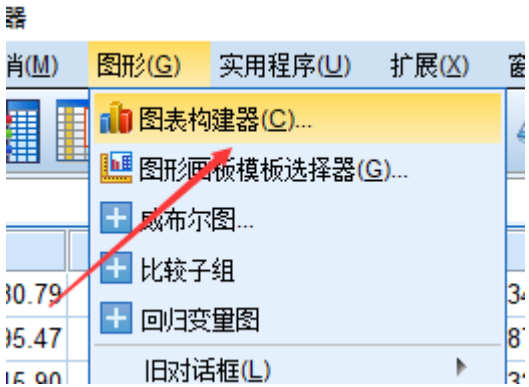
(2)从图中可以看出， K值从1到5时，畸变程度变化最大。超过5以后，畸变程度变化显著降低。因此肘部就是 K=5，故可将类别数设定为5.（当然，K=3也可以解释）

数学建模学习交流

确定K后保存聚类结果并画图



重新聚类一次，确定聚类个数

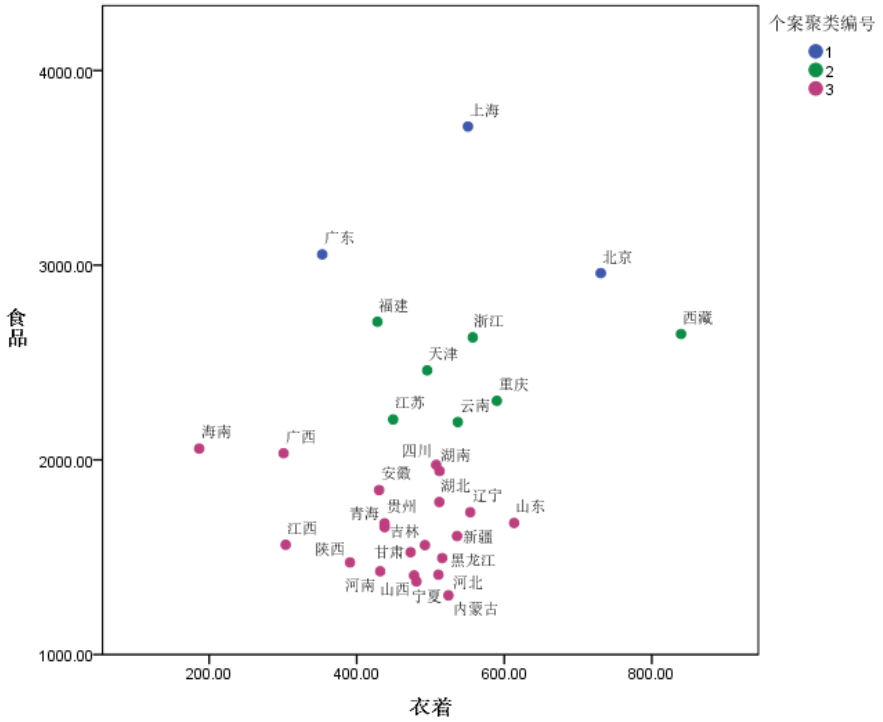


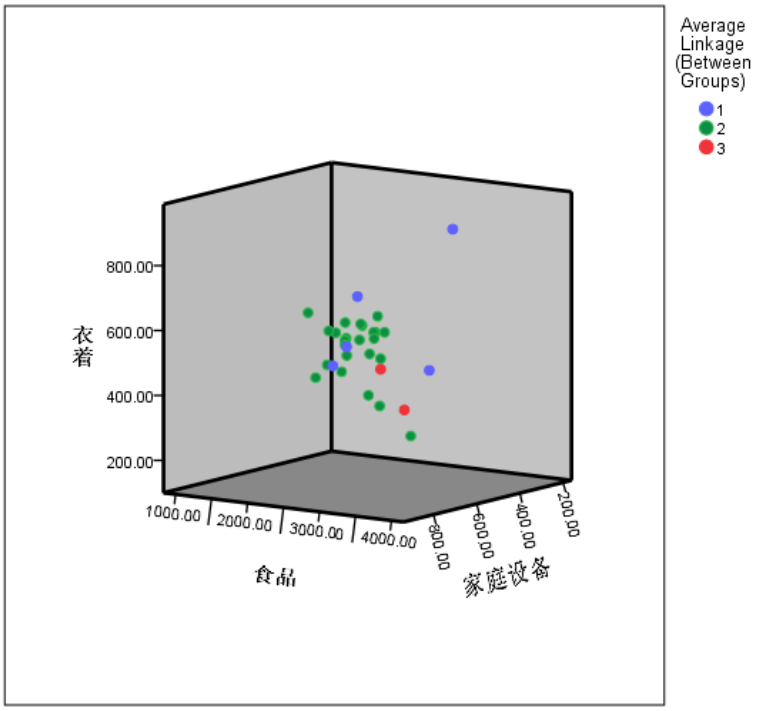
聚类结束后画图

数学建模学习交流

示意图

最好是不用默认的（太丑了，特别是那个背景颜色）
双击图中的任意元素，可对其进行调整。



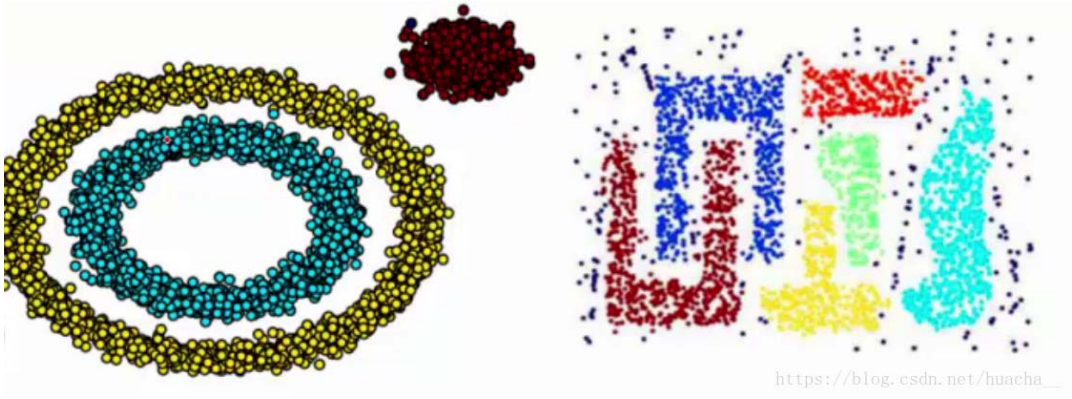


注意：只要当指标个数为2或者3的时候才能画图，上面两个图纯粹是为了演示作图过程，实际上本例中指标个数有8个，是不可能做出这样的图的。

数学建模学习交流

DBSCAN算法

DBSCAN(Density-based spatial clustering of applications with noise)是Martin Ester, Hans-PeterKriegel等人于1996年提出的一种基于密度的聚类方法，聚类前不需要预先指定聚类的个数，生成的簇的个数不定（和数据有关）。该算法利用基于密度的聚类的概念，即要求聚类空间中的一定区域内所包含对象（点或其他空间对象）的数目不小于某一给定阈值。该方法能在具有噪声的空间数据库中发现任意形状的簇，可将密度足够大的相邻区域连接，能有效处理异常数据。



DBSCAN：具有噪声的基于密度的聚类方法

谁和我挨的近，我就是谁兄弟
兄弟的兄弟，也是我的兄弟

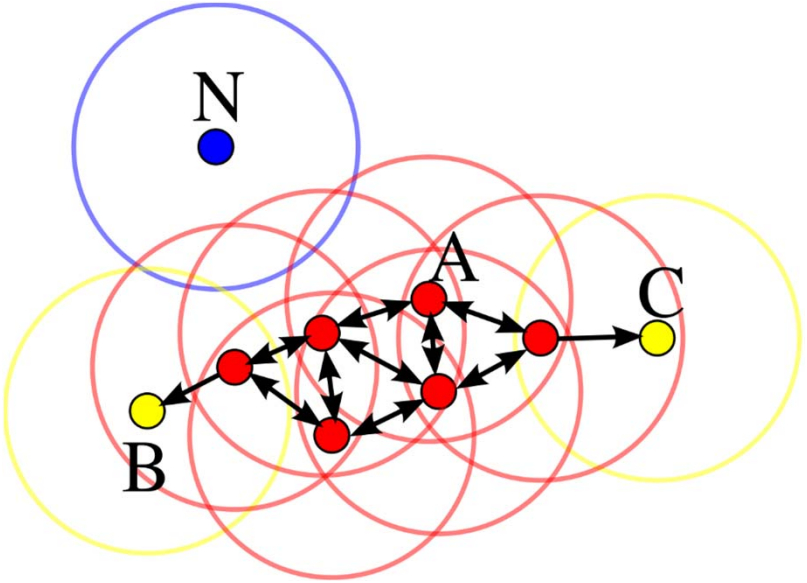
数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

基本概念

DBSCAN算法将数据点分为三类：

- 核心点：在半径Eps内含有不少于MinPts数目的点
- 边界点：在半径Eps内点的数量小于MinPts，但是落在核心点的邻域内
- 噪音点：既不是核心点也不是边界点的点



在这幅图里，MinPts = 4，点 A 和其他红色点是核心点，因为它们的 ε-邻域（图中红色圆圈）里包含最少 4 个点（包括自己），由于它们之间相互相可达，它们形成了一个聚类。点 B 和点 C 不是核心点，但它们可由 A 经其他核心点可达，所以也和A属于同一个聚类。点 N 是局外点，它既不是核心点，又不由其他点可达。

数学建模学习交流

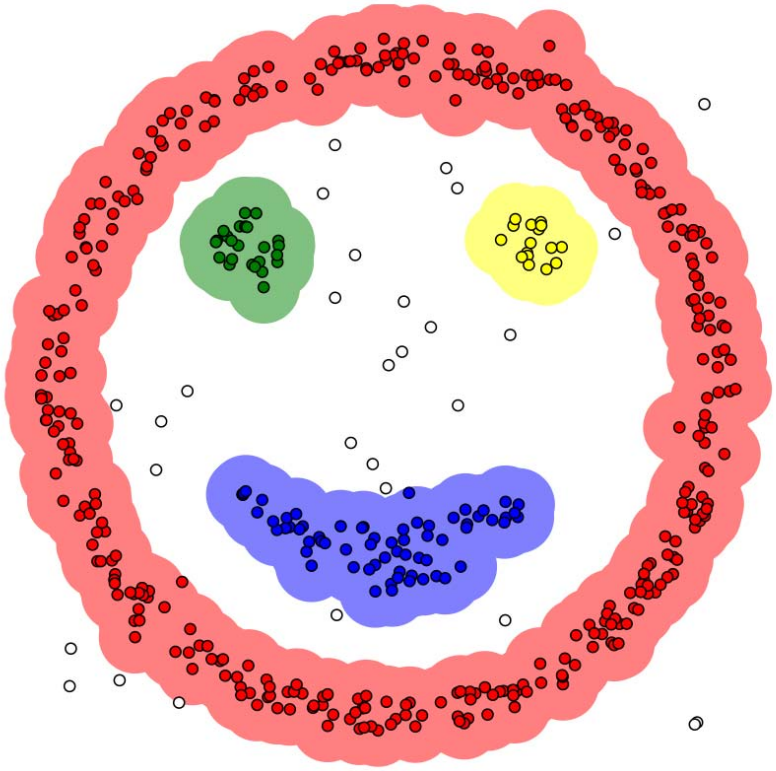
关注微信公众号"数学建模学习交流"获取更多优质资料

DBSCDN可视化

DBSCAN算法可视化

https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

（如果网页失效的话也没关系，仅供演示用，不影响后面的学习）



数学建模学习交流

伪代码（维基百科）

```
DBSCAN(D, eps, MinPts) {
   C = 0
   for each point P in dataset D {
      if P is visited
         continue next point
      mark P as visited
      NeighborPts = regionQuery(P, eps)
      if sizeof(NeighborPts) < MinPts
         mark P as NOISE
      else {
         C = next cluster
         expandCluster(P, NeighborPts, C, eps, MinPts)
      }
   }
}

expandCluster(P, NeighborPts, C, eps, MinPts) {
   add P to cluster C
   for each point P' in NeighborPts {
      if P' is not visited {
         mark P' as visited
         NeighborPts' = regionQuery(P', eps)
         if sizeof(NeighborPts') >= MinPts
            NeighborPts = NeighborPts joined with NeighborPts'
      }
      if P' is not yet member of any cluster
         add P' to cluster C
   }
}

regionQuery(P, eps)
   return all points within P's eps-neighborhood (including P)
```

伪代码（Pseudocode）是一种非正式的，类似于英语结构的，用于描述模块结构图的语言。使用伪代码的目的是使被描述的算法可以容易地以任何一种编程语言（Pascal，C，Java等）实现。因此，伪代码必须结构清晰、代码简单、可读性好，并且类似自然语言。

数学建模学习交流

Matlab代码

Matlab官网推荐下载的代码：
https://ww2.mathworks.cn/matlabcentral/fileexchange/52905-dbscan-clustering-algorithm

```
% Copyright (c) 2015, Yarpiz (www.yarpiz.com)
% All rights reserved. Please read the "license.txt" for license terms.
%
% Project Code: YPML110
% Project Title: Implementation of DBSCAN Clustering in MATLAB
% Publisher: Yarpiz (www.yarpiz.com)
%
% Developer: S. Mostapha Kalami Heris (Member of Yarpiz Team)
%
% Contact Info: sm.kalami@gmail.com, info@yarpiz.com
```

数学建模学习交流

优缺点

优点：

1. 基于密度定义，能处理任意形状和大小的簇；
2. 可在聚类的同时发现异常点；
3. 与K-means比较起来，不需要输入要划分的聚类个数。

缺点：

1. 对输入参数 ε 和Minpts敏感，确定参数困难；
2. 由于DBSCAN算法中，变量 ε 和Minpts是全局唯一的，当聚类的密度不均匀时，聚类距离相差很大时，聚类质量差；
3. 当数据量大时，计算密度单元的计算复杂度大。

我的建议：

只有两个指标，且你做出散点图后发现数据表现得很“DBSCAN”，这时候你再用DBSCAN进行聚类。

其他情况下，全部使用系统聚类吧。

K-means也可以用，不过用了的话你论文上可写的东西比较少。

数学建模学习交流

课后作业

（1）代码题(选做)：请结合DBSCAN的伪代码给Matlab官网推荐下载的代码添加上注释。（代码要是怕被查重，最容易的方法就是加上自己的注释，当然也可以修改变量名）
注意：里面可能有部分函数你之前没有见过，例如pdist2和false这两个函数，请自己查阅相关资料，并带上一些小例子在注释中，以防日后忘记。

（2）用亿图、PPT或者Visio等软件中的一种画出系统聚类的流程图，画出来的图以后国赛就可以直接用了，而且它的版权属于你自己，查重你也不用怕。

（3）完成一篇论文作业，“各国森林、草原资源聚类.doc”，提示：指标共三个，量纲不同哦，论文中要交代清楚K的选择，当然你也可以画出你的聚类结果的三维图。

（4）体验下DBSCAN代码，我会给你一个新的数据集合，你可以试试对于这个数据集，选择多大的Eps和MinPts会使得聚类的效果最好。(smile.mat)

数学建模学习交流