Advances in Applied Mathematics 应用数学进展, 2020, 9(11), 2029-2035
Published Online November 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.911235

Hans汉斯

基于改进K-Means算法的手机处理器聚类分析

林　挺，蔡　静，李东升

贵州民族大学数据科学与信息工程学院，贵州　贵阳
Email: 963856774@qq.com

收稿日期：2020年11月1日；录用日期：2020年11月18日；发布日期：2020年11月25日

摘　要

提出了一种改进的K-means聚类算法，并对2015至2019年手机市场主流的51款处理器进行聚类分析。首先使用手肘法改进K-means算法中K值的选取，求出最佳K值；其次利用欧氏距离求得各样本到聚类中心的距离，并将各样本归类到离其最近的聚类中心所在的簇中；重新计算新簇的聚类中心，若与旧聚类中心相同，则停止运算，否则重新计算各样本到新聚类中心的距离，重新归类直至新聚类中心与旧聚类中心相同；最后得到4个簇，分别包含7、16、14、14个样本，并将其分为高端、中端、中低端、低端处理器。

关键词

K-Means，手肘法，K值，聚类中心，处理器

Clustering Analysis of Mobile Processor Based on Improved K-Means Algorithm

Ting Lin, Jing Cai, Dongsheng Li

School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang Guizhou
Email: 963856774@qq.com

Received: Nov. 1st, 2020; accepted: Nov. 18th, 2020; published: Nov. 25th, 2020

Abstract

An improved K-means clustering algorithm is proposed, and 51 mainstream processors in the mobile phone market from 2015 to 2019 are analyzed. Firstly, the elbow method is used to improve the selection of K value in k-means algorithm, and the best K value is obtained. Secondly, the Euclidean distance is used to find the distance from each sample to the cluster center, and all samples are gradually classified into the nearest cluster. Then the new cluster centers are recalculated.

文章引用: 林挺, 蔡静, 李东升. 基于改进 K-Means 算法的手机处理器聚类分析[J]. 应用数学进展, 2020, 9(11): 2029-2035. DOI: 10.12677/aam.2020.911235

If the new cluster centers are the same as the old ones, the operation is stopped. Otherwise, the distance from each sample to the new cluster centers is recalculated and reclassified until the new cluster centers are the same as the old ones. Finally, four clusters are obtained, including 7, 16, 14 and 14 samples, which are divided into high-end, middle-end, middle-low-end and low-end processors according to the original data.

Keywords

K-Means, Elbow Method, K Value, Clustering Center, Processor

1. 引言

自从苹果公司于 2007 年发布第一代 iPhone 以来，智能手机极大地改变了人们的日常生活，手机性能的提升增加了手机的应用场景，伴随着互联网技术的发展，智能手机的作用越来越大。拍照、录像、网络社交、游戏、视频通话等许多功能都依赖于手机来实现。手机配置的提升，要求手机处理器要有相当强大的运算能力。因此人们对软件应用需求的增大也反过来迫使厂商增强手机性能。

董骐瑞[1]提出 K-means 最优聚类数的确定存在不足，从而改进了 K-means 算法。凌静[2]等人提出一种结合模拟退火算法的遗传 K-means 聚类方法，以此避免聚类陷入局部最优。黄晓辉[3]等人发现 K-means 算法主要考虑簇内距离，而忽略了簇间距离的作用，并提出一种新的集成簇内和簇间距离的加权 K-means 算法。王巧玲[4]等人针对 K-means 算法随机选择初始聚类中心和 K 值所导致的精度不高问题，提出了一种基于聚合距离的改进 K-means 算法。唐泽坤[5]等人针对 K-means 算法对初始聚类中心和噪声敏感的缺点，提出了 *d*-K-means 算法，能够更好地选择聚类中心。刘越[6]提出了一种改进的 K-means 以降低算法的时间复杂度。黄继超[7]针对 K-means 中 K 值的模糊性，提出了使用距离代价函数来确定准确的 K 值。

处理器对智能手机的重要性相当于大脑对人类的重要性。本文选用了 2015 年至 2019 年手机市场主流的 51 款处理器，涉及到 13 种指标，利用 K-means 算法对其进行聚类。旨在将不同年度、不同定位、不同产商的处理器聚合到适合各自的簇中，以此可以对市场上铺天盖地的各款手机性能有更好的理解，对采用了不同簇的处理器的手机定位也能有一个参考。

2. 改进 K-Means 聚类算法

2.1. 传统 K-Means 聚类算法

作为一种无监督学习算法，K-means (K 均值聚类算法，也称作快速聚类法)由于其简单高效、聚类效果良好等特点，被广泛运用于各种领域的样本聚类。

K-means 聚类算法最早是由 Macqueen 于 1967 年提出，K 指的是类的个数，即聚类簇数。其基本思想是对给定的样本集，先确定聚类簇数 K 以及 K 个聚类中心(均值)，再计算各样本到聚类中心的距离，将其划分到最近中心点所在的簇。K-means 是一种迭代算法，大致步骤如下[8]：

步骤 1：确定 K 值，在 K-means 聚类算法中，K 是唯一的参数；

步骤 2：将样品初步分成 K 类，一般是从数据集中随机选取 K 个样本作聚类中心(明显聚类中心为向量)；

步骤 3：逐个计算样本到各聚类中心的欧式距离，并将其派送到最近的聚类中心所在的簇，直至所有样本都分配完成；

步骤 4：重新计算各簇的聚类中心(即各簇内的样本均值)，若与上一个聚类中心相同，则算法运行结束，若不同，则用新聚类中心取代旧聚类中心，重复进行步骤 3，直至聚类中心相同。

并且一个样本只能属于一个簇，每个簇至少要有一个样本。

2.2. 手肘法改进 K-Means 算法

设数据集的样本总数为 n，K-means 算法的迭代次数为 t，可知该算法的复杂度为 $O(nKt)$。对于 K 值的选取，可以由个人经验所得，也可由手肘法[9]计算得出，手肘法是利用 SSE(误差平方和)：

$$\mathrm{SSE}=\sum_{q=1}^{K}\sum_{l=1}^{n_q}\left\|\boldsymbol{p}_{ql}-\overline{\boldsymbol{p}}_q\right\|^2 \tag{1}$$

以此来得出 K 值。其中 n_q 为各簇中样本的数量，满足 $\sum_{q=1}^{K}n_q=n$，$\boldsymbol{p}_{ql}$ 为第 q 个簇里的第 l 个样本，$\overline{\boldsymbol{p}}_q$ 为第 q 个簇里的聚类中心。

手肘法的思想是：随着 K 值逐渐增大，样本的划分会更加细致，随着各个簇聚合程度的逐步提高，SSE 也会随之变小。同时，当 K 值小于最优聚类簇数时，由于 K 的增大会较大幅度地增加每个簇的聚合程度，所以 SSE 的下降幅度会比较大；而当 K 值到达最优聚类簇数时，再增加 K 值，簇增加的聚合程度就没有之前高，SSE 的下降幅度会大幅变小，逐渐趋于平稳。所以 SSE 与 K 值之间的关系图类似手肘的形状，肘部对应 K 值就是最优聚类簇数。K 值是 K-means 算法中唯一的参数，选取最优 K 值可以在一定程度上改进 K-means 算法。

2.3. 计算距离并分配到簇

先将原始数据集标准化，以消除量纲的影响。本文采用的是 Z-score 标准化：

$$z_{ij}=\frac{x_{ij}-\overline{x}_i}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(x_{ij}-\overline{x}_i\right)^2}},j=1,2,\cdots,n \tag{2}$$

z_{ij} 是标准化后生成的新数据，x_{ij} 为原始数据，$\overline{x}_i$ 为第 i 个指标均值，对本文数据而言，$i=1,2,\cdots,13$。

从标准化后的数据中，按照 2.2 节手肘法确定 K 值，随机选取 K 个样本 $\{\overline{\boldsymbol{p}}_1,\overline{\boldsymbol{p}}_2,\cdots,\overline{\boldsymbol{p}}_K\}$ 作为初始聚类中心，计算样本 $\boldsymbol{p}_j$ 到各聚类中心的欧氏距离[10]：

$$d_{jq}=\left\|\boldsymbol{p}_j-\overline{\boldsymbol{p}}_q\right\|,q=1,2,\cdots,K \tag{3}$$

对于样本 $\boldsymbol{p}_j$，选取离其最近的聚类中心，并标记 $\delta_j=\arg\min_{q\in\{1,2,\cdots,K\}}d_{jq}$，将样本 $\boldsymbol{p}_j$ 划分到簇 C_{δ_j} 中。直至所有样本分配到各簇中，进行步骤 4 即可。

3. 运用改进的 K-Means 聚类算法对手机处理器进行聚类分析

目前主流的手机处理器公司有美国高通(骁龙 Snapdragon 系列)、苹果(A 系列)，韩国三星(Exynos 系列)，中国华为海思(麒麟 Kirin 系列)、联发科(Helio 系列)。其中苹果、三星与华为自家都有手机产品，而高通与联发科仅研发处理器，没有手机品牌。

本文选取了全球市场上 51 款主流手机处理器进行分析，共涉及到 5 个品牌：高通、苹果、三星、华为海思、联发科；13 个指标：CPU 单核跑分 $\boldsymbol{x}_1$、CPU 多核跑分 $\boldsymbol{x}_2$、GPU 跑分 $\boldsymbol{x}_3$、内存带宽 $\boldsymbol{x}_4$ (GB/s)、

网络下行速度 x_5 (Mbps)、网络上行速度 x_6 (Mbps)、制程工艺 x_7 (nm)、闪存读写速度 x_8 (GB/s)、内存主频 x_9 (MHz)、支持摄像头数量 x_{10}、最大内存容量 x_{11} (GB)、最高摄像头像素 x_{12} (万)、屏幕最大显示像素 x_{13}。

其中 CPU 单核跑分、CPU 多核跑分取自 Geekbench 官网；GPU 跑分取自 3D Mark 官网；内存带宽、网络下行速度、网络上行速度、制程工艺、闪存读写速度、内存主频、支持摄像头数量、最大内存容量、最高摄像头像素、屏幕最大显示像素取自各产商官网。数据见表 1：

Table 1. Indicators of mobile processor
表 1. 手机处理器各项指标

处理器	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}
HP10	784	3006	338	7.464	300	50	28	0.4	933	2	4	2100	2073600
HP20	876	3997	642	25.6	300	50	16	0.4	1600	3	6	2400	2073600
HP22	816	3551	374	25.6	300	50	12	0.4	1600	3	6	2100	1152000
HP35	910	4070	318	25.6	300	50	12	0.4	1600	3	6	2500	2592000
HP60	1497	5735	923	28.8	300	150	12	1.5	1800	3	8	3200	2332800
HP65	1840	6024	1287	28.8	300	150	12	1.5	1800	4	8	4800	2721600
HX25	1592	4134	684	14.928	300	50	20	0.4	933	3	4	3200	4096000
HX30	1788	5817	1630	29.856	450	150	10	1.5	1866	3	8	3200	4096000
S430	674	2496	296	6.4	150	75	28	0.4	800	2	4	2100	2304000
S435	676	2583	309	6.4	300	100	28	0.4	800	2	4	2100	2304000
S439	864	3060	453	6.4	150	75	12	0.4	800	3	4	2100	2462400
S450	775	3457	433	7.464	300	150	14	0.4	933	3	4	2400	2462400
S625	846	3969	465	7.464	300	150	14	0.4	933	2	4	2400	2462400
S626	911	4373	462	7.464	300	150	14	0.4	933	3	4	2400	2462400
S630	882	4291	841	21.328	600	150	14	1.5	1333	3	8	2400	3145728
S632	1256	4832	512	21.328	300	150	14	0.4	1333	3	8	2400	2462400
S636	1330	4894	937	21.328	600	150	14	1.5	1333	3	8	2400	2462400
S652	1419	3305	870	14.928	300	150	28	0.4	933	2	4	2100	4096000
S653	1234	3805	935	14.928	600	150	28	0.4	933	2	8	2100	4096000
S660	1624	5523	1346	29.856	600	150	14	1.5	1866	3	8	4800	4096000
S665	1496	5455	1131	29.856	600	150	11	1.5	1866	4	8	4800	2721600
S670	1710	5805	1831	29.856	600	150	10	1.5	1866	3	8	19200	4096000
S675	2381	6302	1082	29.856	600	150	11	1.5	1866	4	8	19200	2721600
S710	1811	5631	1811	29.856	800	150	10	1.5	1866	4	8	19200	4838400
S712	1890	5981	2067	29.856	800	150	10	1.5	1866	4	8	19200	4838400
S730	2556	7027	2164	29.856	800	150	8	1.5	1866	4	8	19200	2721600
S808	1153	2353	1068	14.928	450	50	20	0.2	933	2	3	2100	4096000
S810	1269	2922	1311	24.88	450	50	20	0.2	1555	2	4	5500	8294400

Continued

S820	1782	3902	2460	28.848	600	150	14	0.725	1803	2	6	2800	8294400
S821	1843	4051	2376	28.848	600	150	14	0.725	1803	3	6	2800	8294400
S835	1918	6601	3590	29.856	1000	150	10	1.5	1866	3	8	3200	8294400
S845	2435	8429	4700	29.856	1200	150	10	1.5	1866	4	10	4800	8294400
S855	3459	10976	5491	34.128	2000	316	7	3	2133	5	12	19200	8294400
K650	796	3247	334	7.464	300	100	16	0.4	933	3	4	1600	2073600
K659	922	3675	405	7.464	300	100	16	0.4	933	3	4	1600	2462400
K710	1557	5459	952	25.6	600	150	12	1.5	1600	4	6	3200	2527200
K810	2825	7830	2421	25.6	1000	150	7	1.5	1600	5	8	4800	2527200
K950	1714	5750	924	25.6	300	50	16	0.725	1600	3	4	3200	3686400
K960	1887	6274	1878	29.856	600	150	16	1.5	1866	3	6	3200	4096000
K970	1903	6636	2912	29.856	1200	150	10	1.5	1866	4	8	4800	4147200
K980	3350	10068	4182	34.128	1400	200	7	1.5	2133	5	12	4800	5456000
E7420	1455	4430	1266	24.832	300	50	14	0.725	1552	2	4	1600	3686400
E8890	1829	5336	2058	29.856	600	150	14	0.725	1866	2	4	2400	9216000
E8895	2006	6540	2545	29.856	1000	150	10	1.5	1866	3	6	2800	9216000
E9810	3521	8719	3260	29.856	1200	200	10	1.5	1866	3	8	2400	9216000
E9820	4490	9530	4335	29.856	2000	316	8	2.4	1866	5	12	2200	9216000
A8	1573	2694	1053	14.928	150	50	20	0.4	933	2	2	800	3145728
A9	2553	4500	1845	25.6	300	50	16	1.8	1600	2	2	1200	3145728
A10	3569	6086	2302	25.6	450	100	16	1.8	1600	3	3	1200	3145728
A11	4251	10461	3091	25.6	1000	150	10	1.8	1600	3	4	1200	3145728
A12	4822	11406	4187	29.856	1000	225	7	1.8	1866	3	4	1200	3338496

利用式(2)对原始数据进行标准化处理，再利用式(1)求出 SSE 与 K 值之间的关系，结果如图 1 所示：

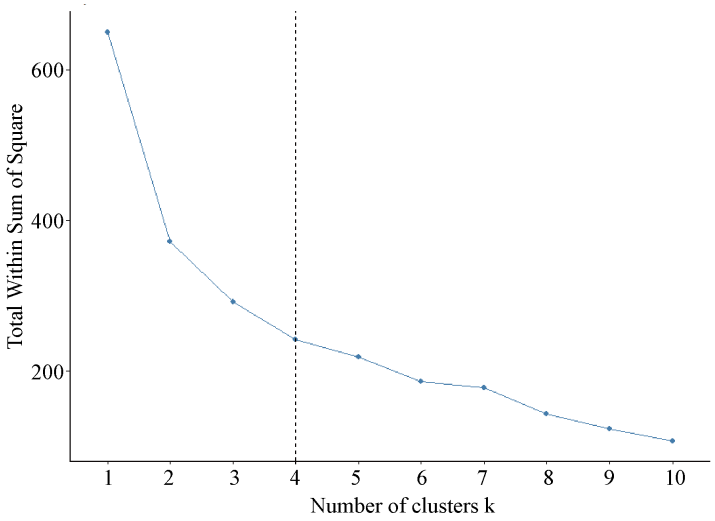

Figure 1. Elbow method to confirm the optimal K value
图 1. 手肘法确认最优 K 值

从图 1 可以看出，肘部对应的 K 值为 4，最佳聚类簇数为 4。利用式(3)计算距离，分配样本到簇，重复进行步骤 3 与步骤 4，得出 4 个簇以及簇中样本，见图 2：

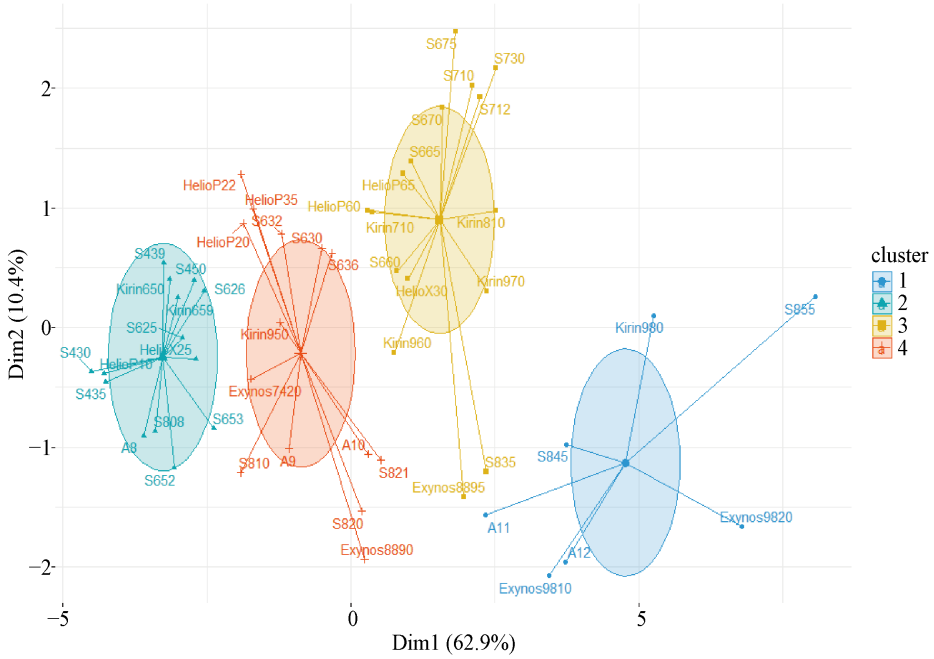

Figure 2. K-means clustering results of mobile processor
图 2. 手机处理器 K-means 聚类结果

4 个簇中分别包含了 7、16、14、14 个样本，其中 1 类处理器的各项指标大多数领先其他类的处理器，可以看出 1 类为高端处理器，事实上 1 类中的 A12、Kirin980、S855、Exynos9820 分别是苹果、华为海思、高通、三星在 2019 年的旗舰处理器，而 A11、S845、Exynos9810 则是 2018 年的旗舰处理器，符合事实。而 2 类则是中端与旧高端处理器的集合，它们的 CPU 多核、内存带宽等方面相差不大，但旧高端处理器(如 2017 年的 Kirin970、S835、Exynos8895)在 GPU 以及网络上行速度、下行速度方面有着明显的领先；而中端处理器(如 2019 年的 Kirin810、S730)在 CPU 单核、最高摄像头像素方面则领先。4 类则是中低端处理器与年代更为久远的高端处理器的集合。3 类处理器的各指标大多数落后于其他类的处理器，是低端处理器的集合。

4. 结论

本文改进了 K-means 算法中 K 值的选取，并对 2015 年至 2019 年手机市场主流的 51 款处理器进行聚类，得到了 4 个类别，并结合原始数据将其分为高端(1 类)、中端(2 类)、中低端(4 类)、低端(3 类)处理器。由此可以对市场上不同品牌不同处理器的手机做横向纵向对比，对采用了某款处理器的手机在市场上的定位也能提供重要参考。但是 K-means 算法只能将样本分为指定的 K 个类，无法对类别做出分级，因此对于聚类结果的分析，需要依赖人为主观因素，如本文将聚类结果分成不同等级。这也说明 K-means 算法依旧存在需要改进的地方。

基金项目

贵州省教育厅青年科技人才成长项目(黔教合 KY 字[2018] 142)。

参考文献

[1] 董骐瑞. k-均值聚类算法的改进与实现[D]: [硕士学位论文]. 长春: 吉林大学, 2015.

[2] 凌静, 江凌云, 赵迎. 结合模拟退火算法的遗传 K-Means 聚类方法[J]. 计算机技术与发展, 2019, 29(9): 61-65.

[3] 黄晓辉, 王成, 熊李艳, 曾辉. 一种集成簇内和簇间距离的加权 k-means 聚类方法[J]. 计算机学报, 2019, 42(12): 2836-2848.

[4] 王巧玲, 乔非, 蒋友好. 基于聚合距离参数的改进 K-means 算法[J]. 计算机应用, 2019, 39(9): 2586-2590.

[5] 唐泽坤, 朱泽宇, 杨裔, 李彩虹, 李廉. 基于距离和密度的 d-K-means 算法[J]. 计算机应用研究, 2020, 37(6): 1719-1723.

[6] 刘越. K-means 聚类算法的改进[D]: [硕士学位论文]. 桂林: 广西师范大学, 2016.

[7] 黄继超. k-means 算法若干改进和应用[D]: [硕士学位论文]. 长沙: 中南大学, 2013.

[8] 何晓群. 多元统计分析[M]. 第 4 版. 北京: 中国人民大学出版社, 2015: 64-65.

[9] 吴广建, 章剑林, 袁丁. 基于 K-means 的手肘法自动获取 K 值方法研究[J]. 软件, 2019, 40(5): 167-170.

[10] Liberti, L., Lavor, C., Maculan, N., *et al.* (2012) Euclidean Distance Geometry and Applications. *Quantitative Biology*, **56**, 3-69. https://doi.org/10.1137/120875909