设计研究与应用

K 均值优化算法综述

邓滨玥

（重庆文理学院 电子信息与电气学院，重庆 402160）

摘 要： k-means 算法源于信号处理中的一种向量量化方法，现在则更多地作为一种聚类分析方法流行于数据挖掘领域。在数据挖掘技术中常常使用聚类方法，而 k-means 算法作为最典型、最常见、实用度最广的一种聚类算法，具有简单易操作等优点。但此算法需要人工设定聚类中心的数量，初始聚类中心，容易陷入局部最优，使得算法的时间复杂度变得较大，得到的聚类结果易受到 k 值与设定的初始聚类中心的影响，针对这些问题，本文介绍了 k-means 算法的改进方法，分析其优缺点并提出了优化算法的下一步研究方向。

关键词： k-means 算法；聚类算法；聚类中心；误差平方和；无监督学习

中图分类号： TP391 **文献标识码：** A **DOI：** 10.3969/j.issn.1003-6970.2020.02.041

本文著录格式： 邓滨玥. K 均值优化算法综述[J]. 软件，2020，41（02）：188–192

A Survey on Advanced K-means Algorithm

DENG Bin-yue

(*School of Electronic Information and Electrical Engineering, Chongqing University of Arts and Sciences, Chongqing* 402160, *China*)

【Abstract】: K-means algorithm originated from a vector quantization method in signal processing and is now more popular in the field of data mining as a clustering analysis method. Clustering method is often used in data mining technology, and k-means algorithm, as the most typical, the most common and the most practical clustering algorithm, has the advantages of simple and easy operation. But this algorithm need to manually set the number of cluster centers, the initial clustering center, easy to fall into local optimum, makes the time complexity of the algorithm is larger, the clustering results are susceptible to k value and setting of the influence of the initial clustering center, to solve these problems, this paper introduces the improvement methods of k - means algorithm, analyzes the advantages and disadvantages and puts forward the optimization algorithm of the next research direction.

【Key words】: K-means; Clustering algorithm; Cluster center; SSE; Unsupervised learning

0 引言

在这个数据库技飞速发展的大数据时代，指数型增长的数据对数据的处理分析技术的要求越来越高，人们希望能通过计算机自动智能地在大型数据中，发现有用的信息并预测未来的样本观测结果。随着不断地探索研究，数据挖掘技术在处理数据方面发展已经较为成熟，它在常规数据分析方法的基础上配合复杂算法来处理大规模的数据，已在各个领域的应用中取得了丰硕的成果。

聚类分析将数据划分为有效可使用的组（簇），使得每一个簇内的数据点特征相似。与预测模型不同，聚类中没有明显的目标变量作为数据的属性存在。聚类分析在理解数据与数据预处理领域中都发挥了很大的作用，也是数据挖掘中常为应用的一种算法。

k 均值聚类算法（k- means clustering algorithm）是聚类分析方法中常被使用的一种迭代求解的无监督学习算法，它对数据挖掘应用与大量的模式向量十分重要。因为其步骤简单快速，对大数据效率较高、可伸缩性强，K-means 算法被大量运用在数据挖掘的任务中。但 K-means 的弊端也十分明显，算法常会陷入局部最优，初始质心以及 K 值都需要人为设定，其选择对最后结果影响较大，针对此问题，许多学者对 K-means 算法进行了提升与优化。

本文将介绍 K-means 算法的基本思想和传统 K-means 优化的算法，以及现在学者针对 K-means 主要问题的改进。

作者简介：邓滨玥(1999–)，女，本科生，主要研究方向：信息工程。

1 K-means 算法的基本思想和流程

对于输入样本集 $\{x_1, x_2, \cdots, x_n\}$，K-means 算法在初始化阶段随机生成 K 个数据点作为初始质心 $\{\mu_1, \mu_2, \cdots, \mu_k\}$，计算每一个样本点到各个质心之间的欧式距离，并将每个样本点划到最近的质心所在簇 $\{C_1, C_2, \cdots, C_k\}$，重新计算出质心，不断进行迭代更新质点位置，降低簇的误差平方和 SSE 直至 SSE 不再变化或目标函数收敛时。其中，欧氏距离 d、质心点 μ_i 及 SSE 的计算公式如下：

$$d(x_i, \mu_j) = \sqrt{\sum_{i=1}^{n}(x_i - \mu_j)^2}$$

$$\mu_i = \frac{1}{N_i}\sum_{x \in C_i} x$$

$$SSE = \sum_{J=1}^{k}\sum_{x_i \in C_j} |x_i - \mu_j|^2$$

其中 N_i 为每个簇的样本个数。

K-means 算法的基本步骤如下图：

```
Function K-means()
初始化k个质心μi，使每一个聚类Ci与μi相对应
repeat
    for 每一个输入数据点xi do
        利用欧氏距离将xi分配给最近的质心μi所属聚类Ci
    for 每一个聚类Ci do
        利用质心计算公式更新所有聚类的质心点μi
    计算簇的误差平方和SSE
Until
    SSE不再明显改变或聚类质心点不再变化
```

图 1 K-means 算法步骤

Fig.1 K-means algorithm steps

此算法起始质心的选择对最终结果造成影响较大，其 K 值为用户定义，难以选择达到目标函数最优解的 K 值。

2 经典 K-means 优化算法

2.1 k-means++

由于 k-means 初始质心的选取会对结果造成较大影响，根据文献[1]中描述，在质心选取时让每个质心位置尽可能分散，使其在不同的簇的内部，更便于其优化。

算法的主要流程为：选取 k 个初始质心，利用质心选取公式更新每一个质心所在位置，经过不断迭代直至质心点位置不再变化，其中，质心选取公式如下：

$$P = \frac{D(x)^2}{\sum_{x \in X} D(x)^2}$$

其中 $D(x)^2$ 为每个点到其前一个质心的距离的平方。K-means++能显著的改善分类结果的最终误差。

2.2 二分 k-means

为了减少初始划分情况对聚类结果的影响，以及改进 k-means 算法收敛于局部的问题，提出了二分 k-means 算法，此算法为分层聚类中自顶向下进行分裂的一种方法。

算法的主要思想为：将所有数据点作为一个簇堆，并将其一分为二，计算所有簇堆的误差平方和，并反复选择误差平方和偏大的簇，使用 k-means 算法将其划分，直到簇的数量等于用户所给定的 k 值。步骤图解如图 2 所示。

而由于二分 K-means 算法需要多次采用多次 K-means 方法聚类，增加了其复杂度，刘广聪等[2]提出了用层次聚类与 Chameleon 算法对二分算法进行改进，随机抽取初始聚类中心，寻找离质心最近与最远的两个数据点作为新的聚类中心重新聚类，并通过计算簇间的相似度，建立相似度矩阵来进行优化，提高算法的效率。

2.3 K-medoids

由于 K-means 算法取质点时计算的为当前簇中所有数据点的平均值，K-means 算法对异常值十分敏感，在此问题上，K-medoids 算法对其做出了改进。

在 K-medoids 中，选取当前簇中到同一簇其他数据点距离之和最小的点作为质心，并使用绝对差值和（Sum of Absolute Differences，SAD）代替 SSE 作为衡量聚类结果的标准。SAD 的计算公式如下：

$$SAD = \sum_{m=1}^{k}\sum_{p_i \epsilon C_i} dist(p_i, o_i) = \sum_{m=1}^{k}\sum_{p_i \in C_i}\sqrt{\sum_{j=1}^{n_{c_i}}(p_{ij} - o_{ij})^2}$$

其中，$dist(p_i, o_i)$ 为簇内每个数据点到其质心的曼哈顿距离，k 为质心点数量，p_i 第 i 个簇内的每一数据点，o_i 为第 i 个簇的质心点。

K-medoids 算法较好的解决了 K-means 算法对噪点敏感的问题，但是算法的时间复杂度较大，只适用于样本小的情况。

对此，Cao 等人提出一种基于 CF 树的 K-medoids

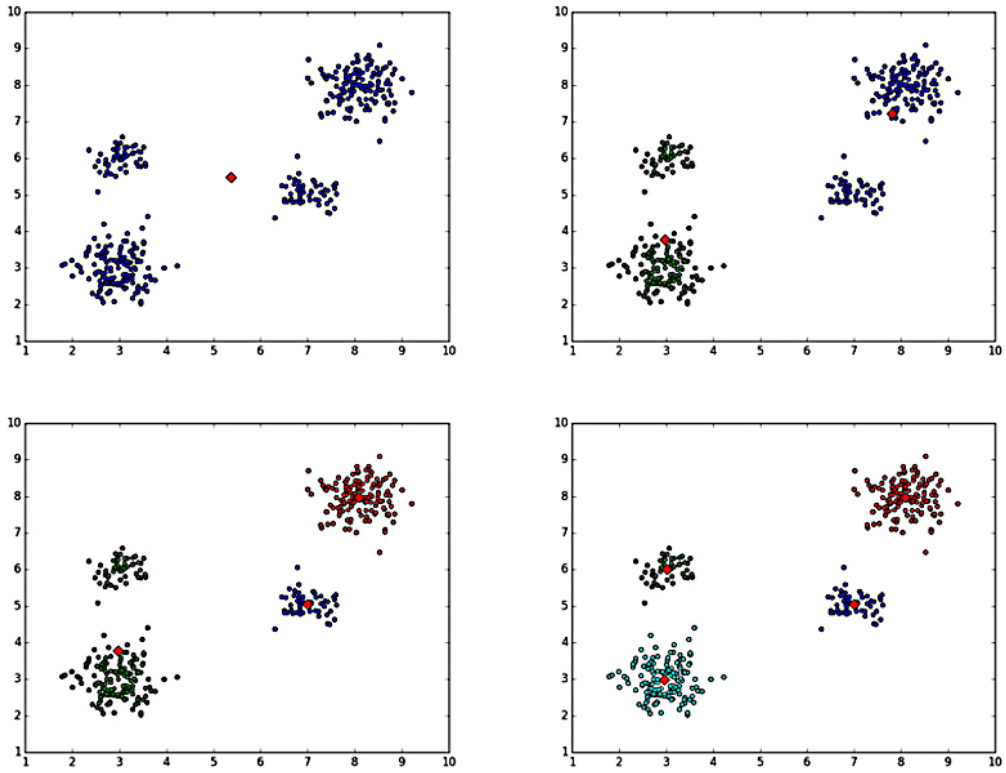

图 2 二分 K-means 聚类过程

Fig.2 Binary K-means clustering process

算法[3]，增强算法的可伸缩性，Park 等人提出了快速 K-medoids 算法[4]，根据数据点聚集密度大的 K 个点作为初始聚类中心，提高聚类效果。针对 K-medoids 对初始聚类中心敏感问题，谢娟英等[5]提出一种基于局部方差的 K-medoids 优化算法，用局部方差度量样本分布密度程度，用局部标准差作为邻域半径计算出样本的邻域集合，加快算法的收敛速度。此算法中样本 x_i 的局部方差及标准差 L_1 计算公式如下：

$$F\left(x_i\right)=\frac{\sum_{j=s_i}^{S_{Num}}[d(i,j)-aver(xi)]^2}{Num-1}$$

$$L_1=\sqrt{F(x_i)}$$

其中，$aver\left(x_i\right)=\frac{\sum_{j=s_i}^{S_{Num}}d(i,j)}{Num}$，$d(i,j)$ 为某一样本点到数据及其他样本点的距离，并按距离进行升序排序。

文献[6]针对快速 K-medoids 初始聚类中心可能位于同一类簇及传统 K-medoids 算法的缺陷，提出基于粒计算的 K-medoids 聚类算法，利用等价关系产生粒子，并根据粒子包含的样本个数定义粒子密度，从而选择密度较大的 K 个例子作为初始聚类中心，使得此算法聚类结果更加稳定，并可适用于大规模的数据集。郝占刚[7]等提出一种基于遗传算法和 K-medoids 算法的聚类新算法，此算法采用遗传算法中的锦标赛选择法随机选择一定数目的样本，并结合 k-medoids 对选择出的个体进行优化，代替原有个体，不断进化直到结果符合要求，这种算法可以很好地解决 k-medoids 算法局部最优与孤立点的问题，并加快了遗传算法的收敛速度。

3 k-means 算法改进

3.1 基于 k 值选择

在 K-means 算法中，由于初始质心点数 k 需要使用者指定，不同 k 值选择所得出的聚类结果也不一样，如何确定最优 k 值或让算法自动获取 k 值成为学者改进 k-means 算法的一个目标。

1998 年，Razaee[8][8]提出最佳的聚类数应该在 2 与 $\sqrt{n}$ 之间，其中 n 为所有数据点的个数，此结论为后期 k 值选取优化方面提供了基础。由于传统二分 k-means 时间复杂度较高，张忠平[9]等人对此算法进行改进，提出基于二分思想的确定聚类数目的算法，由于每次二分只需要两个初始聚类中心，减少聚类

中心对其算法结果的影响。KDM 算法[11][10]通过最大最小距离法和数据存储方法，划分时不断调整聚类中心实现了 k 值的自动确定，同时实现了初始中心的选择。

之后有学者提出使用“手肘法”选择肘点作为最优的 K 值，此方法简单直观但可能会出现不明显的“肘点”或是特殊情况使得 K 值的选择出现偏差，文献[11]ET-SSE 算法对此进行了 k 值选择的优化，引入偏执项调节变量改进总误差平方和，通过对权重的调节得出最终 k 值。

3.2 基于局部最优问题

由于 K-means 算法对初始点以及噪点十分敏感，常常会收敛到局部最小值而引起聚类结果的偏差，通过算法对噪点的处理以及迭代过程中划分规则的改变可以解决此问题以达到全局最优。

陈慧萍等[12]采用模拟退火思想提出了一种全局寻优的 K-means 方法，设定目标函数及控制参数，不断迭代调整控制参数 t（各聚类中心的值）直到得出当前近似最优解，得以得到最优解。PBK-means 算法[13]提出基于距离与密度，计算数据集的平均样本距离，根据数据点之间的距离计算数据权重，从而选取最大权重数据作为第一个中心点，将数据集进行分类，并建立满二叉树，合并叶子结点得到 k 个初始聚类中心，快速处理中小型规模的数据集。

3.3 初始中心选择

K-means 一般采取随机选择的方式确定初始质心，而这样不仅会使得算法的时间复杂度增大，并且可能会选取到离群点导致结果差异很大，现代学者更偏向通过与其他算法相结合的方式获得较准确初始质心。

Redmond[14]等人最早提出通过 kd-tree 从带划分的数据集中筛选密度大又相互分离的数据作为初始中心，而由于此方法在估计数据密度方面存在缺陷，基于此方法，后代学者提出了对应的改进。文献[15]提出基于最小支撑树，选中密度大且足够分离的数据稠密区中的点作为初始聚类中心，使得算法可以在选出处在不同类的数据作为初始中心。文献[16]提出一种利用关系矩阵和度数中心度的分析方法来选取初始中心点，减少聚类过程的迭代次数得到更稳定的聚类结果，但此方法在处理大规模数据问题上还存在局限性。

3.4 其他改进方法

Dan Pelleg[17]等在 2000 年提出一种 x-means 的聚类方法，运用统计学标准将样本的似然函数最大化，通过计算 BIC score 来决定是否将簇二分，算法的主要步骤图如下：

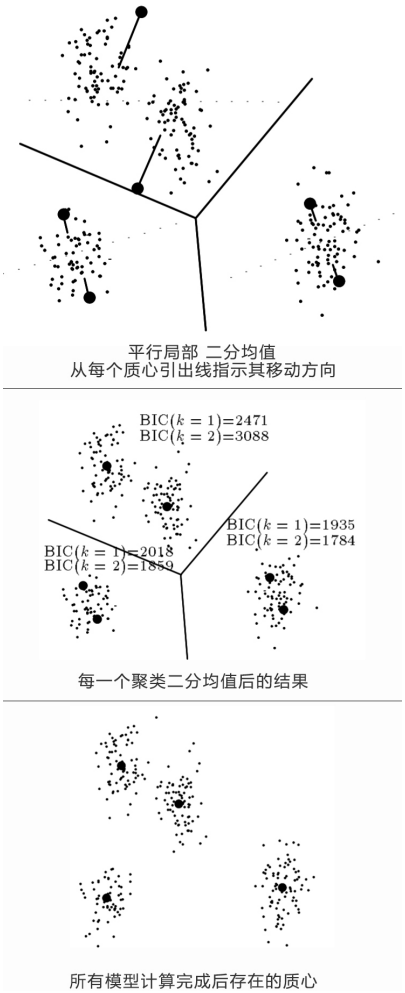

图 3 x-means 算法聚类过程

Fig.3 x-means algorithm clustering process

其中 BIC score 以及最大似然函数计算公式如下：

$$BIC(M_j)=\hat{i}_j(D)-\frac{P_j}{2}logR$$

$$\sigma^2=\frac{1}{R-K}\sum_i(x_i-\mu_{(i)})^2$$

其中 $\hat{i}_j(D)$ 为数据模型的似然函数，P_j 为模型的复杂度，R 为输入集中的数据点数。

此方法不用预先指定 k 的个数，只需要给出 k 值范围，很好地解决了 k-means 算法 k 值难以确定的问题，对大规模的数据也具有很好的效率，但是不适用于高维数据中。

此外，还有很多学者分别提出了基于 Spark 框架[18]、MapReduce 框架[19]、Hadoop[20]框架等常见数据计算平台来改进 K-means 算法，通过并行计算提

高聚类提速。

在 d 维空间中找到 k-均值聚类问题的最优解的计算复杂度：

- NP-hard：一般欧式空间中，即使目标聚类数仅为 2
- NP 困难：平面中，不对聚类数目 k 作限制
- 如果 k 和 d 都是固定的，时间复杂度为，其中 n 为待聚类的观测点数目

4 结束语

作为聚类算法中较为经典的 K-means 算法，因为计算快速方便被广泛应用在数据挖掘等大数据处理方面，由于其缺点也十分明显，在提出后便不断有学者针对这些问题进行优化与改进，但在对算法进行改进时将会牺牲其他各方面的指标。所以在优化 k-means 算法三个主要问题的同时，如何有效地缩短算法的复杂度、使算法能够适用于多维度问题以及大规模数据问题等将成为学者们的下一步的研究方向，尤其是在机器学习技术的日益丰富的背景下，各种聚类算法与机器学习相结合，各种优化方案等更是以后的攻坚工程。

参考文献：

[1] Agarwal M, Jaiswal R, Pal A. k-means++ under Approximation Stability[C]//International Conference on Theory and Applications of Models of Computation. Springer, Berlin, Heidelberg, 2015.

[2] 刘广聪，黄婷婷，陈海南．改进的二分K均值聚类算法[J]. 计算机应用与软件, 2015(2): 261-263.

[3] 曹丹阳，杨炳儒，李广原，等．一种基于CF树的k-medoids聚类算法[J]. 计算机应用研究, 2011(9): 66-69.

[4] PARK H S, JUN C H. A simple and fast algorithm for K-medoids clustering[J]. Expert Systems with Applications, 2009, 36(2): 3336-3341.

[5] 谢娟英，高瑞．Num-近邻方差优化的K-medoids聚类算法[J]. 计算机应用研究, 2015, 32(1).

[6] 马箐，谢娟英．基于粒计算的K-medoids聚类算法[J]. 计算机应用, 2012, 32(7): 1973-1977.

[7] 郝占刚，王正欧，HaoZhangang，等．基于遗传算法和k-medoids算法的聚类新算法[J]. 现代图书情报技术, 2006(5).

[8] Rezaee M R, Lelieveldt B B F, Reiber J H C. A new cluster validity index for the fuzzy c-mean[M]. Elsevier Science Inc. 1998.

[9] 张忠平，王爱杰，柴旭光．简单有效的确定聚类数目算法[J]. 计算机工程与应用, 2009, 45(15): 166-168.

[10] 徐克圣，王澜，XUKe-sheng，等．一种自动获得k值的聚类算法[J]. 大连交通大学学报, 2007(4).

[11] 王建仁，马鑫，段刚龙．改进的K-means聚类k值选择算法[J]. 计算机工程与应用, 2019, 55(8): 33-39.

[12] 陈慧萍，贺会景，陈岚峰，等．基于模拟退火思想的优化k-means算法[J]. 河海大学常州分校学报(4): 33-36+44.

[13] 魏文浩，唐泽坤，刘刚．基于距离和密度的PBK-means算法[J/OL]. 计算机工程: 1-9[2019-11-17].

[14] Redmond S J, Heneghan C. A method for initializing the K-means clustering algorithm using kd-trees[J]. Patten Recognition Letter, 2007, 28: 965-973.

[15] 李春生，王耀南．聚类中心初始化的新方法[J]. 控制理论与应用, 2010, 27(10): 1435-1440

[16] 郁启麟．K-means算法初始聚类中心选择的优化[J]. 计算机系统应用, 2017(5).

[17] Pat Langley. Proceedings of the Seventeenth International Conference on Machine Learning[C]//2000.

[18] 宋董飞，徐华．基于Spark的K-means改进算法的并行化实现[J]. 计算机系统应用.

[19] 毛典辉，北京工商大学计算机与信息工程学院，北京，．基于MapReduce的Canopy-Kmeans改进算法[J]. 计算机工程与应用, 2012, 48(27): 22-26.

[20] 卢胜宇，王静宇，张晓琳，等．基于Hadoop平台的K-means聚类算法优化研究[J]. 内蒙古科技大学学报, 2016, 35(03): 264-268.