**北京邮电大学软件学院**

**2019 -2020 学年第 一 学期实验报告**

**课程名称： 数据挖掘**

**实验名称： 实验一聚类实验**

**实验完成人：**

**姓名：**\_\_黄莹\_\_**学号：**\_\_2017211953\_\_**成绩：**\_\_\_\_\_\_\_\_

**指导教师：**\_\_\_\_\_\_\_\_\_\_\_\_\_牛琨\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**日 期： 2019 年 10 月 23 日**

1. **数据描述报告**

该数据集一共有1411条数据，一共有36个属性，其中产品ID、型号、品牌等属性不是数值型，其余全为数值型，具体含义如下：

产品颜色有七个分类；产品上市时间为2003到2012；产品市场定位为0经济实用，1新潮炫酷，2娱乐小资，3商务功能，4品味尊贵，5老年机，6儿童机；芯片平台分为0高通，1威盛；G网分为0否，1是；芯片主频为40-2400；AP分为0否，1是；频段数量为1-5；零售价格为184-9380；外观类型分为0直板，1翻盖，2滑盖，3旋转，4座机；厚度为9-85；产品重量48.4-790.2；屏幕数量分为1单，2双；主屏幕尺寸分为0-7；显示分辨率为6240-921600；触摸屏分为0无，1电阻，2电容；键盘类型0无，1数字，2全；RAM为1-4096；ROM为0-16384；Flash闪存0-16384；摄像头为0-1300；定位分为0否，1是；FM广播分为0否，1是；电视分为0否，1是；Modem分为0否，1是；红外分为0否，1是；蓝牙分为0否，1是；WLAN分为0否，1是；电池容量为100-4000；重力感应器分为0否，1是；方向感应器分为0否，1是；文字输入方法数为1-3；智能系统分为0否，1是。

该数据集收集了2003年-2012年之间上市的1411款手机的型号品牌、颜色价格、市场定位、硬件信息、外观类型、厚度和重量、屏幕尺寸等信息，其中零售价格均值为1117元，平均厚度为14.9，平均重量为107.6。

1. **K值改变给聚类结果所来的变化**

在测试之前，我们对数据进行了预处理，删除了id属性，以防止该无关属性对聚类结果产生影响，并设置seed=10作为初始值。

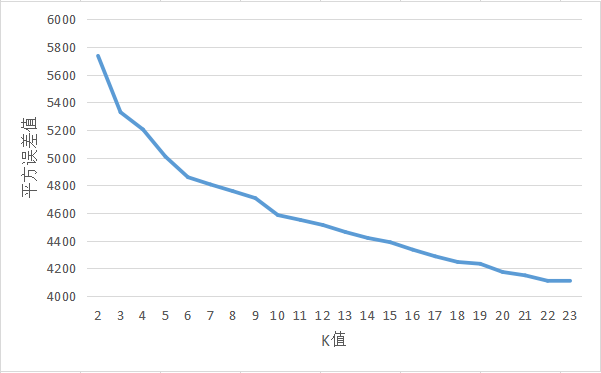
我们测试了2-23的K值给聚类结果产生的变化，实验结果证明：**K值越大，聚类结果的误差平方和越小**。这是评价聚类好坏的标准，数值越小说明同一簇实例之间的距离越小。

|  |  |
| --- | --- |
| K | squared error |
| 13 | 4462.66827026774 |
| 14 | 4419.35531097935 |
| 15 | 4388.52231236273 |
| 16 | 4334.77582392397 |
| 17 | 4286.62661735115 |
| 18 | 4245.75140679751 |
| 19 | 4232.68422245722 |
| 20 | 4173.17727573977 |
| 21 | 4148.98839053949 |
| 22 | 4109.51168816217 |
| 23 | 4108.8497746592 |

我们结果表格如下：

|  |  |
| --- | --- |
| K | squared error |
| 2 | 5736.571028148646 |
| 3 | 5328.687461920233 |
| 4 | 5204.117772492127 |
| 5 | 5007.561701596787 |
| 6 | 4858.832210131204 |
| 7 | 4806.56002707882 |
| 8 | 4757.841521779878 |
| 9 | 4707.425270841811 |
| 10 | 4584.805199302372 |
| 11 | 4549.994132583084 |
| 12 | 4513.259344228589 |

绘制成图表后：



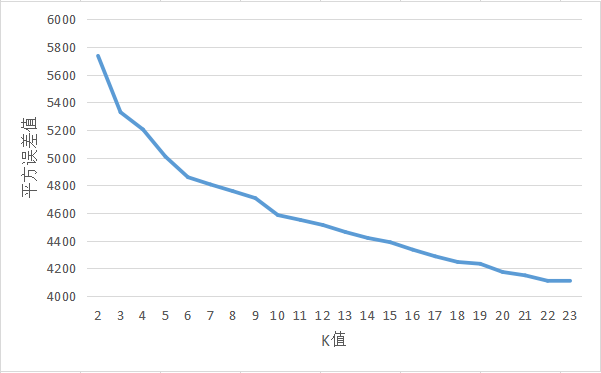
如图所示，K值越大，SSE值越小，但是其下降速度随着K值的增长而变小。

1. **最优聚类数目选取**

我们采用两种方法来共同决定最优聚类数目，以手肘法作主要方法，用轮廓系数法作为辅助。

1. 手肘法。

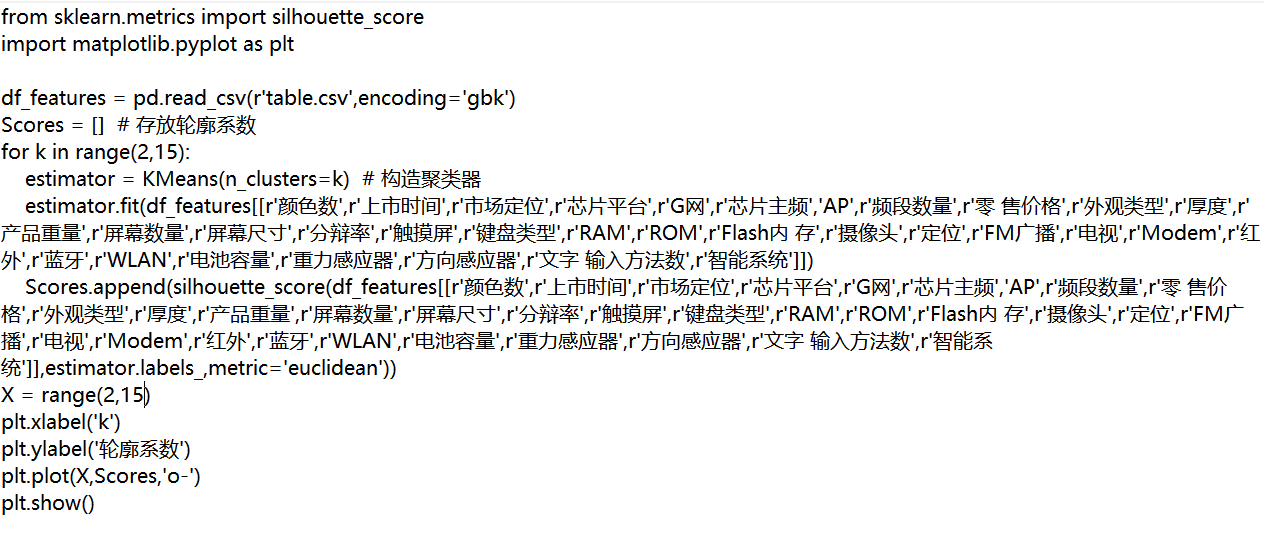
手肘法的核心思想是：随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。并且，当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，也就是说SSE和k的关系图是一个手肘的形状，而这个肘部对应的k值就是数据的真实聚类数。

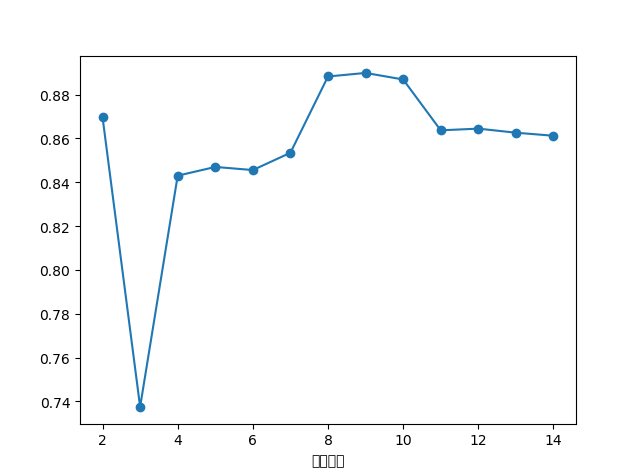


根据该图，手肘位置大概是k=10的位置。

1. 轮廓系数法

簇内样本的距离越近，簇间样本距离越远，平均轮廓系数越大，聚类效果越好。那么，很自然地，平均轮廓系数最大的k便是最佳聚类数。





由该图得到，k=8-10时，平均轮廓系数最大，而且此时SSE也足够小，因此我得出结论：

**K=10的模型，为簇心固定时的最佳模型。**

1. **簇心大小给聚类结果所来的变化**

由上一节可知，我们已经决定将k=10作为簇心固定时的最佳模型，现在我们保持K=10,不变，研究簇心大小与聚类结果的变化。

seed参数是要设一个随机种子，依次产生一个随机数，用来得到k均值算法中初始的第k个簇中心的位置。

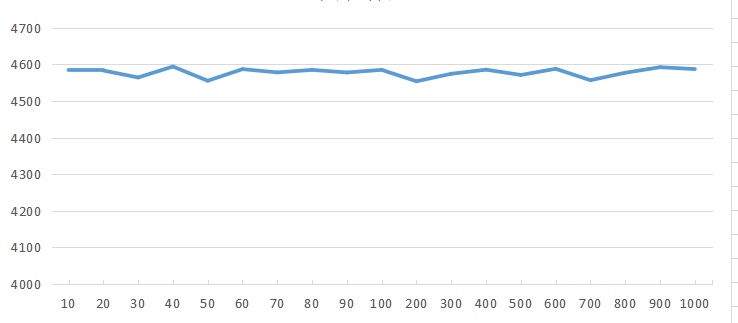
我们测试了10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000的seed值给聚类结果产生的变化，实验结果证明：seed值越大与聚类结果的误差平方和没有太大关系。Seed值的改变会让聚类结果微微波动，但不会有很大变化。

结果表格如下：

|  |  |
| --- | --- |
| Seed | 误差平方和 |
| 10 | 4584.805199 |
| 20 | 4583.380617 |
| 30 | 4564.302948 |
| 40 | 4594.223062 |
| 50 | 4555.381116 |
| 60 | 4587.105034 |
| 70 | 4578.220192 |
| 80 | 4585.086585 |
| 90 | 4578.030713 |
| 100 | 4585.147708 |

|  |  |
| --- | --- |
| Seed | 误差平方和 |
| 200 | 4553.928946 |
| 300 | 4574.451972 |
| 400 | 4585.673981 |
| 500 | 4571.228478 |
| 600 | 4587.991924 |
| 700 | 4556.879314 |
| 800 | 4577.273428 |
| 900 | 4592.39948 |
| 1000 | 4587.217526 |

结果绘成折线图如下：



可以看出，聚类结果与初始簇心选择并无明显关系。

1. **最终模型最优分析及选择**

本模型最终选取K=10，seed=10的模型作为最佳模型，因为当K=10的时候误差平方和足够小，变化足够缓慢，即接近真实的聚类数目，并且其轮廓系数最大，而seed取值与聚类结果无明确关系，因此仍以seed=10作为模型参数。