

基于密度优化初始聚类中心的 K-means 算法

王艳娥¹, 安健², 梁艳¹, 康晶晶³

(1. 西安思源学院 理工学院, 陕西 西安 710038;

2. 西安交通大学深圳研究院, 广东 深圳 518057;

3. 山西农业大学 信息学院, 山西 晋中 030800)

摘要:针对 K-means 算法随机选择初始聚类中心,对噪音和异常点比较敏感,聚类结果过多依赖于专家经验从而缺乏一定客观性的问题,提出一种新的度量样本密度的方法优化 K-means 算法对初始聚类中心的选择。该方法基于样本实际分布,以最优超球体中样本个数与超球体中样本相似性作为度量样本密度的关键,能够有效选出较优的聚类中心,使得选择的初始聚类中心更接近样本集的实际分布。算法在乳腺癌数据集、常用 UCI 数据集以及人工模拟数据集上进行测试,实验结果表明,与已有同类方法相比,该算法在各数据集上的聚类评价指标均有提高,而且运行速度更快,聚类结果更稳定,聚类准确率更高:在乳腺癌数据集 wdbc 上的准确率为 91.04%,提高了 6%。在 Iris 数据集上的准确率为 94%,提高了 5%。

关键词:K-means 算法;密度;去噪;最优超球体;均方差;噪声数据

中图分类号:TP181

文献标识码:A

文章编号:1673-629X(2020)12-0099-07

doi:10.3969/j.issn.1673-629X.2020.12.018

K-means Algorithm Based on Density Optimization Initial Clustering Center

WANG Yan-e¹, AN Jian², LIANG Yan¹, KANG Jing-jing³

(1. School of Technology, Xi'an Siyuan University, Xi'an 710038, China;

2. Shenzhen Research Institute of Xi'an Jiaotong University, Shenzhen 518057, China;

3. School of Information Engineering, Shanxi Agricultural University, Jinzhong 030800, China)

Abstract: The K-means algorithm randomly selects the initial clustering center, which is sensitive to noise and outliers. The clustering results are too dependent on expert experience and thus lack of objectivity. In order to solve the problem, we propose a new method of measuring sample density to optimize the selection of the initial clustering center by K-means algorithm. Based on the actual distribution of samples, this method takes the number of samples in the optimal hypersphere and the similarity of samples in the hypersphere as the key to measure the sample density, and can effectively select the optimal clustering center, so that the selected initial clustering center is closer to the actual distribution of the sample set. The algorithm is tested on the breast cancer data set, UCI data set and artificial simulation data set. The experiment shows that compared with the existing similar methods, the proposed algorithm improves the clustering evaluation index on each data set, and runs faster, with more stable clustering results and higher clustering accuracy. The accuracy rate on wdbc is 91.04%, increased by 6%. The accuracy on Iris is 94%, up 5%.

Key words: K-means algorithm; density; de-noisy; optimal super sphere; mean square error; noise data

0 引言

聚类是数据挖掘中一种无监督学习分析数据的方法,基于“物以类聚”的思想,根据相似性原则将相似性较高数据划归同一类,相似性较低数据划分为不同

类^[1]。聚类分析的无监督特性,使聚类在医疗诊断、交通检测、图像处理、环境检测和大数据等方面得到广泛的应用。聚类分析方法可分为:基于划分式、基于网格、基于密度、基于层次和基于模型等五种类型^[2-3]。

收稿日期:2020-01-21

修回日期:2020-05-22

基金项目:深圳市科技计划项目(JCYJ20170816100939373);陕西省教育科学研究计划项目(18JK1100);陕西省高等教育科学研究项目(XGH19236)

作者简介:王艳娥(1979-),女,硕士,讲师,CCF会员(B6397M),研究方向为数据挖掘。

1 K-means 算法和研究现状

1.1 K-means 算法

K-means 算法^[4]核心思想是随机选取 k 个样本作为初始聚类中心,以欧氏距离作为相似度指标,两个样本之间距离越远相似性越低,距离越近相似性越高,通过不断迭代聚类中心,将相似性高的样本划分为同一类,相似性低的样本划分为不同类。K-means 具有明显的缺陷:(1)需随机选择初始聚类中心;(2)对噪声数据和异常点比较敏感;(3)需提前指定划分类数,使得聚类结果常陷于局部最优。因此关于 K-means 算法的优化,现有文献和相关学者主要是从这三方面展开。文中算法主要研究的是初始聚类中心的选择和噪声数据。

1.2 K-means 算法研究现状

为了解决 K-means 算法的缺陷,众多学者提出了基于密度优化的解决方案。文献[5]通过准则函数确定样本集的最佳聚类数,基于密度选择初始聚类中心,在一定程度克服了 K-means 算法需要预先输入类数和随机选择初始聚类中心的缺陷,聚类结果稳定,但在选择初始聚类中心时需根据经验输入样本邻域半径和最小样本密度两个参数使得算法的聚类结果缺少客观性;文献[6]算法划分出样本空间的高密度区域,在高密度区域选择距离最远的高密度样本作为初始聚类中心,但高密度区域仍需要人为输入样本邻域半径和最小样本密度,也使聚类结果受人为作用干扰大;文献[7]以最大最小距离法为基础,提出离积法的优化 K-means,该算法克服最大最小距离法易导致聚类中心稠密问题,但最大最小距离法将样本空间划分为高密度区域和低密度区域需要人为输入两个参数,这缺点文献[7]并没有克服;文献[8]提出噪声点优化 K-means 算法,在剔除噪声点需要根据经验设定两个参数:样本集最佳噪声样本数和判断样本是否为噪声样本的距离调节系数;文献[5-8]手动输入参数需要历史经验,聚类结果受人为干扰较大,使算法的普适性受到限制。文献[9-10]提出将方差作为选择初始聚类中心的指标,选择数据集中方差最小且处于不同区域的数据对象作为初始聚类中心,该算法的聚类结果稳定,且对噪声数据具有一定的免疫性,但选择的初始聚类中心与数据集实际类中心存在差异,且没有考虑噪声样本在聚类过程中的影响;文献[11]使用平均距离作为计算样本密度的指标,在一定程度避免将噪声点作为初始聚类中心,但选择的初始聚类中心同样与样本集实际中心分布相差较大。

该文在研究上述算法的基础上,提出基于样本规模的最优超球体计算样本密度,使样本密度的计算具有一定的客观性,克服文献[5-8]根据经验输入参数

的缺陷;文献[9-11]虽然确保初始聚类中心不会落在噪声样本,但导致密度最大的样本往往位于多个类的相交处,而不是数据集实际类中心。

2 基于密度去噪的 K-means 算法

2.1 DDK-means 算法相关概念

设 RP 为待聚类的样本空间,含有 n 个样本的样本集 $D = \{x_i \in D^p, |i = 1, 2, \dots, n\}$, 样本空间可划分为 k 类, 设 k 个聚类中心为数据集 $C = \{c_i \in C | i = 1, 2, \dots, k\}$ 。文中算法采用欧氏距离来衡量样本相似度。距离越远相似度越低,反之相似性越高。

(1) 样本 x_i 距离均值 $dm(x_i)$ 如下:

$$dm(x_i) = (1/(n-1)) * \sum_{j=1}^n \text{dist}(x_i, x_j) \quad (1)$$

其中, $j = 1, 2, \dots, n$, 且 $i \neq j$, $\text{dist}(x_i, x_j)$ 为样本 x_i 和 x_j 的距离。

(2) 样本集的均方差 msd 如下:

$$msd = \sqrt{\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (\text{dist}(i, j) - dm(x_i))^2} \quad (2)$$

(3) 样本集的超球体 v 的函数表示如下:

$$v = \pi R^3 \quad (3)$$

其中, $R = \mu * msd$, μ 为调节系数,初始值等于 1。 v 的大小应该与样本集 n 的大小和类簇数 k 相关。假设样本集中所有样本被均匀分配给 k 个类,那么每个类中应包含样本的个数 n/k ,考虑到噪声数据,规定每类样本的个数必须小于 n/k ,实际上不管样本集中的样本是否均匀分配给 k 类,通过规定超球体内样本个数不超过 n/k 都能计算出每个样本的最佳 μ 和最佳局部密度。

(4) 样本 x_i 的密度函数 $\text{density}(x_i)$ 如下:

$$\text{density}(x_i) = \rho(x_i) + \frac{1}{1 + \frac{1}{\rho(x_i)} \sum \text{dist}(x_i, x_j)} \quad (4)$$

其中, $\rho(x_i)$ 表示落入以 x_i 为中心的超球体 v 中的样本个数, $\frac{1}{\rho(x_i)} \sum \text{dist}(x_i, x_j)$ 表示超球体内所有样本与 x_i 的距离均值。 $\rho(x_i)$ 与超球体 v 密切相关,当 v 过小会导致落入 v 中的样本很少甚至没有, v 过大样本密度度量就会过于粗糙,因此选择合适的 v 至关重要。

从式(4)可以看出, $\text{density}(x_i)$ 值与 $\rho(x_i)$ 密切相关,当 $\rho(x_i)$ 的值越大说明落入以 x_i 为中心的超球体的样本越多,样本 x_i 越接近类中心。当 $\rho(x_i)$ 相同时,超球内样本与 x_i 距离越近,距离均值越小,该类样本密集度越高,则 x_i 越接近高密度区域的类中心。作为样本 x_i 的密度 $\text{density}(x_i)$ 的值越大, x_i 成为初始聚

类中心的权重越大。

(5) 样本集的密度值 meanD 表示如下:

$$\text{meanD} = \frac{1}{n} \sum_{i=1}^n \text{density}(x_i) \quad (5)$$

(6) 样本集聚类误差平方和 SSE 表示如下:

$$\text{SSE} = \sum_{i=1}^k \sum_{j \in C_i} \text{dist}(x_j, c_i) \quad (6)$$

2.2 DDK-means 算法原理

均方差在概率统计中用于测量样本集的分布程度,对于数据集可以通过均方差测量数据集的整个离散程度,当均方差的值越大说明数据集越分散,均方差越小数据集越集中。文中以均方差作为计算最优超球体的基础,将整个聚类分为两个阶段:第一阶段计算每个样本的局部密度。在大小相同的超球体内,某个样本的超球体内样本个数越多,则说明该样本处于高密度区域,作为初始聚类中心的权重就越大。根据式(3)计算所有样本的局部密度,当多个样本的超球体内的样本数相同时,则某个样本的超球体内样本紧密度起作用,越紧密,样本的密度越大,样本作为初始聚类中心的权重越大。当各个样本的超球体内的样本数不同时,则超球体内的样本数起作用,样本的超球体内样本数越多,样本密度越大,该样本作为初始聚类中心的权重越大。

第二阶段根据密度选取最佳的聚类中心,完成整个样本集的划分。选择大于样本集平均密度的样本作为初始聚类中心的候选集,同时非初始聚类中心候选集中选取样本密度较低的样本作为噪声样本,将整个样本集划分为非噪声样本集和噪声样本集;接着在候选样本集中同样以均方差为基础,通过可控的伸缩尺度调节样本的距离,选出 k 个密度较大且处于不同密度区域的样本作为初始聚类中心,然后对非噪声样本集进行聚类,完成非噪声样本的划分;最后对噪声样本集中的样本,根据它们与 k 个中心的相似度,将噪声样本划分给对应的类。

2.3 DDK-means 算法实现

根据 DDK-means 算法原理,算法实现步骤分如下两步:

第一步,算法 1:根据新定义的样本密度,将初始样本集划分为初始聚类中心候选样本集、非初始聚类中心候选集、噪声样本集和非噪声样本集。求解样本密度的算法描述如下:

输入: $x_i, \{x_i \in D | i = 1, 2, \dots, n\}$, D 为样本集; k ; 密度调节系数 $\mu = 1$; 初始聚类中心候选集 $D_1 = \varphi$; 非初始聚类中心候选集 $D_2 = \varphi$; 非噪声数据集 $D_3 = \varphi$; 噪声数据集 $D_4 = \varphi$ 。

输出: n 个样本的密度、 D_1, D_2, D_3 和 D_4 , 其中 D_1

$\cup D_2 = D, D_1 \cap D_2 = \emptyset, D_3 \cup D_4 = D, D_3 \cap D_4 = \emptyset$ 。

第 1 步:根据式(1)、式(2)计算样本集的均方差 msd 。

第 2 步:根据式(3)计算样本集的超球体。

第 3 步:根据式(4)计算每个样本的密度。如果样本的最大密度远远小于 n/k , 转到第 2 步,增大式(3)中的 μ 的值,重新计算超球体,使得超球体内样本个数增大,增大到刚好小于或等于 n/k , 转到第 4 步。如果样本最大密度远远大于 n/k , 转到第 2 步,减少式(3)中 μ 的值,重新计算超球体,使得超球体内样本个数减少,减少到刚好小于或等于 n/k , 转到第 4 步。

第 4 步:计算样本集的密度 meanD 。

第 5 步:构造初始聚类中心候选集 $D_1, \{x_i \in D_1 | \text{density}(x_i) > \text{meanD}, i = 1, 2, \dots, n\}$, 非初始聚类中心候选集 $D_2 = D - D_1$ 。

第 6 步:构造噪声数据集 D_4 和非噪声数据集 D_3 。其中 $D_4 = \rho * D_2, 0 \leq \rho \leq 1$, 即在 D_2 中选择样本密度最小的前 $\rho * |D_2|$ 样本作为噪声样本;构造非噪声样本集 $D_3, D_3 = D - D_4$ 。

第 7 步:算法 1 结束。

第二步,算法 2 根据算法 1 的结果,通过不断调节不同聚类中心之间的距离,在初始聚类中心候选集中选择密度最高且处于不同区域的样本作为初始聚类中心。再根据选择的最优初始聚类中心,先针对非噪声数据完成聚类,再将非噪声数据划分到不同的类簇中,从而剔除噪声数据对聚类过程产生的影响。

算法 2:具体实现的步骤如下:

输入:构造 k 空集合 S_1, S_2, \dots, S_k , 初始化为 $c_1 \in S_1, c_2 \in S_2, \dots, c_k \in S_k$; n 个样本的密度、 D_1, D_2, D_3 和 D_4 。

输出:样本集的 k 个划分。

第 1 步:在 D_1 中选择密度最大的样本作为第一个初始聚类中心 c_1 。

第 2 步:在 D_1 选择样本 x_i 作为第二个初始聚类中心 c_2, x_i 满足 $\text{dist}(x_i, c_1) > \text{msd}$ 。

第 3 步:在 D_1 选择样本 x_r 作为第 $r+1$ 个聚类中心, x_r 满足条件 $\text{dist}(x_r, c_1) > \text{msd}/(r-1) \&\& \text{dist}(x_r, c_2) > \text{msd}/(r-1) \&\& \dots \&\& \text{dist}(x_r, c_{r-1}) > \text{msd}/(r-1)$, 其中 $2 \leq r \leq k$ 。直到选择出第 k 个初始聚类中心。

第 4 步:根据每个样本与聚类中心的距离将非噪声数据划分到 K 个类中,重新计算 K 个类的聚类中心。

第 5 步:根据式(6),计算 SSE, 如果 SSE 发生变化转到第 3 步,否则转到第 6 步。

第 6 步:根据噪声数据与聚类中心的聚类,将噪声数据划分到 K 个类中,完成聚类。

3 DDK-means 算法仿真实验

为验证文中算法的有效性,分别在乳腺癌数据集、UCI^[12]数据库中常用的几个数据集以及人工数据集中进行测试,并与传统的 K-means 方法、文献[9,11]中的算法进行比较。所有算法的实验环境为:Win7 操作系统、COREi5 处理器、2G 内存、Matlab R2012a 处理软件。

3.1 实验数据集

3.1.1 乳腺癌数据集

用于测试的乳腺癌数据集为 wdbc 和 breast-cancer-wisconsin。breast-cancer-wisconsin 数据集包

表 1 人工模拟数据集各项参数

参数	第 1 类	第 2 类	第 3 类
均值	$\mu_x^1 = 3 \mu_y^1 = 3$	$\mu_x^2 = 6 \mu_y^2 = -2$	$\mu_x^3 = 9 \mu_y^3 = 3$
标准差	$\sigma^1 = 3.24$	$\sigma^2 = 1$	$\sigma^3 = 0.5 \sigma^1 = 4$

用于进行算法测试的人工模拟数据集包含 6 组数据集,6 数据集各包含 1 800 个样本,类别数为 3,每类簇包含 600 个样本,每类数据集按照不同的高斯分布生成。按照表 1 所示的各项参数生成含有不同噪声比的数据集,噪声比分别为 0%,10%,20%,30%,40%,50%,其中噪声产生在第 3 类,噪声数据的标准差为 4。

3.2 实验结果与分析

文中算法在乳腺癌数据集、UCI 数据集和人工模拟数据集的测试结果分析,通过常用的聚类效果评价指标:聚类误差平方和、聚类时间、聚类准确率^[13]、Rand index^[14]、Jaccard coefficient^[15]、Adjusted rand index^[16]进行比较。传统 K-means 算法,随机选择初

始聚类中心,聚类结果不稳定,

3.1.2 UCI 数据集和人工模拟数据集

为验证文中算法的普适性,在 UCI 数据库中选择机器学习用来进行测试的数据集进行验证,包括 Iris、Wine、Ionosphere、Soybean-small 和 Seed 数据集。

为进一步验证文中算法的合理性,生成包含不同噪声比的人工模拟数据集。关于人工模拟数据集高斯分布的相关参数如表 1 所示。

始聚类中心,聚类结果不稳定,

为加强 K-means 算法评价指标的稳定性,采取在测试数据集上重复执行 K-means 算法 100 次,K-means 算法的各项评价指标是执行 100 次后的平均值。

为验证文中算法能够很好地克服以上算法存在的缺陷,将文中算法与传统 K-means 算法、文献[9,11]提出的算法进行对比。

3.2.1 乳腺癌数据集与 UCI 数据集聚类结果分析

K-means 算法、文献[9]、文献[11]和文中算法在乳腺癌数据集和 UCI 数据集上的聚类误差平方和、运行时间如表 2 和表 3 所示。

表 2 四种算法在 UCI 数据集上的聚类误差平方和比较

数据集	K-means	文献[9]	文献[11]	文中
wdbc	7.794 3e+007	7.794 3e+007	7.794 3e+07	2.544 8e+07
breast-cancer-wisconsin	1.932 3e+04	1.932 3e+04	1.932 3e+04	1.304 3e+04
Iris	92.593 2	78.945 1	78.940 8	13.578 3
Wine	2.420 6e+006	2.370 7e+006	2.370 7e+06	6.541 8e+05
Ionosphere	2.382 3e+003	2.387 3e+003	2.387 3e+03	585.523 4
Seeds	203.206 1	203.206 1	203.206 1	183.247 7

表 2 中加粗数据表示该算法的聚类误差平方和评价指标最佳。从表 2 中的实验结果数据可以看出,文献[9]、文献[11]在 Iris 和 Ionosphere 数据集的聚类误差平方和明显优于 K-means 算法,在其他数据集中与 K-means 算法相同;文中算法在乳腺癌数据集以及几个常用的 UCI 数据集上的聚类误差平方和均明显低于 K-means 算法、文献[9]和文献[11];结果说明,文中算法能够将相似性高的样本划分为同一类,相似性低的样本划分为不同类,聚类的结果更符合数据集的

原始分布。

表 3 是四种算法在样本集上运行时间比较。从表 3 可以看出 K-means 算法在聚类时间上明显优于文献[9]、文献[11]和文中算法,结果产生的原因是其他三种算法在选择最优的初始聚类中心时有一定的时间开销;但文中算法在运行时间上明显优于文献[9]和文献[11],文中算法在对样本进行聚类时,减少反复聚类时的样本集规模,噪声样本并没有参与反复聚类的过程,当对非噪声样本完成聚类后,噪声样本一次性直

接划分给相似性高的类;同时由于文中算法选择的初始聚类中心更接近样本集实际中心的分布,使得反复聚类的迭代次数减少,进一步降低了时间开销。

表 3 UCI 数据集四种算法聚类时间比较

数据集	K-means	文献[8]	文献[10]	文中
wdbc	0.004 3	0.124 3	0.146 3	0.058 6
breast-cancer-wisconsin	0.108 6	0.120 7	0.161 6	0.059 8
Iris	0.004 9	0.140 9	0.129 6	0.071 8
Wine	0.004 7	0.096 2	0.112 7	0.071 1
Ionosphere	0.003 7	0.124 3	0.090 7	0.064 7
Seeds	0.005 4	0.115 0	0.121 9	0.059 4

图 1 是 K-means、文献[9]、文献[11]和文中算法

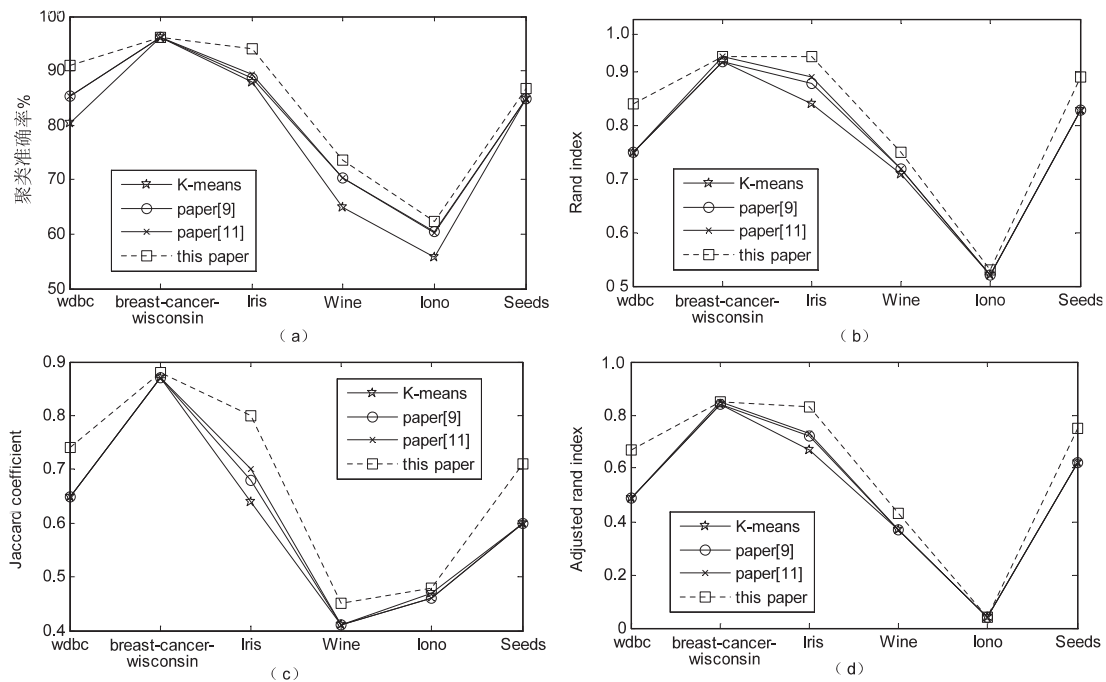


图 1 四种算法在 UCI 数据集上的结果比较

通过在乳腺癌数据集和常用的 UCI 数据集进行聚类结果的比较,证明文中提出的优化 DDK-means 算法的聚类效果明显优于其他三种聚类方法,其中 K-means 算法的聚类效果最差,文献[9]和文献[11]的聚类结果相似,文中算法有效地克服了优化后初始聚类中心与样本实际类中心差异较大的缺陷。

3.2.2 人工数据集结果分析

在人工模拟数据集上对 K-means 算法、文献[9]、文献[11]和文中算法进行测试。除了在六种聚类效果评价指标进行对比外,对四种算法选择的初始聚类中心进行了比较,四种算法选择的初始聚类中心如图 2 所示。图 2 中黑白相间的圆表示不同算法在不同噪声比数据集上选择的初始聚类中心。

K-means 算法的初始聚类中心是随机产生,初始聚类中心不稳定,图 2 中的 K-means 初始聚类中心是随机选取其中一次的结果;文献[9]、文献[11]和文中

在乳腺癌数据集和 UCI 数据集上在聚类准确率、Rand index、Jaccard coefficient 和 Adjusted rand index 参数指标的比较折线图。图 1(a)中,文中算法在这几个数据集上的聚类准确率最优,K-means 算法的聚类结果最差;图 1(b)中,文中算法的 Rand index 明显优于其他三种算法,K-means 算法的聚类效果最差;图 1(c)中,文中算法的 Jaccard coefficient 均优于其他三种算法,而且在 wdbc、Iris 和 Seeds 样本集的优势明显;图 1(d)中,文中算法的 Adjusted rand index 在 wdbc、Iris、Wine、Seeds 数据上明显优于其他三中算法,在 breast-cancer-wisconsin 和 Ionosphere 数据上也具有一定的优势。

算法选择的初始聚类中心稳定。图 2 选取具有代表性的无噪声数据集、20% 噪声数据集、50% 噪声数据集,在这三个数据集上运行 K-means 算法、文献[9]、文献[11]和文中算法;图 2(a)~(d)分别是 K-means 算法、文献[9]、文献[11]和文中算法在三个数据集上选择的初始聚类中心。图 2(a)是 K-means 算法选择的初始聚类中心,随机选择的初始聚类使得初始的中心往往不够理想,不同类簇的初始聚类中心可能位于在同一类中,甚至可能为噪声数据,这样极大概率导致 K-means 聚类结果不稳定且趋于局部最优;图 2(b)是文献[9]选择的初始聚类中心,文献[9]基于方差优化后选择的初始聚类中心稳定,能够保证聚类中心分布在不同区域,且初始聚类中心稳定,但从图中可以看出文献[9]选择的初始聚类中心偏离数据集真实的聚类中心;图 2(c)是文献[11]选择的初始聚类中心,图 2(c)能够保证初始聚类中心选择稳定,且处于不同的

区域,但初始聚类中仍然偏离数据集真实中心;图2(d)是文中算法的结果,可以看出文中算法选择的初

始聚类中心分别位于三类样本密集区域,初始聚类中心更接近样本集实际类中心。

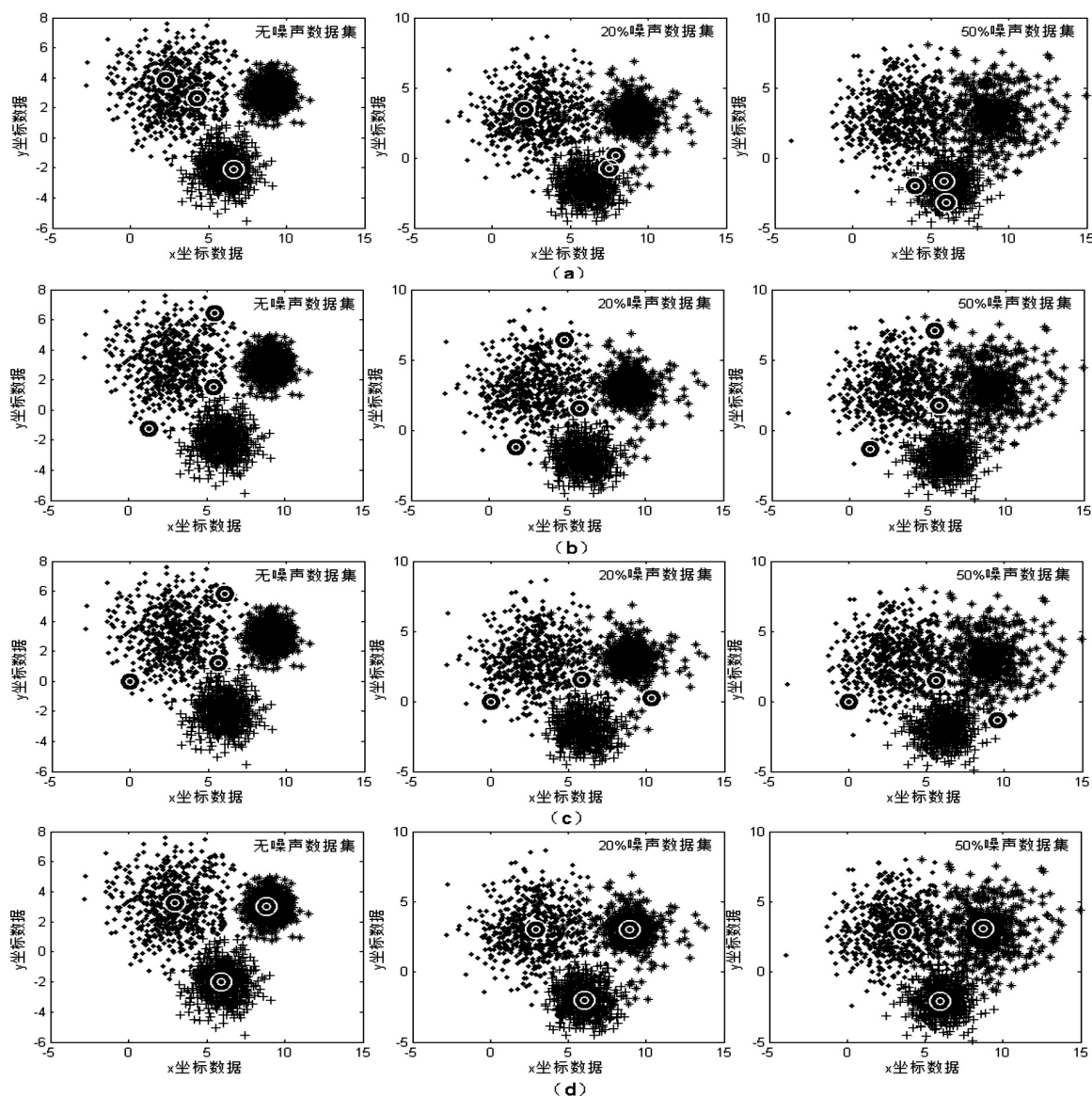


图2 四种算法选择的初始聚类中心

表4和表5是四种算法在不同噪声比的6组人工模拟数据集上的聚类误差平方和比较和算法运行时间

表4 人工模拟数据集聚类误差平方和比较

噪声数据	K-means	文献[9]	文献[11]	文中
0%	5.321 6e+03	5.321 6e+03	5.321 6e+03	2.381 5e+03
10%	5.622 7e+03	5.622 7e+03	5.622 7e+03	2.491 7e+03
20%	5.753 0e+03	5.753 0e+03	5.753 0e+03	2.645 2e+03
30%	6.093 0e+03	6.093 0e+03	6.093 0e+03	2.851 0e+03
40%	6.628 7e+03	6.628 7e+03	6.628 7e+03	3.172 9e+03
50%	6.935 5e+03	6.935 5e+03	6.935 5e+03	3.334 3e+03

表4中用加粗数据表示该算法的聚类评价指标最佳。从表4中提供的数据可以看出,文中算法在不同噪声比的人工模拟数据集上的聚类误差平方和均明显优于K-means算法、文献[9]和文献[11];文献[9]和

文献[11]在人工模拟数据集上的聚类误差平方和与K-means相同。

表5中K-means算法在不同噪声比人工模拟数据集的运行时间明显均优于其他三种算法,但文中算

法的运行时间均优于文献[9]和文献[11]。

表 5 人工模拟数据集运行时间比较

噪声数据	K-means	文献[9]	文献[11]	文中
0%	0.004 1	0.226 3	0.490 7	0.162 5
10%	0.004 4	0.233 7	0.497 5	0.164 6
20%	0.004 7	0.228 1	0.455 9	0.163 2
30%	0.004 5	0.238 1	0.522 1	0.159 1
40%	0.004 0	0.240 9	0.539 3	0.161 4
50%	0.004 2	0.244 7	0.488 6	0.159 5

图 3(a)~(d)分别是 K-means、文献[9]、文献[11]和文中算法在不同噪声比的人工模拟数据集上在聚类准确率、Rand index、Jaccard coefficient 和 Adjusted rand index 四种评价指标的比较折线图,可以看出文中算法在四种聚类评价指标上均明显优于其他三种算法。

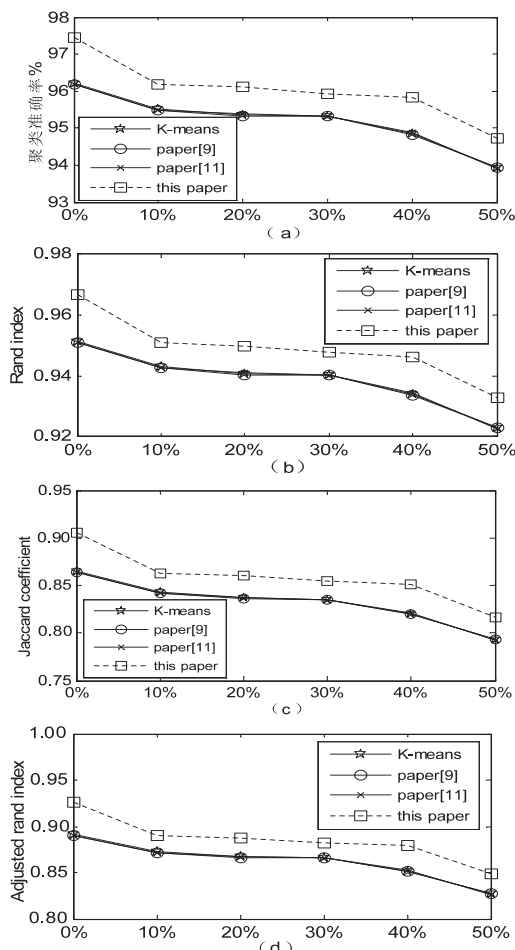


图 3 四种算法在不同噪声比人工数据集上的运行结果
人工模拟数据集上的聚类结果进一步说明,文中算法能够克服选择的初始聚类中心与数据集实际中心分布差异较大的问题。

4 结束语

针对现有基于密度优化 K-means 算法存在的问

题,提出密度去噪的 DDK-means 算法,通过样本集的规模和样本类簇数对样本密度的最大值进行限定,同时根据样本集的密度均值剔除样本集中的噪声样本,克服需要手动输入参数以及噪声样本参与整个聚类的缺陷。与同类文献对比,实验结果证明文中算法不仅在乳腺癌数据集的聚类结果稳定、聚类准确率提高明显、对噪声数据不敏感,且在其他 UCI 数据集上也具有较优的聚类效果。

参考文献:

- [1] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报, 2008,19(1):48-61.
- [2] JAIN A K, MURTY M N, FLYNN P J. Data clustering: a review[J]. ACM Computing Surveys, 1999, 31(3):264-323.
- [3] 陈黎飞,姜青山,王声瑞. 基于层次划分的最佳聚类数确定方法[J]. 软件学报, 2008,19(1):62-72.
- [4] MACQUEEN J B. Some methods for classification and analysis of multivariate observations[C]//Proceeding of the fifth Berkeley symposium on mathematical statistics and probability. Berkeley: University of California Press, 1967:281-297.
- [5] 张琳,陈燕,汲业,等. 一种基于密度的 K-means 算法研究[J]. 计算机应用研究, 2011,28(11):4071-4073.
- [6] 傅德胜,周辰. 基于密度的改进 K 均值算法及实现[J]. 计算机应用, 2011,31(2):432-434.
- [7] 熊忠阳,陈若田,张玉芳. 一种有效的 K-means 聚类中心初始化方法[J]. 计算机应用研究, 2011,28(11):4188-4190.
- [8] GAN Guojun, KWOK-PONG M. K-means clustering with outlier removal[J]. Pattern Recognition Letters, 2017, 90:8-14.
- [9] 谢娟英,王艳娥. 最小方差优化初始聚类中心的 K-means 算法[J]. 计算机工程, 2014,40(8):205-211.
- [10] 周炜奔,石跃祥. 基于密度的 K-means 聚类中心选取的优化算法[J]. 计算机应用研究, 2012,29(5):1726-1728.
- [11] 张素洁,赵怀慈. 最优聚类个数和初始聚类中心点选取算法研究[J]. 计算机应用研究, 2017,34(6):1617-1620.
- [12] FRANK A, ASUNCION A. UCI machine learning repository [D]. Irvine, USA: University of California, 2010.
- [13] SUN Y, ZHU Q M, CHEN Z X. An iterative initial-points refinement algorithm for categorical data clustering[J]. Pattern Recognition Letters, 2002,23(7):875-884.
- [14] RAND W M. Objective criteria for the evaluation of clustering methods[J]. Journal of the American Statistical Association, 1971,66(336):846-856.
- [15] HALKIDI M, BATISTAKIS Y, VAZIRGIANNIS M. On clustering validation techniques[J]. Journal of Intelligent Information Systems, 2001,17(2-3):107-145.
- [16] TOPCHY A, JAIN A K, PUNCH W. A mixture model for clustering ensembles [C]//The 2004 SIAM international conference on data mining. Florida: [s. n.], 2004:379-390.