文章编号: 1001-0920(2017)04-0759-04 DOI: 10.13195/j.kzyjc.2016.0274

# 基于平均差异度优选初始聚类中心的改进$K$-均值聚类算法

李 武†, 赵娇燕, 严太山

(湖南理工学院 信息与通信工程学院，湖南 岳阳 414006)

**摘 要:** 针对$K$-均值聚类算法对初始聚类中心存在依赖性的缺陷,提出一种基于数据空间分布选取初始聚类中心的改进算法. 该算法首先定义样本距离、样本平均差异度和样本集总体平均差异度;然后将每个样本按平均差异度排序,选择平均差异度较大且与已选聚类中心的差异度大于样本集总体平均差异度的样本作为初始聚类中心. 实验表明,改进后的算法不仅提高了聚类结果的稳定性和正确率,而且迭代次数明显减少,收敛速度快.

**关键词:** $K$-均值聚类；初始聚类中心；样本差异度

**中图分类号:** N945 **文献标志码:** A

# Improved $K$-means clustering algorithm optimizing initial clustering centers based on average difference degree

*LI Wu†, ZHAO Jiao-yan, YAN Tai-shan*

(College of Information and Communication Engineering，Hu'nan Institute of Science and Technology，Yueyang 414006，China)

**Abstract:** Aiming at the dependence on initial clustering centers of the $K$-means clustering algorithm, an improved algorithm is proposed. In the improved $K$-means algorithm, the initial clustering centers are selected according to the distribution of data spatial. The distance between two samples, the average difference of each sample, and total average difference of sample set are defined. Then the average difference of each sample is sorted. The sample with larger average difference is selected as the initial clustering center if its difference from the selected cluster is larger than average difference. Experimental results show that the stability and accuracy of the clustering results are increased by using the improved algorithm, and the convergence speed is also accelerated.

**Keywords:** $K$-means clustering； initial clustering center； sample difference

## 0 引 言

聚类分析是指将物理或抽象对象的集合分成由类似对象组成的多个类的过程,它是研究分类问题的一种统计分析方法,同时也是数据挖掘的一种重要方法. $K$-均值算法作为一种基于划分的动态聚类算法,获得了广泛应用[1-3]. 然而,$K$-均值算法对初始聚类中心的依赖性使得最终聚类结果不稳定,正确率较低,从而影响算法性能. 一些学者对$K$-均值算法进行了研究,提出了多种改进算法[4-14]. 这些改进算法取得了较好的聚类结果,但没有从根本上摆脱算法对初始聚类中心的依赖性,而初始聚类中心选取的随机性和盲目性使得算法的性能无法得到显著提高.

为此,本文提出基于平均差异度优选初始聚类中心的改进$K$-均值聚类算法. 为了寻找与数据分布相一致的初始聚类中心,首先计算样本两两之间的距离,然后将每个样本的平均差异度排序,选择平均差异度较大且与已选聚类中心的差异度大于平均差异度的样本作为初始聚类中心. 实验表明,改进后的算法不仅提高了聚类结果的稳定性和正确率,而且迭代次数明显减少,收敛速度快.

## 1 改进的$K$-均值聚类算法

### 1.1 基本思想

为便于表述,首先定义样本距离、每个样本的平均差异度、样本集的总体平均差异度3个概念.

设样本集$X=\{X_1,X_2,\cdots,X_N\}$, $X_i$为$m$维向量, $X_i\in S_k(t)$, $S_k(t)$为第$t$次迭代的第$k$个类, $N$为

**收稿日期:** 2016-03-08；**修回日期:** 2016-06-01.

**基金项目:** 国家自然科学基金项目(61473118)；湖南省自然科学基金项目(2015JJ2074)；湖南省高校创新平台开放基金项目(13K102)；湖南省科技计划项目(2016TP1021).

**作者简介:** 李武(1977−), 男, 教授, 博士, 从事决策分析、复杂系统建模与优化等研究；赵娇燕(1991−), 女, 硕士生, 从事智能信息处理的研究.

†通讯作者. E-mail: liwu0817@163.com

样本个数, $Z = \{Z_1(t), Z_2(t), \cdots, Z_K(t)\}$ 为 $K$ 个初始聚类中心, $Z_k(t)$ 为第 $t$ 次迭代第 $k$ 个类的聚类中心.

**定义1** 样本 $X_i$ 与 $X_j$ 的距离为 $d_{ij}$. 设 $X_i$、$X_j$ 为两个 $m$ 维模式样本, 若

$$X_i = [x_{i1}, x_{i2}, \cdots, x_{im}],$$
$$X_j = [x_{j1}, x_{j2}, \cdots, x_{jm}], \tag{1}$$

则 $X_i$ 与 $X_j$ 的距离定义[15]为

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{im} - x_{jm})^2}. \tag{2}$$

**定义2** 样本 $X_i$ 的平均差异度为

$$d_i = \frac{\sum_{j=1}^{N} d_{ij}}{N}. \tag{3}$$

**定义3** 样本集的总体平均差异度为

$$M = \frac{\sum_{i=1}^{N} d_i}{N}. \tag{4}$$

作为初始聚类中心的数据对象, 不仅应具有较大的平均差异度, 而且聚类中心之间的差异度要大于样本集的总体平均差异度. 因此, 首先将平均差异度最大的样本作为第一个聚类中心; 然后找出该样本以外的样本集中平均差异度最大的样本, 若该样本与已选定的所有初始聚类中心的差异度均大于整个样本集的总体平均差异度, 则该样本作为第2个聚类中心, 否则针对该样本集中平均差异度第2大的样本进行判断, 直至选出第2个聚类中心, 如此循环, 直至选出所有聚类中心作为初始聚类中心; 最后按传统的 $K$-均值聚类算法进行聚类.

### 1.2 具体算法

根据上述分析, 得到基于平均差异度优选初始聚类中心的改进 $K$-均值聚类算法流程如图1所示, 具体步骤如下.

Step 1: 确定模式类别数 $K, 1 < K < N$, $N$ 为样本个数, $k$ 为第 $k$ 个聚类中心, $t$ 为第 $t$ 次迭代, 令 $k = 1$, $t = 1$.

Step 2: 计算数据对象两两之间的距离 $d_{ij}$.

Step 3: 计算每个数据对象 $X_i$ 的平均差异度 $d_i$ 和样本集的总体平均差异度 $M$.

Step 4: 将平均差异度最大的样本作为第1个聚类中心 $Z_1(t)$, 并将该样本从样本集中删除, $k = k + 1$.

Step 5: 寻找剩余样本集中平均差异度最大的样本 $X_i$, 计算其与已有聚类中心 $Z_1(t), Z_2(t), \cdots, Z_{k-1}(t)$ 的距离, 并将该样本从样本集中删除.

Step 6: 若 $X_i$ 与 $Z_1(t), Z_2(t), \cdots, Z_{k-1}(t)$ 的距离均不小于 $M$, 则 $Z_k(t) = X_i$, 否则返回Step 5.

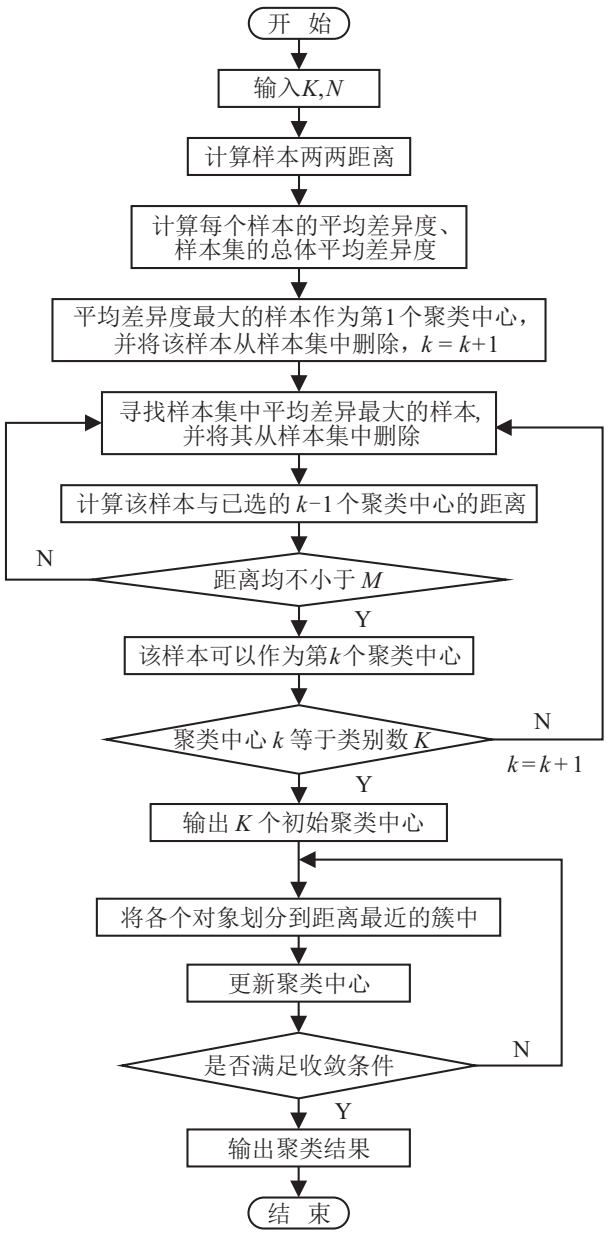


**图1 改进 $K$-均值聚类算法流程**

Step 7: 如果聚类中心个数 $k$ 等于类别数 $K$, 则输出 $K$ 个初始聚类中心 $Z_1(t), Z_2(t), \cdots, Z_K(t)$, 否则, $k = k + 1$, 返回Step 5.

Step 8: 按最小距离原则将其余样本分配到 $K$ 个聚类中心中的某一个, 若

$$\min\{\|X - Z_i(t)\|, i = 1, 2, \cdots, K\} = \|X - Z_k(t)\| = D_k(t), \tag{5}$$

则

$$X \in S_k(t). \tag{6}$$

Step 9: $t = t + 1$, 计算各个聚类中心的新向量值 $Z_k(t)$, 有

$$Z_k(t) = \frac{1}{N_k} \sum_{X \in S_k(t-1)} X, \quad k = 1, 2 \cdots, K. \tag{7}$$

以每类的均值向量作为新的聚类中心, 其中 $N_k$ 为第

$k$类的元素个数.

Step 10: 如果$Z_k(t) \doteq Z_k(t-1), k = 1, 2, \cdots, K$, 则算法收敛, 计算完毕, 否则返回Step 8.

## 2　算法复杂度分析

本文算法主要包括初始聚类中心优选与后续聚类迭代两部分,因此可从这两方面进行算法复杂度分析.

首先计算样本两两距离、样本平均差异度和样本集的总体平均差异度,需要时间复杂度为$O(N^2)$; 然后寻找平均差异度最大,且与已选的聚类中心的聚类均大于总体平均差异度的样本作为聚类中心, 最坏的情况时间复杂度为$O(KN), K < N$; 最后利用传统$K$-均值聚类算法进行迭代,需要的时间复杂度为$O(NKt)$. 所以本文改进算法总的时间复杂度为$O(N^2) + O(KN) + O(NKt)$.

## 3　实验分析

为了验证本文改进算法的有效性,将本文算法、传统$K$-均值聚类算法、其他研究者提出的改进$K$-均值聚类算法进行对比仿真分析. 仿真实验环境如下: 操作系统Windows7, 处理器Intel(R) Core(TM) i3-2120 CPU, 内存4 GB, 仿真软件Matlab 2014a.

实验所用的测试数据集为UCI数据库中用于测试聚类的Iris数据集、wine数据集、glass数据集, 各数据集的基本特征描述见表1.

**表 1　各个数据集的基本特征**

| 数据集 | 样本个数 | 数据样本的维数 | 分类 |
|---|---|---|---|
| Iris | 150 | 4 | 3 |
| wine | 178 | 13 | 3 |
| glass | 214 | 9 | 7 |

将本文算法、传统$K$-均值聚类算法、其他研究者的相关改进算法, 分别针对Iris、wine、glass数据集进行反复实验, 并对聚类结果进行对比分析. 聚类结果的有效性用分类的正确率表示, 结果的效率性用迭代次数表示, 分类正确率的计算式定义为

$$\mathrm{Rate} = \frac{\mathrm{Rt}}{\mathrm{Sm}} * 100\%. \tag{8}$$

其中: Rt为聚类结果中正确分类的数据样本个数, Sm为数据集中数据样本总数.

**表2　改进前后的算法正确率比较**　%

| 数据集 | 传统$K$-均值聚类算法 | | | 本文改进算法正确率 |
|---|---|---|---|---|
| | 最高正确率 | 最低正确率 | 平均正确率 | |
| Iris | 88.67 | 51.33 | 82.37 | 88.67 |
| wine | 64.61 | 56.61 | 61.04 | 62.36 |
| glass | 54.21 | 39.25 | 44.39 | 51.40 |

改进前后的$K$-均值聚类算法的比较结果如表2和表3所示.

**表3　改进前后的算法迭代次数比较**

| 数据集 | 传统$K$-均值聚类算法 | | | 本文改进算法迭代次数 |
|---|---|---|---|---|
| | 最高迭代次数 | 最低迭代次数 | 平均迭代次数 | |
| Iris | 11 | 3 | 7.1 | 1 |
| wine | 12 | 6 | 9.13 | 2 |
| glass | 17 | 11 | 14.33 | 3 |

由表2可见:

1) 采用传统$K$-均值聚类算法进行聚类分析, 对于不同的聚类中心会得到不同的聚类结果, 聚类正确率波动性大, 很不稳定. 这是因为传统$K$-均值聚类算法的初始聚类中心是随机选择的, 并没有考虑到数据的分布情况, 只是单单给出了初始聚类中心.

2) 经多次实验计算其平均正确率发现, 改进后的$K$-均值聚类算法可以得到稳定且较高的正确率. 这是因为改进后的算法是通过启发式算法寻找初始聚类中心, 考虑到了数据集的空间分布, 得到的初始聚类中心更符合实际情况.

由表3可见: 本文提出的改进算法大幅减少了迭代次数, 且迭代次数稳定, 收敛速度快. 这是因为传统的$K$-均值算法的初始聚类中心是随机的, 迭代次数不稳定, 且较差的初始聚类中心会增加迭代次数, 改进后的算法对初始聚类中心的选择已经很接近实际初始聚类中心, 在$K$-均值迭代过程中, 容易满足收敛条件, 收敛速度便会很快.

本文改进的$K$-均值聚类算法与其他文献改进算法的比较结果如表4所示.

**表4　本文改进算法与其他文献算法的迭代次数比较**

| 数据集 | 文献[9] | 文献[10] | 文献[11] | 文献[12] | 文献[13] | 文献[14] | 本文 |
|---|---|---|---|---|---|---|---|
| Iris | 4 | 8.95 | 2 | 8 | 2 | 8.3 | 1 |
| wine | 未提及 | 未提及 | 2 | 9 | 2 | 3.4 | 2 |
| glass | 11 | 11.02 | 未提及 | 未提及 | 4 | 22 | 3 |

由表4可见: 本文改进算法的迭代次数明显低于其他文献的迭代次数, 算法的效率进一步提高, 是一种有效的改进算法.

## 4　结　论

本文针对$K$-均值聚类算法对初始聚类中心的依赖性造成的缺陷, 提出了一种启发式算法寻找初始聚类中心, 有效克服了初始聚类中心选取的随机性和盲目性. 实验结果表明, 改进后的$K$-均值聚类算法显著提高了聚类结果的稳定性和正确率, 且迭代次数减少, 收敛速度较快, 更适合实际数据的分类, 相对于其他诸多改进算法, 具有较明显的聚类优势.

## 参考文献(References)


[1] 黄月, 吴成东, 张云洲, 等. 基于$K$均值聚类的二进制传感器网络多目标定位方法[J]. 控制与决策, 2013, 28(10): 1497-1501.
(Huang Y, Wu C D, Zhang Y Z, et al. Multi-objective localization method based on $K$-means clustering in binary sensor networks[J]. Control and Decision, 2013, 28(10): 1497-1501.)

[2] Bishnu P S, Bhattacherjee V. Software fault prediction using quad tree-based $k$-means clustering algorithm[J]. IEEE Trans on Knowledge & Data Engineering, 2012, 24(6): 1146-1150.

[3] Wu J, Liu H, Xiong H, et al. $K$-means-based consensus clustering: A unified view[J]. IEEE Trans on Knowledge & Data Engineering, 2015, 27(1): 155-169.

[4] Bagirov A M, Ugon J, Webb D. Fast modified global $K$-means algorithm for incremental cluster construction[J]. Pattern Recognition, 2011, 44(4): 866-876.

[5] Tzortzis G, Likas A, Tzortzis G. The minmax $K$-means clustering algorithm[J]. Pattern Recognition, 2014, 47(7): 2505−2516.

[6] Li Y, Wang Q, Chen J, et al. $K$-means algorithm based on particle swarm optimization for the identification of rock discontinuity sets[J]. Rock Mechanics & Rock Engineering, 2015, 48(1): 375-385.

[7] Lin C H, Chen C C, Lee H L, et al. Fast $K$-means algorithm based on a level histogram for image retrieval[J]. Expert Systems with Applications, 2014, 41(7): 3276-3283.

[8] 施侃晟, 刘海涛, 宋文涛. 基于词性和中心点改进的文本聚类方法[J]. 模式识别与人工智能, 2012, 25(6): 996-1001.
(Shi K S, Liu H T, Song W T. A text clustering method based on speech to text and improved center selection[J]. Pattern Recognition and Artificial Intelligence, 2012, 25(6): 996-1001.)

[9] Amorim R C D, Makarenkov V. Applying subclustering and Lp distance in weighted $K$-Means with distributed centroids[J]. Neurocomputing, 2016, 173(3): 700-707.

[10] Lingras P, West C. Interval set clustering of web users with rough $k$-means[J]. J of Intelligent Information Systems, 2004, 23(1): 5-16.

[11] 邢长征, 谷浩. 基于平均密度优化初始聚类中心的$K$-means算法[J]. 计算机工程与应用, 2014, 50(20): 135-138.
(Xing C Z, Gu H. $K$-means algorithm based on average density optimizing initial cluster centre[J]. Computer Engineering and Applications, 2014, 50(20): 135-138.)

[12] 郑超, 苗夺谦, 王睿智. 基于密度加权的粗糙$K$-均值聚类改进算法[J]. 计算机科学, 2009, 36(3): 220-222.
(Zheng C, Miao D Q, Wang R Z. Improved rough $K$-means clustering algorithm with weight based on density[J]. Computer Science, 2009, 36(3): 220-222.)

[13] 何云斌, 肖宇鹏, 万静, 等. 基于密度期望和有效性指标的$K$-均值算法[J]. 计算机工程与应用, 2013, 49(24): 105-111.
(He Y B, Xiao Y P, Wan J, et al. Improved $K$-means algorithm based on expectation of density and clustering validity index[J]. Computer Engineering and Applications, 2013, 49(24): 105-111.)

[14] 赖月霞, 刘建平, 杨国兴. 基于遗传算法的$K$-均值聚类分析[J]. 计算机工程, 2008, 34(20): 200-205.
(La Y X, Liu J P, Yang G X. $K$-means analysis based on genetic algorithm[J]. Computer Engineering, 2008, 34(20): 200-205.)

[15] 徐泽水. 基于相离度和可能度的偏差最大化多属性决策方法[J]. 控制与决策, 2001, 16(增): 818-821.
(Xu Z S. Maximum deviation method based on deviation degree and possibility degree for uncertain multi-attribute decision making[J]. Control and Decision, 2001, 16(S): 818-821.)


（责任编辑：郑晓蕾）