K-Means聚类算法研究综述

杨俊闯,赵 超
河北工程大学 信息与电气工程学院,河北 邯郸 056038

摘 要:*K*-均值(*K*-Means)算法是聚类分析中一种基于划分的算法,同时也是无监督学习算法。其具有思想简单、效果好和容易实现的优点,广泛应用于机器学习等领域。但是*K*-Means算法也有一定的局限性,比如:算法中聚类数目*K*值难以确定,初始聚类中心如何选取,离群点的检测与去除,距离和相似性度量等。从多个方面对*K*-Means算法的改进措施进行概括,并和传统*K*-Means算法进行比较,分析了改进算法的优缺点,指出了其中存在的问题。对*K*-Means算法的发展方向和趋势进行了展望。

关键词:*K*-Means;聚类算法;聚类中心;离群点
文献标志码:A **中图分类号:**TP301 doi:10.3778/j.issn.1002-8331.1908-0347

杨俊闯,赵超.*K*-Means聚类算法研究综述.计算机工程与应用,2019,55(23):7-14.
YANG Junchuang, ZHAO Chao. Survey on *K*-Means clustering algorithm. Computer Engineering and Applications, 2019, 55(23):7-14.

Survey on *K*-Means Clustering Algorithm

YANG Junchuang, ZHAO Chao
College of Information and Electrical Engineering, Hebei University of Engineering, Handan, Hebei 056038, China

Abstract: The *K*-Means algorithm is a partition-based algorithm in cluster analysis. With an unsupervised learning algorithm, its advantages of simple thinking, good effect and easy implementation are widely used in fields such as machine learning. But the *K*-Means algorithm also has certain limitations. For example, the *K* number of clusters in the algorithm is difficult to determine how to choose the initial cluster center, how to detect and remove outliers and the distance and similarity measure. This paper summarizes the improvement of *K*-Means algorithm from several aspects, and compares it with the classical *K*-Means algorithm. In addition, it analyzes the advantages and disadvantages of the improved algorithm, and points out the problems. Finally, the development direction and trend of *K*-Means algorithm are prospected.

Key words: *K*-Means; clustering algorithm; cluster center; outliers

1 引言

近几年时间,大数据时代的到来促使机器学习技术飞速发展。聚类分析作为传统机器学习算法中常用方法之一,由于其实用、简单和高效的特性而广受青睐,它已成功应用于许多领域,如:文档聚类[1-2]、市场细分[3-4]、图像分割[5-7]、特征学习[8-9]等。聚类也是数据挖掘中一个重要的概念[10],其核心是寻找数据对象中隐藏的有价值的信息。

典型的聚类算法分为三个阶段,如图1所示[11]:特征选择和特征提取,数据对象间相似度计算,根据相似度将数据对象分组。聚类目的是将数据对象分成多个类或簇,同一簇中的对象具有较高的相似度,而不同簇中的对象差别较大[12]。聚类算法可以分为两大类:层次聚类算法和划分聚类算法[13]。层次聚类算法通过不同类别间的数据对象的相似度试图构建一个高层次的嵌套聚类树结构,聚类树的构建有两种类型:凝聚型层次聚类(自底向上的方式构建树结构)和分裂型层次聚类(自顶向下的方式构建树结构)。划分聚类算法需要预先指定聚类数目和聚类中心,通过优化一些损失函数,将数据集分成若干互不相交的簇。

基金项目:河北省高等学校科学技术研究项目(No.QN2018109)。
作者简介:杨俊闯(1995—),男,硕士,研究领域为机器学习;赵超(1986—),通讯作者,男,博士,讲师,研究领域为机器学习与量子计算,E-mail:zc57@163.com。
收稿日期:2019-08-22 **修回日期:**2019-10-12 **文章编号:**1002-8331(2019)23-0007-08
CNKI **网络出版:**2019-10-15, http://kns.cnki.net/kcms/detail/11.2127.TP.20191015.1136.006.html

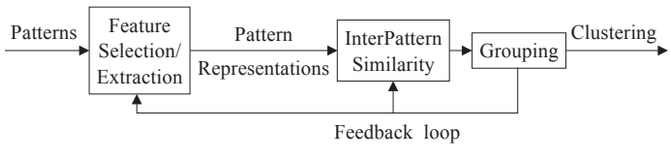


图1　聚类的三个阶段

K-Means算法作为聚类算法中最流行的算法，在1967年被MacQueen[14]首次使用，相较于其他的聚类算法，*K*-Means算法以效果较好、思想简单的优点在聚类算法中得到了广泛的应用。但是，*K*-Means算法也有其自身的局限性，比如算法中聚簇个数 k 需要事先确定，初始聚类中心由随机选取产生，离群点对聚类结果的影响等。针对上述的缺点，各个领域的学者提出了不同的改进算法。

文中首先介绍了传统*K*-Means算法的原理和流程，然后针对*K*-Means算法的上述缺点，从多个改进方向分别详细讲述了不同的改进算法。最后，根据当前研究的热点问题，提出对*K*-Means算法未来研究方向的展望。

2　传统*K*-Means算法

2.1　*K*-Means算法原理

K-Means算法是一种无监督学习，同时也是基于划分的聚类算法[15]，一般用欧式距离作为衡量数据对象间相似度的指标，相似度与数据对象间的距离成反比，相似度越大，距离越小。算法需要预先指定初始聚类数目 k 以及 k 个初始聚类中心，根据数据对象与聚类中心之间的相似度，不断更新聚类中心的位置，不断降低类簇的误差平方和(Sum of Squared Error，SSE)，当SSE不再变化或目标函数收敛时，聚类结束，得到最终结果。

K-Means算法的核心思想是：首先从数据集中随机选取 k 个初始聚类中心 $C_i(1\leqslant i\leqslant k)$，计算其余数据对象与聚类中心 C_i 的欧氏距离，找出离目标数据对象最近的聚类中心 C_i，并将数据对象分配到聚类中心 C_i 所对应的簇中。然后计算每个簇中数据对象的平均值作为新的聚类中心，进行下一次迭代，直到聚类中心不再变化或达到最大的迭代次数停止。

空间中数据对象与聚类中心间的欧式距离计算公式为：

$$d(x,C_i)=\sqrt{\sum_{j=1}^{m}(x_j-C_{ij})^2} \tag{1}$$

其中，x 为数据对象，C_i 为第 i 个聚类中心，m 为数据对象的维度，x_j，C_{ij} 为 x 和 C_i 的第 j 个属性值。

整个数据集的误差平方和 SSE 计算公式为：

$$SSE=\sum_{i=1}^{k}\sum_{x\in C_i}\left|d(x,C_i)\right|^2 \tag{2}$$

其中，SSE 的大小表示聚类结果的好坏，k 为簇的个数。

2.2　*K*-Means算法流程

K-Means聚类算法是一个不断迭代的过程[16]，如图2所示，原始数据集有4个簇，图中 x 和 y 分别代表数据点的横纵坐标值，使用*K*-Means算法对数据集进行聚类，在对数据集经过两次迭代后得到最终的聚类结果。

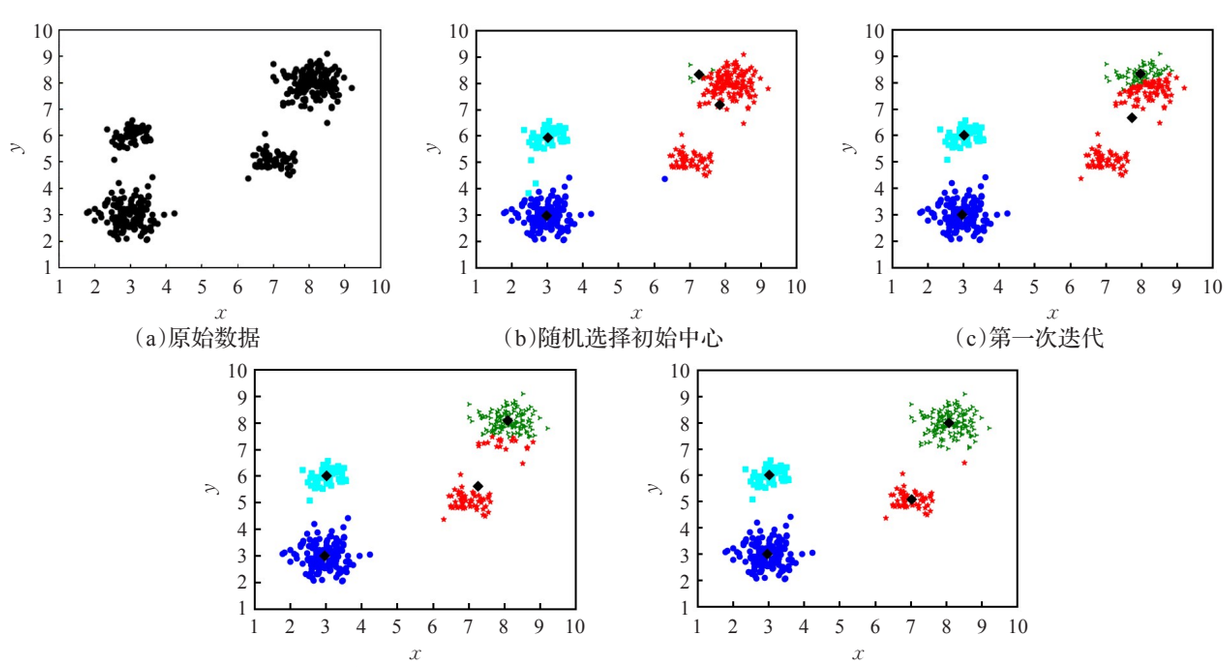


图2　*K*-Means算法迭代过程

K-Means算法的流程图如图3所示。

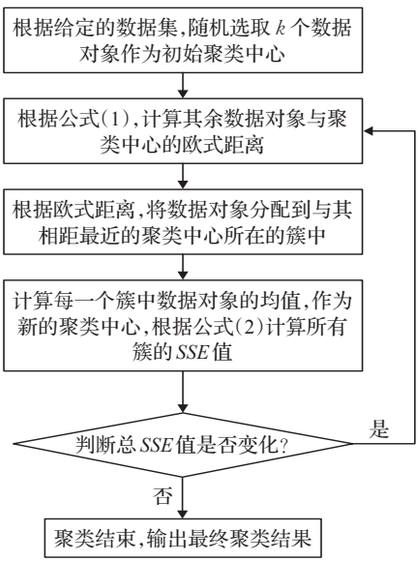


图3　*K*-Means算法流程图

K-Means聚类算法对于大数据集有高效的聚类效果，其算法复杂度为 $O(nmkT)$。其中，n 为数据集大小，m 为数据对象特征维数，k 为指定的簇的数目，T 为总的迭代次数。

3　改进*K*-Means聚类算法

在数据挖掘中，聚类是最常用的方法之一。*K*-Means算法由于其本身的聚类效果好、思想简单、聚类速度快的优点得到了广泛的应用。但是，*K*-Means算法有两个众所周知的缺点，即对初始值的敏感和易陷入局部最优解[17]。因此，许多学者不断地提出改进算法克服这些缺点。目前，对*K*-Means算法的改进方法主要集中于以下几个方向：算法中初始 k 值的选取、初始聚类中心点的选取、离群点的检测和去除、距离和相似性度量等。

3.1　初始 k 值的选取

在传统*K*-Means算法中，聚簇个数 k 要求事先确定，但在实际中，往往因为数据量过大和缺乏经验导致 k 值难以确定，若 k 值选取得过小，则会导致同一簇内数据对象差异很大，若 k 值选择过大，则会导致不同簇间差异很小。同时，k 值选取不当也会使最终的聚类结果陷入局部最优，这往往是使用传统*K*-Means算法最为诟病的地方。早在1998年，Rezaee等[18]提出了最佳 k 值是在 $(1,\sqrt{n})$ 范围内，n 为数据集大小，这也为后来对传统*K*-Means算法的改进提供了方向。

基于簇数 k 与误差平方和 SSE 的关系，文献[19]根据不同 k 值时 SSE 的变化趋势，如图4所示，选择肘点对应的 k 值作为最优簇数。文献[20]针对簇数 k 与 SSE 的关系图像“肘点”模糊不明确的问题，结合指数函数、权重项、偏置项等参数来确定最佳 k 值。由于最优簇数 k 需要通过人工分析决策图来确定[21]，文献[22]针对此问题结合统计学中的方法，利用线性回归来拟合决策图中的点，根据观测值与实际值的差来确定最优 k 值和初始聚类中心。

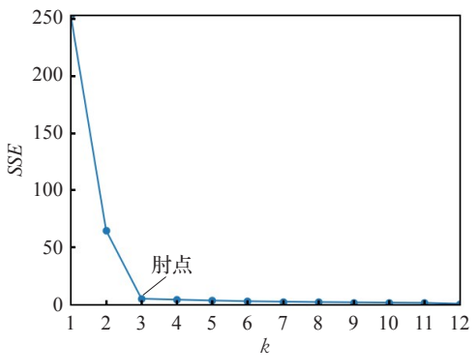


图4　*SSE*与 k 值的关系

层次 k 均值聚类算法[23]基于层次聚类的思想，与空间中的层次结构相结合对传统*K*-Means算法进行改进。该算法能够自适应地得到最佳或者接近最佳的聚簇个数 k，避免人工地选择 k 值。在算法开始时与传统*K*-Means算法一样，随机选取 k 个初始聚类中心进行一次迭代，然后计算聚类后的聚类测度值：

$$J=\sqrt{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-c_i\right)^2}{n-1}} \tag{3}$$

进行更细层次的聚类。其中，c_i 为第 i 个类的聚类中心，x_{ij} 为第 i 个类中第 j 个数据对象，n_i 为第 i 个类中数据对象的数目，k 为簇的个数，n 为数据集大小。聚类中心在每一次迭代后找出所有簇中半径最大的簇，在该簇中选择相距最远的两个样本点作为新的聚类中心，并和其他的聚类中心重新进行迭代。每次迭代结束后计算聚类测度值之比：

$$\varepsilon=\frac{J^{(t)}-J^{(t-1)}}{J^{(t-1)}} \tag{4}$$

其中，t 为迭代次数。如果 ε 大于人为设定的经验值 Δ，则继续进行聚类，否则算法停止，输出聚类结果和最佳簇数 k。结果表明，相较于*K*-Means算法聚类精度有很大提升，该算法也能得到接近实际簇数的 k 值。

Wang等[24]基于图像分割的思想，利用分水岭算法将原始数据集划分为多个区域以确定最佳簇数 k。该改进算法首先对数据集进行预处理，利用公式：

$$p_i=\sum_{j=1}^{n}\left(-\gamma\left\|x_i-x_j\right\|\right) \tag{5}$$

计算各个数据对象的密度，其中，p_i 表示数据对象 x_i 的密集程度，γ 为经验值需要人为设定。将数据对象的空间密度绘制成灰度图像并使用分水岭算法进行分区操作，假设灰度图像的二维图像如图5中曲线 $L1$ 所示，虚

线 $L2$ 为水位线，将数据集中数据对象的密度从小到大排序作为纵坐标，横坐标为密度对应的数据点，水位线从 x_0 到 x_2 的过程中生成 A,B,C 三个区域，选择每个区域的中心为*K*-Means算法的初始聚类中心，最佳簇数 k 为区域的数量，然后进行*K*-Means迭代得到最终聚类结果。

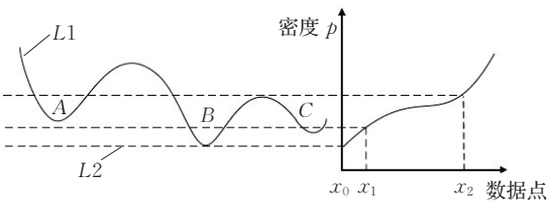


图5　分水岭算法示意图

3.2　初始聚类中心点选取

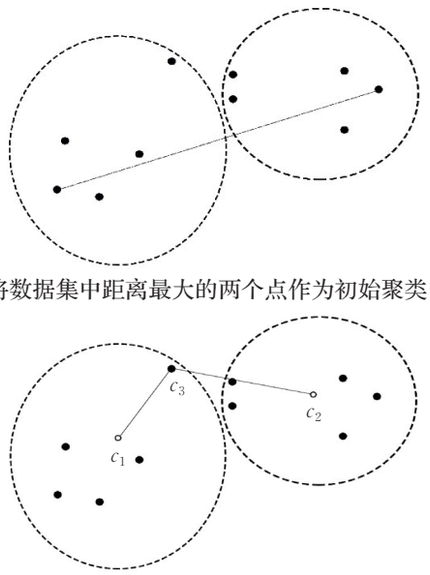
初始聚类中心的选取对*K*-Means算法聚类结果的影响起着决定性的作用。*K*-Means算法对聚类中心的初始位置十分敏感，对于每次迭代，初始聚类中心选取的不同往往会导致不同的聚类结果[25-26]。

Xiong等[27]首先计算所有数据对象的密度，求出数据集的密度平均值，将大于密度平均值的数据对象作为高密度点集，从密度点集中选取密度值最大的数据对象作为第一个初始聚类中心，剩余聚类中心选取依据与前面选定的聚类中心的距离最大的原则进行。同样是基于密度的改进方法，为了避免将一个簇中的两个高密度点同时作为初始聚类中心，Du等[28]基于聚类中心点与其他中心点之间的距离相对较大的基本思想，结合与高密度点的相对距离进行中心点选取。Tanir等[29]首先选取数据集中距离最大的两个点作为初始聚类中心，如图6(a)所示，将剩余数据对象依据到聚类中心点距离的远近分配到相应的簇中，并更新聚类中心。继续寻找与聚类中心距离最远的点作为下一个中心点，在图6(b)中，点 c_3 与聚类中心 c_1 和 c_2 的距离最远，所以选择 c_3 作为第三个聚类中心，直到聚类中心点个数到 k 个为止。为了避免聚类结果陷入局部最优解，文献[30]采用聚类中心搜索算法对数据集进行多次随机划分取样，如图7所示，右图对左图进行了多次取样，并且取样后的右图依然保持了原始数据集的特征，在取样后的数据集上选取中心点进行*K*-Means聚类，降低了算法运行的时间复杂度。

文献[31]基于数据对象的个体轮廓系数选取优秀样本，并从优秀样本中自适应地选择初始聚类中心进行*K*-Means聚类。个体轮廓系数的公式为：

$$S_i=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{6}$$

$$a(i)=\frac{1}{n_c-1}\sum_{i,j\in C_c,i\neq j}d(i,j) \tag{7}$$



(a)将数据集中距离最大的两个点作为初始聚类中心

(b)与聚类中心距离最远的点作为下一个聚类中心点

图6　最大距离法选择初始聚类中心

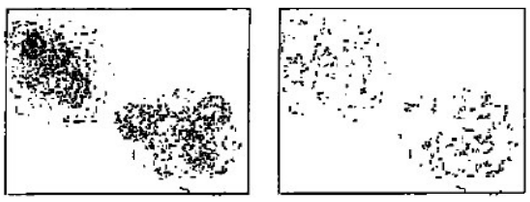


图7　原始数据集与取样后数据集的分布

$$b(i)=\min_{p,p\neq c}\left\{\frac{1}{n_p}\sum_{i\in C_c,j\in C_p}d(i,j)\right\} \tag{8}$$

其中，n_c 表示 c 类中数据对象的数目，$d(i,j)$ 为数据对象 i，j 之间的欧式距离，$a(i)$ 表示和 i 同类中所有数据对象之间的平均距离，$b(i)$ 表示和 i 不同类中所有数据对象之间的最小平均距离。S_i 在 $[-1,1]$ 内取值，其值越大，说明该数据对象聚到某个类越合理。该算法首先将传统*K*-Means算法迭代 H 次，记录每次迭代的距离矩阵 $dist(j)$ 并计算 k 个聚类中心的个体轮廓系数相加得到 $Sil(j)$，对个体轮廓系数进行降序排列，取其对应的距离矩阵 $dist(j)$ 的前 $Q\%$ 的样本标记为候选优秀样本。对候选优秀样本进行计数，删除数量小于 k 的样本，将其余样本作为优秀样本调用传统*K*-Means算法进行聚类得到聚类结果。该算法的缺点在于参数 H，Q，M 都是根据经验人为设定的，会影响最终的聚类精度。

将图论中的最小生成树算法和传统*K*-Means算法相结合，基于最小生成树的层次*K*-Means算法[32]，首先使用prim算法将数据对象生成一棵最小生成树，然后根据最大距离分裂原则将最小生成树划分成 m 个子簇，从 m 个子簇中找到数据对象最多的 $k(k\leqslant m)$ 个簇，计算每个簇的均值作为初始聚类中心进行传统*K*-Means迭代处理。根据评价函数：

$$I=\sqrt{\frac{\sum_{j=1}^{k}\sum_{i=1}^{m}dist(p_i,c_j)}{S}} \tag{9}$$

判断是否进行两个簇之间的合并，其中，p_i 是聚类中心 c_j 中包含的数据对象，S 为聚到该簇的数目，$dist(p_i,c_j)$ 为数据对象 p_i 和 c_j 的欧式距离。利用公式：

$$\Delta=\frac{I^{(t)}-I^{(t-1)}}{C^{(t)}-C^{(t-1)}} \tag{10}$$

对评价函数进行评价分析，其中，C 为簇数目，t 为迭代次数。如果 $\Delta>\omega$（$\omega=-0.19$，为经验值）则继续进行*K*-Means迭代，否则聚类结束。该算法的缺点在于最后判断迭代结束条件时，采用与人为设定的经验值相比较，难免会影响聚类结果和算法的准确性。

Gu[33]使用减法聚类确定初始聚类中心，并与Locality Sensitive *K*-Means（LSKM）[34]算法相结合改进传统*K*-Means算法。减法聚类是一种密度聚类算法，首先将样本集中的所有数据对象都当成潜在的聚类中心，根据：

$$H(X_i)=\sum_{j=1}^{n}e^{-\alpha\times d(x_j,x_i)} \tag{11}$$

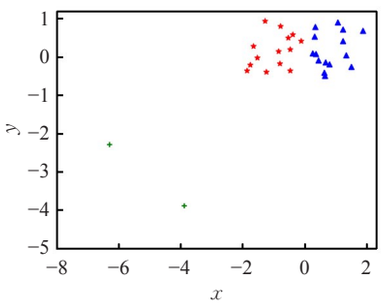
计算每个数据对象的密度指标，其中，$d(x_i,x_j)$ 为数据对象 x_i 和 x_j 之间的欧式距离。找到密度指标最大的数据对象作为第一个初始聚类中心，之后除去已完成的聚类中心的作用，再根据公式：

$$H_p(x_i)=H_{p-1}(x_i)-H_{p-1}^{*}\cdot e^{-\beta\cdot d(x_{p-1}^{*},x_i)} \tag{12}$$

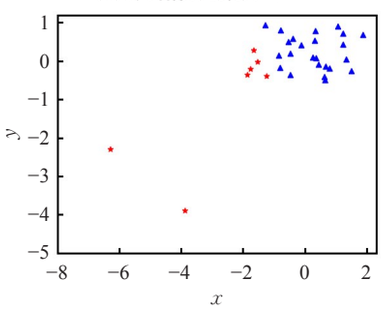
寻找新的聚类中心，其中，H_{p-1}^{*} 是前一次迭代中最大的密度指标，H_{p-1}^{*} 为前一次迭代找到的聚类中心。直到初始聚类中心的个数达到设定的 k 值停止查找，然后将找到的初始聚类中心作为输入运行LSKM算法进行迭代聚类。该算法的缺点是公式中的 α，β 和 k 三个参数的值都需要人工输入，并且离群点对聚类效果的影响较大。

3.3　离群点的检测

离群点（也称为噪声或异常值）是少数异于正常数据集的数据对象，在聚类分析中，离群点通常被认为是远离于聚类中心的点。由于当今各行业的数据集有数量大和多样性丰富的特点，所以对一些学习算法的鲁棒性要求较高。然而很多的传统学习算法并不具备这一特点，它们假设数据集中没有离群点，*K*-Means算法就是如此，算法运行前没有对数据进行预处理。*K*-Means算法对离群点较为敏感[35]，如图8所示，有两个数据集和两个离群点，数据集中的两个族分别用星形和三角形表示，离群点用加号表示，在对数据集运行*K*-Means算法后，由于两个离群点的影响，得到如图8(b)所示情况，并不能得到图8(a)中正确的簇的分布情况。如果将数据集中的离群点作为初始聚类中心，则会使聚类结果陷入局部最优值，影响聚类效果。



(a)原始数据集分布



(b)运行*K*-Means算法后的聚类结果

图8　离群点对*K*-Means算法的影响

文献[36]使用基于密度的离群点的检测算法——LOF算法来筛除离群点，并结合最大最小距离法在筛选后的样本点上选取初始中心。算法首先根据LOF算法[37]，计算数据集中的每一个数据对象的离群因子，离群因子越大，说明该数据对象偏离中心的程度越大，越有可能是离群点。对离群因子进行升序排列，选出对应的前 $\alpha\times n$（$0<\alpha\leqslant 1$，α 的值需要人为设定）个数据对象作为候选聚类中心样本。在候选聚类中心样本上使用最大最小法选出 k 个初始聚类中心，进行*K*-Means迭代输出聚类结果。最大最小法的初始中心点选择公式为：

$$C_i=\max\{Dis_j;j=1,2,\cdots,n\} \tag{13}$$

$$Dis_j=\min_{C_i\in C}\{d(C_i,x_j)\} \tag{14}$$

其中，Dis_j 为数据对象 j 到聚类中心的最小距离，n 为数据集大小，C_i 表示聚类中心点，$d(C_i,x_j)$ 为数据对象 x_j 到聚类中心 C_i 的欧式距离。该算法的缺点在于找到最优的 α 值需要进行大量的实验，增加了算法的时间复杂度。

Fan等[38]基于数据对象的网格密度对传统*K*-Means算法提出改进。假设数据集共有 m 维，计算每一个维度 J 的最大值 $J_{\max}$ 和最小值 $J_{\min}$，每一个维度中网格的长度为：

$$GL_J=\frac{J_{\max}-J_{\min}}{k+1},1\leqslant J\leqslant m \tag{15}$$

其中，k 为聚类中心个数，需要人工输入。将每一维的 GL_J 当作半径，目标数据对象的网格密度为以自身为圆心，GL_J 为半径的圆的范围内的数据对象个数。密度阈值 $Mins$ 需要事先确定，离群点的判定规则为，如果数

据对象的网格密度小于 $Mins$ 则被认定为离群点,应从数据集中剔除。离群点去除后,对每一个维度 J 中的值依次进行排序,将每次排序后的数据划分成 k 段,得到每一段的平均值 $\{p_{J1},p_{J2},\cdots,p_{Jk}\}$。根据平均值得到 k 个初始聚类中心 $\{C_1,C_2,\cdots,C_k\}$,其中 $C_k=\{p_{1k},p_{1k},\cdots,p_{1k}\}$,将每一个数据对象根据聚类中心的距离分配到响应的簇中,重新计算每一个簇的聚类中心,不断进行迭代,直到聚类中心不再变化时算法结束。与传统 K-Means 算法相比较,该改进算法将数据从各个维度中抽离出来,取每个维度的平均值作为初始聚类中心,可以降低随机选取初始聚类中心对聚类结果的影响。

3.4　距离和相似性度量

聚类算法大多数都是通过距离度量计算数据对象的相似性,从而对相关数据对象进行分组[39]。由于每种算法采用的距离度量方式不同,所以,用不同算法计算两个数据对象间的相似度也会有所不同,因此选择合适的距离度量方法在任何聚类算法中都起着至关重要的作用。传统 K-Means 算法中采用欧式距离度量数据对象间的相似性,欧式距离默认数据对象的所有属性值同等重要,但在现实生活中,往往不是这样。这种对属性重要性不加区分的处理方法很可能导致数据对象在欧氏空间中产生距离失真:如果空间中的两点在重要属性上距离很近,但由于其他无关属性对距离的放大作用,这两点在欧氏空间中很可能被度量为最远[40]。

空间相似性度量 K-Means 算法[41]针对传统 K-Means 算法对非簇型数据分类效果不佳的问题,采用空间密度相似性距离弥补欧氏距离不能准确表达流型数据对象间的相似性的缺点,并结合新的聚类中心迭代模型求出聚类结果。文献[42]将 K-Means 算法应用于文本聚类,并分别采用欧式距离、平方欧式距离、曼哈顿距离、余弦距离、谷本距离等方法一一对文本数据对象间的相似性进行测量和比较。根据不同方法的聚类结果进行分析得出结论,欧式距离在测量文本相似度方面还有一定的局限性,应该根据不同的数据集选取不同的相似性测量方法。Xu 等[43]用信息增益(Information Gain)和特征选择算法分别求出数据集中各个特征的权重值,取其平均值作为最终的特征权重值与欧式距离结合作为加权距离进行 K-Means 聚类,也得到了较为不错的结果。

在粗糙集理论和假设检验的基础上,基于加权距离的自适应粗糙 K-Means 聚类算法[44]利用粗糙集中属性约简理论对数据集进行预处理,筛选出属性值比较重要的数据集合,对该集合求每个属性的权重值:

$$W_j=\frac{\sum_{i=1}^{N}\left(x_{ij}-\overline{x_j}\right)^2}{\sum_{q=1}^{m}\sum_{i=1}^{N}\left(x_{ij}-\overline{x_q}\right)^2} \tag{16}$$

其中,x_{ij} 为数据对象 x_i 的第 j 个属性值,$\overline{x_j}$ 为第 j 个属性值的平均值,N 为数据对象个数,m 为数据集中属性的个数。对于 m 维的数据集,加权后距离计算公式为:

$$dist(x_k,x_i)=\sqrt{\sum_{j=1}^{m}\left(x_{kj}-x_{ij}\right)^2} \tag{17}$$

然后,根据粗糙集理论结合改进后的加权欧式距离选取聚类中心点。对每个簇进行统计学检验来判断中心点所在簇的数据对象是否符合高斯分布模型,如果符合,算法结束,输出最优 k 值,否则簇的中心点分裂,继续更新每个数据对象属性的权重和聚类中心点。

针对传统 K-Means 算法中使用欧式距离的相似性度量不能有效地反应非凸数据集的分布情况,如图 9 所示,点 a 和点 b 在同一个簇中,点 a 和点 b 之间的相似性应该高于点 a 和点 c 的相似性。然而,使用欧式距离计算相似性,则会使 a 点更接近 c 点。针对这个问题,Xue 等[45]提出使用空间密度相似距离来替代欧式距离。为了计算相似性,首先计算空间密度相似线长度:

$$L(i,j)=\mathrm{e}^{Dist(i,j)\times A}-1 \tag{18}$$

$$A=\frac{Dist(i,j)}{P} \tag{19}$$

其中,$Dist(i,j)$ 为数据对象 i 和 j 之间的欧式距离,A 为距离调整因子,P 为簇内数据的平均欧式距离。令 $D_G(i,j)$ 表示数据对象 i 和 j 之间的最短距离,距离越小数据对象之间相似性越大。首先对 $D_G(i,j)$ 初始化,如果数据对象 j 是 i 第 k 个邻近点之一,则数据对象 i,j 的最短距离为 $D_G(i,j)=L(i,j)$,否则为无穷大。因此,数据集中任意两个数据对象之间最短距离的计算公式为:

$$D_G(i,j)=\min(D_G(i,j),D_G(i,s)+D_G(s,j)) \tag{20}$$

其中,s 表示数据集中除 i 和 j 外的数据对象,使用 $D_G(i,j)$ 距离替代欧式距离进行 K-Means 算法迭代,直到聚类中心不再变化算法停止。实验证明,该算法对非线性数据集有良好的聚类能力。但该算法的时间复杂度为 $O(n^3kt)$,其中 n 为数据集大小,t 为总的迭代次数,相较于 K-Means 算法时间复杂度更高。

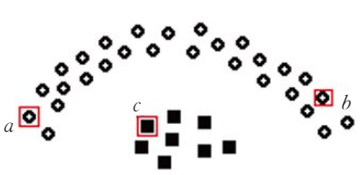


图 9　非凸数据集分布

3.5　K-Means 算法的其他改进

近年来,仿生智能优化算法由于其结构简单,易于实现的特点,多用于求复杂问题的最优解,比较流行的算法有萤火虫算法[46]、森林优化算法[47]、遗传算法[48]等。将这些算法的优点与传统 K-Means 算法相结合,可以提高 K-Means 算法的全局搜索能力,使聚类中心不容易陷

入局部最优解。通常使用智能优化算法寻找*K*-Means算法的初始聚类中心，降低选取初始中心点的随机性。

基于萤火虫优化的加权*K*-Means算法[49]，利用萤火虫优化算法的全局搜索能力强、易收敛的特点，选取*K*-Means算法的初始聚类中心。由于数据属性对聚类结果的影响程度不同，在传统欧式距离的基础上引入权重值，加大了数据的不同属性间的区分程度，消除了数据集中噪声点的影响。该算法很好地克服了传统*K*-Means算法中初始聚类中心难选取和噪声点对聚类结果的影响，提升了聚类的性能。文献[50]提出基于改进森林优化算法的*K*-Means算法，引入衰减因子作为自适应步长加快算法聚类速度，结合算术交叉操作，改进传统森林优化算法易陷入局部最优解、收敛慢的缺点，提高聚类精度和聚类准确率。

Shi等[51]将遗传算法与*K*-Means算法相结合，提高*K*-Means算法的聚类效率与精确度。该算法首先使用近邻排序算法(Sorted-Neighborhood Method，SNM)对原始数据集中的重复数据进行清理，将去重后的数据进行归一化，计算数据集中各个数据对象之间的欧式距离，然后使用公式：

$$AvgDis=\frac{\sum_{i=1,j=1}^{n} Dis\left(s_i,s_j\right)}{A_n^2} \tag{21}$$

求数据集的平均欧式距离，其中，$Dis(s_i,s_j)$为数据对象s_i和s_j之间的欧式距离，A_n为数据对象的数量。数据集中的每个数据对象如果与目标点的距离在$AvgDis$之内，那么认为该数据对象为目标点的邻近点，并统计其邻近点的数量。将数据集中各个数据对象的邻近点的数量按降序排列，取其前k个数据对象作为初始聚类中心进行*K*-Means聚类。然后利用遗传算法对*K*-Means聚类后的结果进行清理，初始种群是由50个01字符生成的基因序列，选择每个基因对应的特征作为*K*-Means聚类算法的结果。适应度函数公式为：

$$f_i=\frac{N-\sum_{i=1}^{k}a_i^k}{N},1\leqslant i\leqslant l \tag{22}$$

其中，f_i为基因i的适应度，N为数据集中数据对象的数目，a_i^k为基因i在聚类结果被分错的数目，l为种群中个体的数目，k为簇的数目。为了计算更加简便，需要将适应度进行归一化：

$$\widehat{f_i}=\frac{f_i-f^{\max}}{f^{\max}-f^{\min}},1\leqslant i\leqslant l \tag{23}$$

其中，$f^{\max}$和$f^{\min}$分别代表了种群中适应度的最大值与最小值。根据个体的适应度的大小选择轮盘对赌区域进行交叉操作和突变操作，消除数据集中无用的属性特征，如果达到最大迭代次数则输出新种群和最优结果，否则利用遗传算法继续进行迭代。

4　结束语

自从*K*-Means算法提出以来，不断有学者提出改进算法，丰富*K*-Means算法内容。*K*-Means算法是一个十分经典的聚类算法，其今后的应用也将会越来越普遍。对于传统*K*-Means算法中存在的一些缺点以及针对这些缺点的改进方法，文中进行了详细的阐述。

K-Means算法有着广泛的应用前景，今后也将面临更多的挑战。对*K*-Means算法的改进绝不止于本文中所阐述的方向。总结前人的经验，未来针对*K*-Means算法还可以对以下方向进行研究：(1)提升*K*-Means算法处理海量或多维数据集的能力。随着大数据时代的到来，所能获取的信息量呈指数式爆炸，如何将*K*-Means更好地用于处理指数级数据的聚类，也是需要研究的方向。(2)降低*K*-Means算法的时间复杂度。改进的*K*-Means聚类算法有着良好的聚类效果，但这是在牺牲了时间的前提下换来的，如何能更好更快地提升聚类能力，需要做更进一步优化。

参考文献：

[1] Sardar T H，Anrisa A.An analysis of MapReduce efficiency in document clustering using parallel *K*-means algorithm[J].Future Computing and Informatics Journal，2018，3(2)：200-209.

[2] Li W，Feng Y，Li D，et al.Micro-blog topic detection method based on BTM topic model and *K*-means clustering algorithm[J].Automatic Control and Computer Sciences，2016，50(4)：271-277.

[3] Tleis M，Callieris R，Roma R.Segmenting the organic food market in Lebanon：an application of *k*-means cluster analysis[J].British Food Journal，2017，119(7)：1423-1441.

[4] Hung P D，Ngoc N D，Hanh T D.*K*-means clustering using RA case study of market segmentation[C]//Proceedings of the 2019 5th International Conference on E-Business and Applications.New York：ACM，2019：100-104.

[5] Dhanachandra N，Manglem K，Chanu Y J.Image segmentation using *K*-means clustering algorithm and subtractive clustering algorithm[J].Procedia Computer Science，2015，54：764-771.

[6] Moriya T，Roth H R，Nakamura S，et al.Unsupervised pathology image segmentation using representation learning with spherical *k*-means[C]//Medical Imaging 2018：Digital Pathology，Houston，2018.

[7] Nasir A，Mashor A S，Mohamed M，et al.Enhanced *k*-means clustering algorithm for malaria slide image segmentation[J].Journal of Advanced Research in Fluid Mechanics and Thermal Sciences，2018，42：1-15.

[8] Tang J L，Wang D，Zhang Z G，et al .Weed identification based on *K*-means feature learning combined with convolutional neural network[J].Computers and Electronics in

Agriculture,2017,135:63-70.
[9] Tang J L,Zhang Z G,Wang D,et al.Research on weeds identification based on *K*-means feature learning[J].Soft Computing,2018,22(22):7649-7658.
[10] Dubey A,Choubey A.A systematic review on *k*-means clustering techniques[J].International Journal of Scientific Research Engineering & Technology,2017,6(6):624-627.
[11] Jain A K,Murty M N,Flynn P J.Data clustering:a review[J].ACM Computing Surveys,1999,3(3):264-323.
[12] 海沫,张书云,马燕林.分布式环境中聚类问题算法研究综述[J].计算机应用研究,2013,30(9):2561-2564.
[13] Chakraborty S,Das S.*k*-Means clustering with a new divergence-based distance metric:convergence and performance analysis[J].Pattern Recognition Letters,2017,100:67-73.
[14] MacQueen J.Some methods for classification and analysis of multivariate observation[C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability.Berkeley:University of California Press,1967:281-297.
[15] Saroj,Kavita.Review:study on simple *k* mean and modified *K* mean clustering technique[J].International Journal of Computer Science Engineering and Technology,2016,6(7):279-281.
[16] Anil K J,Data clustering:50 years beyond *K*-means[J].Pattern Recognition Letters,2010,31(8):651-666.
[17] Hung C H,Chiou H M,Yang W N.Candidate groups search for *K*-harmonic means data clustering[J].Applied Mathematical Modelling,2013,37(24):10123-10128.
[18] Rezaee M R,Lelieveldt B P F,Reiber J H C.A new cluster validity index for the fuzzy *c*-mean[J].Pattern Recognition Letters,1998,19(3/4):237-246.
[19] 成卫青,卢艳红.一种基于最大最小距离和SSE的自适应聚类算法[J].南京邮电大学学报(自然科学版),2015,35(2):102-107.
[20] 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.
[21] Rodriguez A,Laio A.Clustering by fast search and find of density peaks[J].Science,2014,344:1492-1496.
[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].计算机工程与用,2018,54(7):152-158.
[23] 胡伟.改进的层次*K*均值聚类算法[J].计算机工程与应用,2013,49(2):157-159.
[24] Wang X,Jiao Y,Fei S.Estimation of clusters number and initial centers of *K*-means algorithm using watershed method[C]//2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science,Guiyang,2015:505-508.
[25] Celebi M E,Kingravi H A,Vela P A.A comparative study of efficient initialization methods for the *k*-means clustering algorithm[J].Expert Systems with Applications,2013,40(1):200-210.
[26] Lei J,Jiang T,Wu K,et al.Robust *K*-means algorithm with automatically splitting and merging clusters and its applications for surveillance data[J].Multimedia Tools and Applications,2016,75(19):12043-12059.
[27] Xiong C,Hua Z,Lv K,et al.An improved *K*-means text clustering algorithm by optimizing initial cluster centers[C]//2016 7th International Conference on Cloud Computing and Big Data,Macau,2016:265-268.
[28] Du X,Xu N,Zhou C,et al.A density-based method for selection of the initial clustering centers of *K*-means algorithm[C]//2017 IEEE 2nd Advanced Information Technology,Electronic and Automation Control Conference,Chongqing,2017:2509-2512.
[29] Tanir D,Nuriyeva F.On selecting the initial cluster centers in the *K*-means algorithm[C]//2017 IEEE 11th International Conference on Application of Information and Communication Technologies,Moscow,2017:1-5.
[30] 张玉芳,毛嘉莉,熊忠阳.一种改进的*K*-means算法[J].计算机应用,2003(8):31-33.
[31] 张靖,段富.优化初始聚类中心的改进*k*-means算法[J].计算机工程与设计,2013,34(5):1691-1694.
[32] 贾瑞玉,李振.基于最小生成树的层次*K*-means聚类算法[J].微电子学与计算机,2016,33(3):86-88.
[33] Gu L.A novel locality sensitive *k*-means clustering algorithm based on subtractive clustering[C]//IEEE International Conference on Software Engineering and Service Science,Beijing,2017:836-839.
[34] Huang P F,Zhang D Q.Locality sensitive *c*-means clustering algorithm[J].Neurocomputing,2010,73(16/17/18):2935-2943.
[35] Gan G J,Ng M K P.*k*-means clustering with outlier removal[J].Pattern Recognition Letters,2017,90:8-14.
[36] 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.
[37] Breunig M M,Kriegel H P,Ng R T,et al.LOF:identifying density-based local outliers[J].ACM Sigmod Record,2000,29(2):93-104.
[38] Fan Z,Sun Y.Clustering of college students based on improved *K*-means algorithm[C]//2016 International Computer Symposium,Chiayi,2016:676-679.
[39] Visalakshi N K,Suguna J.*K*-means clustering using Max-min distance measure[C]//2009 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS 2009),Cincinnati,2009:1-6.
[40] 李四海,满自斌.自适应特征权重的*K*-means聚类算法[J].计算机技术与发展,2013,23(6):98-101.

(下转第63页)

Management，2016，36(5)：748-758.
[6] Steenbruggen J，Tranos E，Nijkamp P.Data from mobile phone operators[M].Oxford：Pergamon Press，2015，39(3)：335-346.
[7] Ratti C，Frenchman D，Pulselli R M，et al.Mobile landscapes：using location data from cell phones for urban analysis[J].Environment & Planning B Planning & Design，2006，33(5)：727-748.
[8] Becker R A，Caceres R，Hanson K，et al.A tale of one city：using cellular network data for urban planning[J].IEEE Pervasive Computing，2011，10(4)：18-26.
[9] 孔扬鑫.基于手机信令数据的人口流动分析[D].上海：华东师范大学，2017.
[10] Zhong G，Wan X，Zhang J，et al.Characterizing passenger flow for a transportation hub based on mobile phone data[J].IEEE Transactions on Intelligent Transportation Systems，2016，18(6)：1507-1518.
[11] Larijani A N，Olteanu-Raimond A M，Perret J，et al.Investigating the mobile phone data to estimate the origin destination flow and analysis；case study：Paris region ☆[J].Transportation Research Procedia，2015，6：64-78.
[12] Zheng Y.Trajectory data mining：an overview[M].New York：ACM Press，2015，6(3)：1-41.
[13] Liu Z，Yu J，Xiong W，et al.Using mobile phone data to explore spatial-temporal evolution of home-based daily mobility patterns in Shanghai[C]//International Conference on Behavioral，Economic and Socio-Cultural Computing，Krakow.Piscataway，NJ：IEEE，2017：1-6.
[14] Jiang S，Ferreira J，Gonzalez M C.Activity-based human mobility patterns inferred from mobile phone data：a case study of Singapore[J].IEEE Transactions on Big Data，2017，3(2)：208-219.
[15] Do C X，Tsukai M.Exploring potential use of mobile phone data resource to analyze inter-regional travel patterns in Japan[M]//Data mining and big data.Cham：Springer，2017：314-325.
[16] Shi L，Wang W，Cai W，et al.Mobility patterns analysis of Beijing residents based on call detail records[C]//International Conference on Wireless Communications and Signal Processing，Nanjing.Piscataway，NJ：IEEE，2017：1-6.
[17] Xu Y，Shaw S L，Zhao Z，et al.Understanding aggregate human mobility patterns using passive mobile phone location data：a home-based approach[J].Transportation，2015，42(4)：625-646.
[18] Zhao Z，Shaw S L，Xu Y，et al.Understanding the bias of call detail records in human mobility research[J].International Journal of Geographical Information Science，2016，30(9)：1-25.
[19] 胡忠顺，王进，朱亮.基于手机信令数据的大客流监控应用研究[J].电信技术，2017(4)：21-25.
[20] 项译.基于手机信令数据的旅游交通客流特征分析研究[D].南京：东南大学，2017.
[21] 钟尖.基于手机信令的综合交通枢纽客流监测技术研究[D].重庆：重庆交通大学，2017.
[22] Alexander L，Jiang S，Murga M，et al.Origin-destination trips by purpose and time of day inferred from mobile phone data[J].Transportation Research：Part C Emerging Technologies，2015，58：240-250.

（上接第14页）

[41] 薛卫，杨荣丽，赵南，等.空间密度相似性度量*K*-means算法[J].小型微型计算机统，2018，39(1)：53-57.
[42] 陈磊磊.不同距离测度的*K*-Means文本聚类研究[J].软件，2015，36(1)：56-61.
[43] Xu Y，Fu X L，Li H H，et al.A *K*-means algorithm based on feature weighting[C]//2018 2nd International Conference on Electronic Information Technology and Computer Engineering，Shanghai，2018.
[44] 孙志鹏，钱雪忠，吴秦，等.基于加权距离计算的自适应粗糙*K*-均值算法[J].计算机应用研究，2016，33(7)：1987-1990.
[45] Xue W，Yang R，Hong X，et al.A novel *k*-Means based on spatial density similarity measurement[C]//2017 29th Chinese Control and Decision Conference，Chongqing，2017：7782-7784.
[46] Yang X S.Nature-inspired metaheuristic algorithms[M].[S.l.]：Luniver Press，2010：81-89.
[47] Ghaemi M，Feizi-Derakhshi M R.Forest optimization algorithm[J].Expert Systems with Applications，2014，41(15)：6676-6687.
[48] Whitley D.A genetic algorithm tutorial[J].Statistics and Computing，1994，4(2)：65-85.
[49] 陈小雪，尉永清，任敏，等.基于萤火虫优化的加权*K*-means算法[J].计算机应用研究，2018，35(2)：466-470.
[50] 魏康园，何庆，徐钦帅.一种改进森林优化的*K*-means聚类算法[J].贵州大学学报(自然科学版)，2018，35(6)：69-75.
[51] Shi H，Xu M.A data classification method using genetic algorithm and *K*-means algorithm with optimizing initial cluster center[C]//2018 IEEE International Conference on Computer and Communication Engineering Technology，Beijing，2018：224-228.