文章编号:1005-0523(2022)05-0119-08

K-means聚类算法研究综述

王 森,刘 琛,邢帅杰

(华东交通大学理学院,江西 南昌 330013)

摘要:聚类分析是数据挖掘的重要技术,而在5G时代,海量的数据维度高、数据集大,利用K-means算法易受离群点的影响,且K值、初始聚类中心的选取影响聚类结果的稳定性和准确率,甚至导致聚类陷入局部最优,对K-means算法的改进受到众多研究者的关注。主要对K-means聚类的研究现状进行归纳总结。首先,介绍K-means算法的思想原理;其次,针对初始聚类中心点的选取、K值确定、离群点对现有改进算法进行基于密度和距离的分类总结,并对各个改进算法的优势和缺陷进行分析;最后对K-means算法未来可能的研究方向和趋势进行展望。

关键词:K-means;聚类算法;K值;初始聚类中心;离群点;密度;距离

中图分类号:TP301.6 **文献标志码**:A

本文引用格式:王森,刘琛,邢帅杰. K-means聚类算法研究综述[J]. 华东交通大学学报,2022,39(5):119-126.

DOI:10.16749/j.cnki.jecjtu.20220914.001

Review on K-means Clustering Algorithm

Wang Sen, Liu Chen, Xing Shuaijie

(School of Science, East China Jiaotong University, Nanchang 330013, China)

Abstract: Cluster analysis is an important technique for data mining. In the 5G era, massive data has high dimensions and large data sets. The K-means algorithm is susceptible to outliers, and the k value and the selection of initial clustering centers affect the stability and accuracy of the clustering result. It even causes the clustering to fall into the local optimum, so the improvement of the K-means algorithm has attracted the attention of many researchers. This article mainly summarizes the current research status of K-means clustering. Firstly, it introduces the principle of K-means algorithm. Secondly, according to the selection of the initial clustering center point, the determination of the K value, and the outliers, the existing improved algorithms are classified and summarized based on density and distance, and the advantages and disadvantages of each improved algorithm are analyzed. Finally, the K-means algorithm is analyzed and prospects for possible future research directions and trends are discussed.

Key words: K-means; clustering algorithm; K value; initial cluster center; outlier; density; distance

Citation format: WANG S, LIU C, XING S J. Review on K-means clustering algorithm[J]. Journal of East China Jiaotong University, 2022, 39(5): 119-126.

随着计算机和网络技术的日益成熟,大数据时代快速发展,每时每刻都在产生大量的数据信息,并且越来越复杂。而聚类分析是数据挖掘中数据划分和数据分组的重要方法,一直以来受到众多研究者

收稿日期:2021-09-15

基金项目:江西省自然科学基金项目(2019ZACBL20010)

的关注,在生物学和医学[1-4]、市场营销[5-6]、文本分类[7]、机器学习[8-11]等各个领域得到广泛应用。

聚类算法是一种区别于分类的无监督学习算法,即对无标签的样本点根据数据及数据间的信息关系,对数据对象进行分组。聚类的最终目的使组内的对象之间相似,不同组中的对象之间有区别[12]。组内的样本点相似性越大,组间相似性越小,则聚类效果越好。目前广泛研究的聚类算法大致可以分为以下几类:基于层次的聚类,基于密度的聚类,基于原型的聚类,基于划分的聚类。

划分聚类是将数据集划分成不重叠的子簇,使得最终每个数据点恰好位于一个子簇中,而且各个子簇的并恰是整个数据集,经典算法有 K-means 算法,AP 算法。层次聚类是嵌套簇的集族,组织成一棵树。除去叶节点外,树中每一个结点(簇)都是其子簇的并,树根则是包含所有对象的簇,其中聚类树的构建分为凝聚的层次聚类和分裂的层次聚类,经典算法有 AGNES 算法,DIANA 算法。基于密度的聚类则是通过寻找被低密度区域所分离的高密度区域的方法而进行的聚类,算法有 DBSCAN 算法,DPC 算法。

K-means 算法最早是由 Mac[13]于 1967 年提出,是非常经典的算法,目前仍是非常流行的“十大算法”之一。该算法由于其原理简单、效率高而被人们广泛使用,但是聚类结果也易受其他因素的影响,例如 K 值的确定,离群点,初始聚类中心的选取,随意选取不同的点作为初始聚类中心将会导致不同的聚类结果,甚至可能陷入局部最优。

首先,概述 K-means 算法的基本思想原理;其次,分别对关于初始聚类中心的选取、K 值的确定和离群点三方面的改进算法进行基于密度和距离方面的分类总结,分析各个改进算法的优缺点;最后,展望 K-means 算法在未来的研究方向和发展趋势。

1 传统 K-means 聚类算法

1.1 K-means 算法基本思想

K-means 聚类算法是一种简单的迭代性聚类算法,是对一个 n 维向量的数据点集 $\boldsymbol{D}=\{x_i|i=1,\cdots,N\}$进行聚类,其中 x_i 表示第 i 个数据点,最终将集合 $\boldsymbol{D}$ 划分成 k 个类簇。分组的依据主要是“紧密度”或者“相似度”,组内对象越相似、组间差距越大越好[14]。距离的度量有欧几里得距离、曼哈顿距离和切比雪夫距离,聚类算法通常用欧氏距离作为相似性度量,以误差平方和(sum of squared error, SSE)作为度量聚类质量的目标函数,通过最小化目标函数,将数据点按照距离聚类中心的远近分成 k 个簇。

定义 1 数据点之间的欧几里得距离

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{1}$$

定义 2 数据点与聚类中心 c_i 间的欧几里得距离公式

$$d(x,c_i)=\frac{1}{2} \tag{2}$$

式中:c_i 为第 i 个类簇的聚类中心;x 为数据集 D 中的数据对象。

定义 3 误差平方和 SSE 计算公式

$$SSE=\sum_{i=1}^{k}\sum_{n\subset c_i}^{n}|d(x,c_i)|^2 \tag{3}$$

式中:C_i 为第 i 个类簇;c_i 为簇 C_i 的聚类中心;x 为数据集 $\boldsymbol{D}$ 中的数据对象;k 为类簇数目。

K-means 聚类算法是一种动态聚类算法,需要进行不断重复迭代。该算法的基本思想原理是:选择 K 个数据作初始聚类中心,其中 K 为用户指定的类簇个数。计算数据点到各个初始聚类中心的距离,将数据点就近分配到各个初始聚类中心所在的集合中形成一个簇。根据簇中的各个点,更新每个簇的中心。重复分配和更新的步骤,直到簇不在发生变化,或者目标函数满足条件即可。

2 K-means 算法的改进

2.1 初始聚类中心的选取

K-means 算法中,初始聚类中心点的选取对聚类结果的影响非常大,对于不同的初始聚类中心,最终的聚类的结果往往不同,所以 K-means 聚类算法的稳定性较差,并且聚类中心的选择会导致出现聚类结果陷入局部最优的问题。目前针对初始聚类中心的选择的改进研究主要从密度和距离两个方面入手。

2.1.1 基于密度

目前众多学者提出基于密度的 K-means 算法改进方案,主要依据数据集中数据点的密度分布来选取,避免由于随机选取的初始聚类中心过于密集,导致聚类陷入局部最优的情况。

蔡宇浩等[15]提出了 WLV-K-means 算法，通过对局部方差进行加权来优化初始聚类中心。方差表示了数据的离散程度，方差越小则离散程度越小，密度越大。该算法首先计算样本数据的邻域半径 θ 和邻域内部的样本数据到样本中心的方差，以加权的局部方差为新的密度度量。再对各个样本的加权局部方差进行排序，将加权局部方差最小即密度最大的数据点作为第一个初始聚类中心。最后利用改进的最大最小法，即对加权局部方差作倒数作为密度系数对最大最小值进行加权，进而找到各个初始聚类中心。WLV-K-means 算法的优点在于引入加权局部方差，提升了计算的准确性，减少了噪声点对聚类效果的影响，但同时增加了算法时间复杂度和空间复杂度，针对数据集中样本的数目和分布不同，对应的最优半径调节参数 θ 也不同，需要多次调节寻找合适的调节参数值。薛印玺等[16]提出基于样本密度的全局优化 K 均值聚类算法 KMS-GOSD 算法，该算法通过高斯模型得到迭代初始所有聚类中心的预估计密度，在更新聚类中心迭代过程中加入偏移操作，引入衰减因子 $Ra=(m-1)/m$，其中 m 为最大的迭代次数，以此逐步降低预估计密度，加速偏移收敛，通过比较实际密度和逐渐减小的预估计密度，得到密度较高的聚类中心。KMS-GOSD 算法极大地避免了质心作为聚类中心时陷入局部最优的可能，降低了算法复杂度，并且增强了聚类中心点对全局的探索能力，具有较高的准确率和稳定性。

2.1.2　基于距离

目前，有学者基于欧氏距离提出相异度概念，通过构造相异度矩阵进而判断数据样本点之间的差异，并借此判断数据样本点能否作为初始聚类中心。孟子健等[17]定义了两个数据点间的第 k 个属性的相异度且对数据进行了标准化处理

$$r_{if}^{(k)}=\frac{|x_{ik}-y_{ik}|}{\max x_{R_k}-\min x_{R_k}} \tag{4}$$

式中：x_{ik} 和 x_{jk} 分别为数据点 x_i 和 x_j 的第 k 个属性值；x_{R_k} 为 $\boldsymbol{D}$ 中所有数据点的第 k 个属性的所有取值。并由此给出了两数据点间的相异度公式

$$r_{if}^{(k)}=\sum_{k=1}^{N} r_{if}^{(k)} \tag{5}$$

式中：N 为数据的维数。

进而构造相异度矩阵

$$\boldsymbol{R}=\begin{bmatrix} 0 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & 0 & r_{23} & \cdots & r_{2n} \\ r_{31} & r_{32} & 0 & \cdots & r_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \cdots & 0 \end{bmatrix} \tag{6}$$

又定义了样本的平均相异度

$$\bar{r}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij} \tag{7}$$

对于给定的数据点 x_0，通过计算点 x_0 的相异度参数 $N(x_0,\bar{r})$ 即以 x_0 为中心，以 $\bar{r}$ 为半径的区域内点的个数

$$\left.\begin{aligned} \delta(y)&=\begin{cases}1, y\leqslant 0\\ 0, \text{其它}\end{cases}\\ N(x_0,\bar{r})&=\sum_{i=1}^{n}\delta(x_{0i},\bar{r})\end{aligned}\right\} \tag{8}$$

数据点 x_0 的相异度参数 $N(x_0,\bar{r})$ 越大，数据点 x_0 的邻域内的点越多，则 x_0 更有可能为某一簇的初始聚类中心。

计算数据集 $\boldsymbol{D}$ 中的每个数据点的相异度参数，并组成一相异度参数集合 M，从集合 M 中找出最大的参数所对应的数据点作为第一个初始聚类中心，并将该数据点和与该数据点的相异度小于 $\bar{r}$ 的所有数据点从集合 M 中删去，重复计算剩下数据点的相异度参数，直到找出 K 个初始聚类中心为止。该算法消除了聚类对初始聚类中心的敏感性，提高了聚类的准确率。但是该算法的缺点是当最大相异度参数所对应的点不唯一时，不能够合理地选取初始聚类中心。董秋仙等[18]对算法此缺陷进行改进，找出最大的相异度参数值对应的所有样本点 x_i 计算

$$\mathrm{sum}(i)=\sum_{j=1}^{p} r_{ij} \tag{9}$$

式中：$r_{ij}\leqslant\bar{r}$，$j=1,2,\cdots,p$，且构成集合 sum，若 sum(i)=min sum，则第 i 个样本点即为第一个初始聚类中心。通过此算法，可在相异度参数最大值不唯一时，找到合适的第一个初始聚类中心，实验表明，此改进算法比原算法具有更高的准确率，同时减少了迭代次数。廖纪勇等[19]利用欧氏距离定义了数据点间的相异性，均值相异性和总体相异性，构造相异性矩阵，提出了 IK-DM 算法。均值相异性反映了数据点在数据集 $\boldsymbol{D}$ 中的分布情况，值越大，则数据点 x_i 的密度越低，与其他点距离越远。IK-DM 算法通过计算，选取均值相异性最大的点作为第一个初始聚

类中心,选择均值相异性第二大的点作为临时第二个初始聚类中心，通过计算该点与已有初始聚类中心间的相异性,若大于总体相异性,则该点确立为第二个初始聚类中心，否则找均值相异性次大的点进行判断,依次找出所有的初始聚类中心。该算法降低了离群点对聚类结果的影响，有效地选择数据集 $\boldsymbol{D}$ 中相异性较大的数据点作为初始聚类中心,避免选择的中心点过于密集,显著提高了算法的稳定性和准确率，但算法的执行速度较为缓慢,耗时略长。

2.2 K 值的选择

K 值的选取极大地影响了 K-means 聚类结果,而聚类算法中 K 值的选择往往需要事先指定,且多数是根据历史经验或者多次尝试中得到的。Rezaee等[20]根据实验证明,最佳 K 值位于区间[1,n]中。为了得到更准确的类簇数目,学界尝试从各个方面对聚类算法进行深入研究,提出了多种改进算法。

2.2.1 基于密度

贾瑞玉等[21]首先通过计算样本密度选出初始聚类中心,在此基础上计算对应不同的 K 值时,优化传统聚类有效性指标 BWP 为 IBWP 指标，更好地反应了单个数据点的聚类效果,IBWP 指标值越大,聚类效果越好,以此求得最佳聚类数目 K。该算法改进了聚类有效性指标,具有较好的准确性和稳定性,但增加了算法的时间复杂度,运行时间过长。贾瑞玉等[22]提出 CNACS-K-means 算法,该算法重新定义数据点的局部密度，构造数据集 $\boldsymbol{D}$ 的决策图,通过残差分析自动确定类簇数目和初始聚类中心。实验表明,CNACS-K-means 算法在二维和高维数据集上具有较高的准确性,能自动确定聚类类簇数目,该算法的缺点是不同数据集的数据分布会对聚类效果造成一定的影响,对于分布比较稀疏的数据集聚类效果不理想。

2.2.2 基于距离

众多学者基于欧氏距离与误差平方和 SSE 来确定聚类数目 K,随着 K 值增大,则聚类数目越多,类内差距会越来越小且 SSE 会逐渐减小。当 K 值小于真实聚类数时,随着 K 值增大,SSE 下降幅度大,当 K 值等于真实聚类数时,再增加 K 值,则斜率会迅速增大,随着 K 值增大,折线图趋于平缓;因此拐点处对应的 K 值多数为最佳 K 值。但是,对于某些数据集来说,“拐点”并不明显,而得到的是一个“拐点区间”，则无法确定准确 K 值，导致实验出现偏差。王建仁等[23]改进“手肘法”为 ET-SSE 算法,针对“手肘法”中的“肘点”不明确问题进行改进。改进算法利用指数函数的指数爆炸性质,引入偏执项提高聚类误差较大簇的 SSE 的比重,引入权重 θ 进行放缩调节使得该算法能更快更准地确定 K 值。该算法有效地解决了“肘点不明确”的问题,提高准确率的同时,降低时间复杂度,缺点是权重 θ 的调节需要根据实验进行调整，无法根据最佳权重进行实验。王子龙等[24]对欧氏距离进行维度加权的改进,并且引入样本点的权重 w_i 和参数 τ_i，通过改进后的欧氏距离计算样本的密度和权重，选取密度最大的样本点作为第一个初始中心点，再利用样本点的权重和参数 τ_i 以此得到下一个初始中心点。该算法明显改进了正常数据点与离群数据点到聚类中心的距离，使得离群数据点到聚类中心之间的距离变大,放大差异,同时,权重的引入避免选取噪声点作为聚类中心,避免聚类陷入局部最优。在中小型数据上,该算法提高了运行速度,减少了算法的迭代次数,但针对大型数据,提高了算法的时间复杂度,耗时较长。

2.3 离群点筛选

聚类过程中,离群点的存在一定程度上影响了初始聚类中心的选取,导致离群点可能成为初始聚类中心,增加了迭代次数,降低了算法效率。目前对于离群点的研究也成为一个热门方向,众多研究者尝试通过多种方法对数据进行预处理来降低离群点对聚类的影响,以此提升算法效率。

2.3.1 基于密度

唐东凯等[25]提出了 OFMMK-means 算法,通过使用 LOF 算法——基于密度的离群点检测算法来排除离群点的影响，再根据最大最小法选取初始聚类中心,进行聚类。该聚类算法首先对数据集 $\boldsymbol{D}$ 进行离群点检测，计算出每个数据点的离群因子值。数据点的离群因子值分布体现了该点是离群点的概率,值越大,则该点为离群点的概率越大。再根据离群因子值的大小对数据集 $\boldsymbol{D}$ 中的数据点进行升序排列,取前 $\alpha n(0<\alpha\leqslant 1, n=|\boldsymbol{D}|)$ 个数据点作为初始聚类中心的候选集。最后根据最大最小法[26]在初始聚类中心的候选集上选取距离尽可能

远的数据点作为初始聚类中心，再进行 K-means 聚类。OFMMK-means 算法提高了算法的稳定性，具有较高的聚类准确率，同时，相较于传统 K-means 算法，聚类平均迭代次数也较小，但参数 α 的取值需要依据实验确定，无法得到最佳准确值。杨红等[27]利用 LOF 算法进行离群点检测，对数据集进行数据预处理，得到密集点数据集 iris-1 和离群点数据集 iris-2，接着对 iris-1 数据集进行 K-means 聚类，根据准则函数得到各个簇，其中准则函数作了改进

$$SSE_2=\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\frac{SSE}{(c_i-c_j)} \tag{10}$$

最后，把离群点数据集中的点根据距离最近原则分配到各个簇中。该算法主要针对传统准则函数仅仅考虑类内相似性的缺点进行改进，改进准则函数更考虑到类间差异性，放大了聚类效果，进一步优化了聚类结果。刘凤等[28]利用局部密度离群值检测方法检测并去除数据集 $\boldsymbol{D}$ 中离群值，对剩余点进行 K-means 聚类，通过 DB 指标、Dunn 指标、Silhouette 指标进行有效性评价，以此评估数据集聚类结果，提高了聚类稳定性。

2.3.2　基于距离

基于距离的离群点检测方法计算简单，可以用于任何可以计算数据点的数据集中，且与数据集的分布无关。冷泳林等[29]利用基于距离的离群点检测首先随机选一数据点，计算该数据点与其他数据点间的欧氏距离，如果距离大于 d 的数据点的比例大于参数 p，则认为该点为离群点，依次测验数据集中 $\boldsymbol{D}$ 中的数据点，找出所有离群点。接着，在非离群点中随机选取 K 个初始聚类中心，对非离群点进行 K-means 聚类，最后将离群点按照距离最近原则分配到各个簇中。该算法降低了离群点对与聚类的影响，提高了聚类精度和稳定性，但对于高维数据集来说，聚类效果并不十分明显。Milos 等[30]研究了基于距离法的高维数据集的离群点检测，发现基于距离的方法可在高维数据中产生更具对比性的异常值分数，通过比较异常值分数进而得到离群点。此改进算法解决了高维数据集之中基于距离法检测离群点的问题。

3　K-means 改进算法对比分析

表 1、表2 和表 3 分别给出选取初始聚类中心的改进算法对比、确定 K 值的改进算法对比以及筛选离群点的改进算法对比。

表 1　初始聚类中心选取的改进算法对比

Tab.1　Comparison of improved algorithms for selecting initial cluster centers

Improved algorithm	Main idea	Advantage	Shortcoming
WLV-K-means	Weighted local variance, weighted maximum and minimum	Improve clustering accuracy and reduce the influence of noise points	Increase the time complexity, the optimal radius adjustment parameters θ are different
KMS-GOSD	Global optimization	Reduce algorithm complexity, enhance global exploration capabilities, higher accuracy and stability	Increased algorithm time complexity
Improved algorithm based on dissimilarity matrix	Select the initial center according to the maximum dissimilarity parameter value	Improved clustering accuracy	The point corresponding to the maximum value is not unique, and the initial center cannot be selected reasonably
IK-DM	Mean dissimilarity reflects the distribution of data points, and overall dissimilarity is used as the judgment condition	Improve the stability and accuracy of the algorithm to avoid too dense center points	The execution speed is slow and takes a little longer

表2 K值确定的改进算法对比

Tab.2 Comparison of improved algorithms for determining K value

Improved algorithm	Main idea	Advantage	Shortcoming
Improve BWP index based on sample density	The larger the IBWP value, the better the clustering effect	High clustering accuracy and stability	Increase algorithm time complexity
CNACS-K-means	Redefine local density, self-determined by residual analysis	Good accuracy on two-dimensional and high-dimensional data sets, and automatically determine the K value	Poor clustering effect on sparse data sets
ET-SSE	The nature of the exponential function, introducing paranoid terms and weights	Solve "elbow point is not clear", improve accuracy and reduce time complexity	The optimal weight value cannot be determined and needs to be adjusted based on experiments
Improved Euclidean distance algorithm based on dimensionality weighting	Dimensional weighting based on Euclidean distance, introducing sample point weights and parameters	Increase running speed and reduce the number of iterations	Running time is too long for large data sets

表3 筛选离群点的改进算法对比

Tab.3 Comparison of improved algorithms for screening outliers

Improved algorithm	Main idea	Advantage	Shortcoming
OFMMK-means	Fusion of LOF algorithm and max min method	Improve clustering accuracy and stability, reduce the number of algorithm iterations	The value of the parameter depends on the data set
Improved algorithm for outliers based on distance	Outlier detection algorithm based on distance to screen outliers	Reduce the influence of outliers, improve clustering accuracy and stability	The effect is not obvious on high-dimensional data sets

4 结论

K-means算法是一个极其经典的聚类算法,自提出以来,以其思想简单、聚类速度快、结果良好而得到广泛应用。但该算法也存在缺陷,本文主要对初始聚类中心的选取、K值的选择、离群点的筛选问题进行论述,并且对各个缺陷分别基于密度和距离两个方面进行详细阐述,分析各个改进算法的优缺点。

随着近年来K-means算法应用领域的逐渐拓宽,学术界对算法进行的改进也逐渐增多,但多数改进算法聚类效果的提高多是以增加时间复杂度作为代价,对于多维数据应用时,算法有效性尚不显著,在后继优化研究中,可以考虑从以下几个方面入手。

1)在提升聚类效果的同时降低时间复杂度。

2)考虑提升算法处理高维数据集的能力,随着时代快速发展,各种数据信息繁冗复杂,能够利用K-means算法快速高效地处理多维数据也是未来重要的研究方向。

3)目前,越来越多的K-means改进算法中,多数改进算法会引入参数,但对于参数值的确定需要依据实验经验调试,无法确定准确参数值,这需要进一步研究。

参考文献:

[1] 韩庆华,马乾,刘名,等. 温度变化下基于固有频率聚类分析的空间网格结构损伤诊断[J]. 华东交通大学学报,2021,38(4):8-17.
HAN Q H,MA Q,LIU M,et al. Damage diagnosis of space grid structure based on natural frequency clustering analysis under varying temperature effects[J]. Journal of East China Jiaotong University,2021,38(4):8-17.

[2] NAWAZ M,MEHMOOD Z,NAZIR T,et al. Skin cancer detection from dermoscopic images using deep learning and

fuzzy K-means clustering[J]. Microscopy Research and Technique，2021，85(1)：339-351.

[3] GUO Y M，TANG Y，DU Y，et al. Cluster analysis of polymers using laser-induced breakdown spectroscopy with K-means[J]. Plasma Science & Technology，2018(6)：99-103.

[4] RAVI B T，SHAH D U，JOSHI N P. Performance comparison of cluster number of K-means clustering algorithm using mammographic image[J]. Digital Image Processing，2014，6(4)：197-200.

[5] 谭征. 基于 K-Means 和 SEM 的消费者互联网保险购买意愿研究——以 TPB 和 TAM 为分析框架[J]. 重庆理工大学学报(自然科学)，2019，33(2)：198-207.
TAN Z. Study on consumers′ willingness-to-pay for internet insurance based on K-means and SEM：taking TPB and TAM as the analysis framework[J]. Journal of Chongqing University of Technology (Natural Science)，2019，33(2)：198-207.

[6] 许苗村，蒋先刚. 基于均值聚类的银行客户信用关系分析[J]. 华东交通大学学报，2008，25(6)：55-58.
XU M C，JIANG X G. An analysis on credit relationship of bank customer based on K-means cluster[J]. Journal of East China Jiaotong University，2008，25(6)：55-58.

[7] WANG H，ZHOU C D，LI L X. Design and application of a text clustering algorithm based on parallelized K-means clustering[J]. Revued Intelligence Artificielle，2019，33(6)：453-460.

[8] ABO-ELNAGA Y，NASR S. K-means cluster interactive Algorithm-based evolutionary approach for solving bilevel multi-objective programming problems[J]. Alexandria Engineering Journal，2021，61(1)：811-827.

[9] YANG M S，SINAGA K P. A feature-reduction multi-view K-means clustering algorithm[J]. IEEE Access，2019，7：114472-114486.

[10] RONG H，RAMIREZ-SERRANO A，GUAN L，et al. Image object extraction based on semantic detection and improved K-means algorithm[J]. IEEE Access，2020，8：171129-171139.

[11] 朱凡，王印琪. 基于 K-means 与神经网络机器学习算法的用户信息聚类及预测研究[J]. 情报科学，2021，39(7)：83-90.
ZHU F，WANG Y Q. Clustering and prediction of user information based on K-means and neural network machine learning algorithm[J]. Information Science，2021，39(7)：83-90.

[12] PANG N T，MICHAEL S，VIPIN K. Introduction to data mining[M]. Beijing：Beijing the People′s Posts and Telecommunications Press，2011.

[13] MAC Q J. Some methods for classification and analysis of multivariate observations[M]. Berkeley：Berkeley Symposium on Mathematical Statistics and Probability，1967.

[14] 姬强，孙艳丰，胡永利，等. 深度聚类算法研究综述[J]. 北京工业大学学报，2021，47(8)：912-924.
JI Q，SUN Y F，HU Y L，et al. Review of clustering with deep learning[J]. Journal of Beijing University of Technology，2021，47(8)：912-924.

[15] 蔡宇浩，梁永全，樊建聪，等. 加权局部方差优化初始簇中心的 K-means 算法[J].计算机科学与探索，2016，10(5)：732-741.
CAI Y H，LIANG Y Q，FAN J C，et al. Optimizing initial cluster centroids by weighted local variance in K-means algorithm[J]. Journal of Frontiers of Computer Science and Technology，2016，10(5)：732-741.

[16] 薛印玺，许鸿文，李羚. 基于样本密度的全局优化 K 均值聚类算法[J]. 计算机工程与应用，2018，54(14)：143-147.
XUE Y X，XU H W，LI L. Global optimized K-means clustering algorithm based on sample density[J]. Computer Engineering and Applications，2018，54(14)：143-147.

[17] 孟子健，马江洪. 一种可选初始聚类中心的 K 均值算法[J]. 统计与决策，2014(12)：12-14.
MENG Z J，MA J H. A K-means algorithm with optional initial clustering centers[J]. Statistics and Decision，2014(12)：12-14.

[18] 董秋仙，朱赞生. 一种新的选取初始聚类中心的 K-means 算法[J]. 统计与决策，2020，36(16)：32-35.
DONG Q X，ZHU Z S. A new K-means algorithm for selecting initial clustering center[J]. Statistics and Decision，2020，36(16)：32-35.

[19] 廖纪勇，吴晟，刘爱莲. 基于相异性度量选取初始聚类中心改进的 K-means 聚类算法[J]. 控制与决策，2020，554：1-8.
LIAO J Y，WU S，LIU A L. Improved K-means clustering algorithm for selecting initial clustering centers based on dissimilarity measure[J]. Control and Decision，2020，554：1-8.

[20] REZAEE M R，LELIEVELDT B P，REIBER J H. A new cluster validity index for the fuzzy C-means[J]. Pattern Recognition Letters，1998，19(3/4)：237-246.

[21] 贾瑞玉，宋建林. 基于聚类中心优化的 K-means 最佳聚类数确定方法[J]. 微电子学与计算机，2016，33(5)：62-66.
JIA R Y，SONG J L. K-means optimal clustering number determination method based on clustering center optimization[J]. Microelectronics & Computer，2013，33(5)：62-66.

[22] 贾瑞玉,李玉功. 类簇数目和初始中心点自确定的K-means算法[J]. 计算机工程与应用,2018,54(7):152-158.
JIA R Y,LI Y G. K-means algorithm of clustering number and centers self-determination[J]. Computer Engineering and Applications,2018,54(7):152-158.

[23] 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):27-33.
WANG J R,MA X,DUAN G L. Improved K-means clustering K-value selection algorithm[J]. Computer Engineering and Applications,2019,55(8):27-33.

[24] 王子龙,李进,宋亚飞. 基于距离和权重改进的K-means算法[J]. 计算机工程与应用,2020,56(23):87-94.
WANG Z L,LI J,SONG Y F. Improved K-means algorithm based on distance and weight[J]. Computer Engineering and Applications,2020,56(23):87-94.

[25] 唐东凯,王红梅,胡明,等. 优化初始聚类中心的改进K-means算法[J]. 小型微型计算机系统,2018,39(8):1819-1823.
TANG D K,WANG H M,HU M,et al. Optimizing initial cluster center of improved K-means algorithm[J]. Journal of Chinese Computer Systems,2018,39(8):1819-1823.

[26] KATSAVOUNIDIS I,JAY K C,ZHANG Z. A new initialization technique for generalized lloyd iteration[J]. Signal Processing Letters,1994,1(10):144-146.

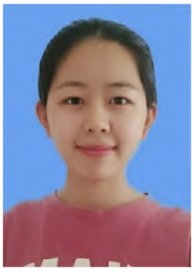
[27] 杨红, 李丹宁, 王雅洁. 基于离群点检测 (LOF) 的K-means算法[J]. 通信技术,2019,52(8):1884-1888.
YANG H,LI D N,WANG Y J. K-means algorithm based on LOF[J]. Communications Technology,2019,52(8):1884-1888.

[28] 刘凤,戴家佳,胡杨. 基于局部密度离群点检测K-means算法[J]. 重庆工商大学学报(自然科学版),2021,38(4):30-35.
LIU F,DAI J J,HU Y. The K-means algorithm based on local density outlier detection[J]. Journal of Chongqing Technology and Business University(Natural Science Edition),2021,38(4):30-35.

[29] 冷泳林,张清辰,赵亮,等. 基于离群点检测的K-means算法[J]. 渤海大学学报(自然科学版),2014,35(1):34-38.
LENG Y L,ZHANG Q C,ZHAO L,et al. K-means algorithm based on outliers detection[J]. Journal of Bohai University(Natural Science Edition),2014,35(1):34-35.

[30] MILOS R,ALEXANDROS N,MIRJANA I. Reverse nearest neighbors in unsupervised distance-based outlier detection[J]. IEEE Transactions on Knowledge and Data Engineering,2015,27(5):1369-1382.

第一作者:王森(1969—),男,教授,硕士研究生导师,研究方向为计算机应用开发。E-mail: 515613251@qq.com。



通信作者:刘琛(1997—),女,硕士研究生,研究方向为数据挖掘。E-mail:852965628@qq.com。

(责任编辑:刘棉玲)