改进的方差优化初始中心的 K-medoids 算法

张晓滨,母玉雪
(西安工程大学 计算机科学学院,陕西 西安 710600)

摘 要:针对传统 K-medoids 算法对于初值敏感、容易陷入局部最优解、稳定性差等缺点和方差优化初始中心的 K-medoids 聚类算法的时间复杂度较高、邻域半径不够精确等问题,提出一种改良的基于方差优化初始中心的 K-medoids 聚类算法。该算法引入了全局方差的概念,并将其作为样本的密度参数,选择部分方差值较小的样本作为候选初始聚类中心样本集,并利用最大距离乘积法从候选初始聚类中心样本集中选取方差值较小且距离较远的 K 个样本当作初始聚类中心,该算法充分兼顾了初始聚类中心的分散性和代表性。在更新簇类中心时,根据样本密度原则逐步扩大搜索范围,代替了传统的随机选取。通过在 UCI 数据集上的实验结果表明,该算法不仅有效优化了初始聚类中心点的选取,同时也有效改进了聚类速度和聚类效果。

关键词:K-medoids 算法;初始聚类中心;方差优化;最大距离乘积法;样本密度
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2020)07-0042-04
doi:10.3969/j.issn.1673-629X.2020.07.010

An Improved K-medoids Algorithm for Initial Center of Variance Optimization

ZHANG Xiao-bin, MU Yu-xue
(School of Computer Science, Xi'an Polytechnic University, Xi'an 710600, China)

Abstract: Aiming at the disadvantages of traditional K-medoids algorithm such as sensitivity to initial value, falling into local optimal solution easily, poor stability and the problems of variance optimization initial center K-medoids algorithm such as high time complexity and inaccurate neighborhood radius, we propose an improved K-medoids clustering algorithm based on the initial center of variance optimization. The concept of global variance is introduced in this algorithm and taken as a sample density parameters. Some smaller values of the variance of sample set are chosen as a candidate for the initial clustering center, and the method of maximum distance product is used to select K samples with small variance and far distance from the candidate initial clustering center set as the initial clustering center. The algorithm gives full consideration to the dispersion and representativeness of the initial clustering center. When updating the cluster center, the search scope is gradually expanded according to the sample density principle, which replaces the traditional random selection. Experimental results on UCI data set show that the proposed algorithm not only effectively optimizes the selection of initial clustering center, but also effectively improves the clustering speed and clustering effect.

Key words: K-medoids algorithm; initial cluster center; variance optimization; maximum distance product method; sample density

0 引 言

聚类分析是机器学习领域的一项关键技术,也是最重要的数据分析方法之一,在模式识别、图像分割、信息检索、疾病诊断、决策支持等方面有广泛的应用[1-2]。常见的聚类算法总体上可细分为:划分法、网格法、层次法、密度法与基于模型法[3-5]五大类别。聚类通过获取给定数据集的内在属性,将数据集划分为多个类簇,使得同一类簇中的样本相似度较大,不同类簇中的样本相似性相对较小,然后把给定数据集样本依据固有信息分别开,从而揭示数据样本集的初始分布。

基于划分的聚类方法首先通过预先指定初始聚类中心和聚类数目,采用迭代重定位技术反复迭代运算,使样本在划分类簇间移动,当目标函数的误差值达到收敛效果时,才获得最终的聚类结果。K-means 算法与 K-medoids 算法是比较典型的基于划分下的无监

收稿日期:2019-08-21 **修回日期**:2019-12-23
基金项目:陕西省自然科学基金(2015JQ5157)
作者简介:张晓滨(1970-),男,硕士,副教授,研究方向为数据库与数据挖掘技术、移动互联个性化服务技术与应用;母玉雪(1996-),女,硕士研究生,CCF 会员(D2073G),研究方向为数据挖掘与机器学习。

督聚类算法,和 K-means 算法相比而言,K-medoids 算法克服了对孤立点敏感的缺陷,有着较强的准确性和鲁棒性。为解决 K-medoids 算法初始化敏感、易陷入局部最优解等问题,提出了一系列 K-medoids 的改进算法[6-15]。Park 等人[7]首次提出快速 K-medoids 聚类算法,通过改良初始聚类中心的选择方式与聚类中心点的更新,缩短了聚类时间. 然而该算法在选取初始聚类中心时,样本密度的定义较为复杂,计算较为耗时,且未充分考虑到数据空间关系和数据集自身分布,使得初始聚类中心有可能位于同一个类簇。文献[8]利用密度信息产生初始聚类中心,但是牺牲了时间复杂度去换取较好的搜索性能,以求达到聚类性能提高的效果。文献[9]提出以方差作为样本分布密集程度的度量标准,分别以样本间距离均值和标准差为邻域半径,解决了邻域半径需要人为给定调节系数的缺陷,但是均值和标准差容易受到异常值影响,即邻域半径并非为最佳邻域半径。文献[10]提出了局部方差概念,引入了近邻概念定义样本的局部方差,但是近邻值需要人为给定,由于数据集大小不一,且属性各不相同,最佳邻域值不易选取。文献[12]提出了一种基于距离不等式的 K-medoids 聚类算法,但是算法复杂度较高。

针对传统 K-medoids 算法和方差优化初始聚类中心 K-medoids 算法的潜在不足,文中改进了样本密度的定义,并利用最大距离乘积法,选取样本密度较高且距离较远的 K 个样本为初始聚类中心。在更新簇类中心时,根据样本密度原则逐步扩大搜索范围,代替了传统的随机选取,在保证聚类质量的条件下提高了聚类速度。通过在 UCI 数据集的实验,证明了该算法的有效性。

1　传统 K-medoids 算法及其改进

1.1　基本概念与定义

假设包含 N 个样本的数据集 $X=\{x_1,x_2,\cdots,x_i,\cdots,x_N\}$,对于其中任一样本 x_i,类簇数目为 K,即 $X=\{C_1,C_2,\cdots,C_K\}$,$O=\{o_1,o_2,\cdots,o_K\}$ 为类簇中心样本集合,其中 $K<N$。

定义 1:任意两个样本 x_i 和 x_j 间的距离采用欧几里得距离,则

$$d(x_i,x_j)=\sqrt{\|x_i-x_j\|^2} \tag{1}$$

定义 2:数据标准化(normalization)不仅能够加快梯度下降求最优解的速度,还有可能提高数据精度。线性数据标准化是对原始数据的一种线性变化,使得结果映射到[0,1]范围内。转换函数为:

$$x'_{i,j}=\frac{x_{i,j}-\min(X_{:,j})}{\max(X_{:,j})-\min(X_{:,j})} \tag{2}$$

其中,max 是样本数据的最大值,min 是样本数据的最小值。

定义 3:样本 x_i 的全局方差定义为:

$$F_{(x_i)}=\frac{\sum_{j=1}^{N}[d(x_i,x_j)-\mathrm{aver}(x_i)]^2}{N-1} \tag{3}$$

其中,$\mathrm{aver}(x_i)=\dfrac{\sum_{j=1}^{N}d(x_i,x_j)}{N}$。

1.2　传统 K-medoids 聚类算法

典型的 K-medoids 算法描述为:

输入:含有 N 个样本的数据集 $X=\{x_1,x_2,\cdots,x_i,\cdots,x_N\}$,类簇的数目 K;

输出:K 个最优簇集合。

处理流程

Step1:随机选取 K 个样本数据 $\{o_1,o_2,\cdots,o_K\}$ 作为初始聚类中心。

Step2:分配样本。根据式(1)计算剩余样本点到聚类中心的距离,并将其划分到离它最近的聚类中心点所代表的类簇。

Step3:更新类簇中心。随机选取非中心点 o_{random},依照平方差函数值减少原则,用 o_{random} 代替 o_j,更新每个簇的聚类中心点,平方差函数可以定义为 $J_c=\sum_{j=1}^{K}\sum_{p=C_i}|p-o_j|^2$,其中 o_j 是聚类中心点,p 是簇 C_i 中的样本。

Step4:重复执行 Step2 和 Step3,直到每个类簇不再发生变化。

Step5:输出划分的 K 个类簇。

1.3　Num-近邻方差优化初始中心的 K-medoids 算法

文献[10]算法是把方差当作样本分布密度的度量值,选取离 x_i 最近的 Num 个样本计算 Num-近邻局部方差,选择方差最小的样本 $x_{\mathrm{sort}(k)}$ 当作第一个初始聚类中心,计算样本标准差 $\mathrm{std}_{\mathrm{sort}(k)}$,并将 $\mathrm{std}_{\mathrm{sort}(k)}$ 作为邻域半径,计算 $x_{\mathrm{sort}(k)}$ 邻域 $\mathrm{Neighbor}_{\mathrm{sort}(k)}$,在样本集中去除 $\mathrm{Neighbor}_{\mathrm{sort}(k)}$,再选取剩余样本集中方差最小的样本作为聚类中心,直至选够 K 个。该算法被称为 SD(standard-deviation as radius of neighborhood)算法,主要的改进点在于初始聚类中心的选择上,解决了初始聚类中心可能位于同一类簇中心的缺陷。

2　文中算法

2.1　算法的改进思想

通常使用方差来衡量各个数据样本点及其均值之间的偏离程度,如果数据分布比较分散,则方差相对较大,亦说明数据对象的波动相对较严重。根据方差的

定义可知,一个数据样本集下方差最小的样本常常处于数据集的中心区域,或者是样本分布较为集中的区域。文中所提算法兼顾考虑到样本密度和其样本集的空间距离特征,表现优良的初始簇类中心点不仅需要所在区域样本数量较多,而且还应该有良好的独立性,各个中心点应该彼此分散而不集中,相互距离应该较远,才能使得各个簇类中心点尽可能代表不同的类簇。

初始聚类中心选取基本思路:首先,计算样本的全局方差,并按照数值大小升序排序,将前 K^2 个低方差样本存入候选聚类中心集合 P 中;然后依照最大距离乘积法,从集合 P 中选取 K 个方差值较小且相对分散的初始中心存入聚类中心集合 O,使得 $O=\{o_1,o_2,\cdots,o_K\}$。将选取好的集合 O 作为初始聚类中心点。文中所提算法既保证初始聚类中心处在样本分布的密集区域,还保证了初始聚类中心处于不同的类簇,从而摆脱了对人为赋值的依赖。

2.2 算法的具体描述

输入:含有 N 个样本的数据集 $X=\{x_1,x_2,\cdots,x_i,\cdots,x_N\}$,类簇的数目 K;

输出:K 个最优簇集合。

处理流程

Step1:利用转换函数式(2)对样本进行数据标准化。

Step2:初始聚类中心的选取。假设候选初始聚类中心集合 P 和初始聚类中心集合 O 均为空集。

Step2.1:利用式(1)计算各个样本间的距离,并根据式(3)计算各个样本的全局方差,并将方差值最小的前 K^2 个样本加入高密度点集合 P,$P=\{p_1,p_2,\cdots,p_{K^2}\}$,集合 P 为候选初始聚类中心点集合。

Step2.2:根据式(1)和式(3),首先从集合 P 中选取方差值最小的样本 p_i 作为第一个初始聚类中心 o_1,然后在集合 P 中选取距离 o_1 最远的点 p_j 作为第二个初始聚类中心 o_2,并将 o_1、o_2 加入到初始聚类中心 O 中,即 $O=O\cup\{o_1,o_2\}$。

Step2.3:在集合 P 中选取满足条件 $\max(d(p_m,o_1)\times d(p_m,o_2))$ 的数据对象 p_m 作为第三个初始簇类中心 o_3 加入集合 O 中。

Step2.4:重复执行步骤 Step2.3,直到集合 O 中的初始聚类中心个数等同于 K,即初始聚类中心集 $O=\{o_1,o_2,\cdots,o_K\}$。

Step3:分配样本。根据式(1)计算剩余样本点到聚类中心的距离,并将其划分到离它最近的聚类中心点所代表的类簇。

Step4:更新类簇中心。对于非中心点 o_{random} 的选取,先对样本分布密度即方差大小进行排序,从簇内开始查找,按照平方差函数值减少原则,用 o_{random} 代替 o_j,逐步将搜索更新范围扩大至所有非中心点。更新每个簇的中心点。

Step5:重复执行 Step3 和 Step4,直到每个类簇不再发生变化。

3 实验结果与分析

为了测试文中所提算法的性能,分别采用传统 K-medoids 算法、快速 K-medoids 算法、文献[10]算法和文中算法在 UCI 机器学习数据集上进行了实验,实验环境为操作系统 Windows 10,4 GB 内存,Matlab 应用软件。

实验所采用的数据集为 UCI 机器学习数据库中经常用来测试聚类算法性能的数据集,数据集的描述如表 1 所示。

表 1 实验数据描述

数据集	样本数量	特征个数	聚类数
wine	178	13	3
iris	150	4	3
wdbc	569	30	2
heart	270	13	2
ionosphere	351	34	2

实验结果均为四种聚类算法分别聚类运行 60 次所得到的平均值,具体如图 1 和图 2 所示。

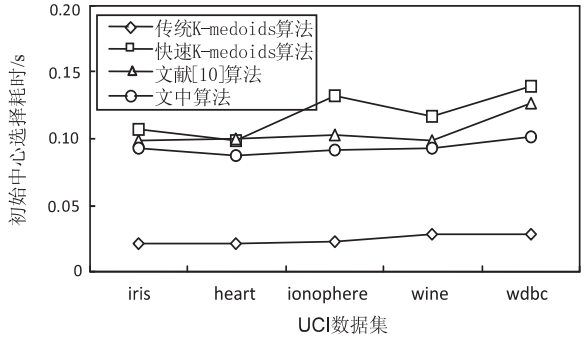


图 1 选取初始聚类中心耗时

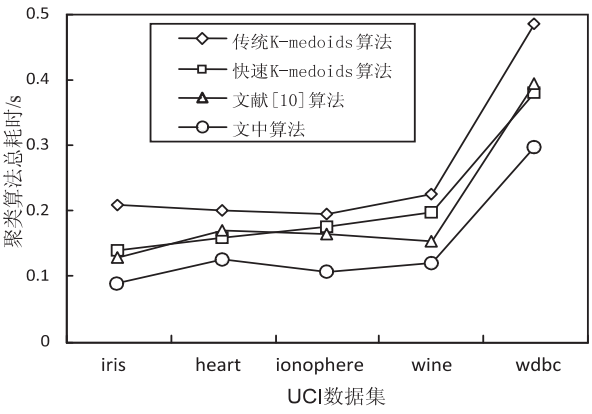


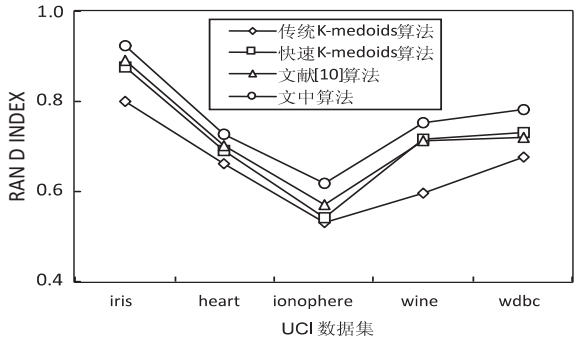
图 2 聚类算法总耗时

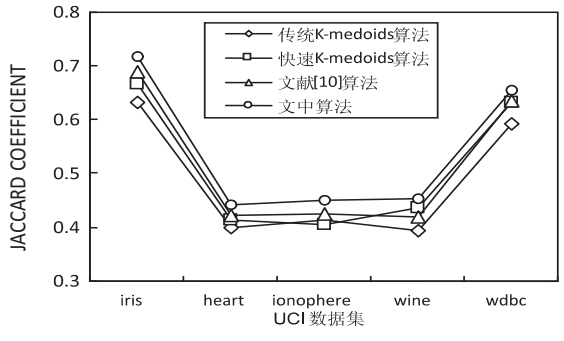
由图 1 可知,在选取初始聚类中心时,文中算法耗时高于传统 K-medoids 算法,低于快速 K-medoids 算法和文献[10]算法。因为传统 K-medoids 算法是随

机选取,并不需要额外的计算,所以耗时较少,且在各个不同大小的数据集中差别不大;文中算法改进了样本密度的衡量标准,简化了选取步骤,所以会略低于快速 K-medoids 算法和文献[10]算法。

由图 2 可知,文中算法的聚类总耗时要低于传统 K-medoids 算法、快速 K-medoids 算法和文献[10]算法。主要原因是传统 K-medoids 算法在初始聚类中心选择时有很大的随机性,导致迭代次数多,从而总耗时多;快速 K-medoids 算法选取的初始聚类中心点有很大可能处在同一个类簇下,容易导致初始聚类中心过于集中;文献[10]算法在邻域半径的计算上容易受噪音数据影响,导致邻域半径很难达到最佳。文中算法在初始时给出聚类中心点的大概位置,而选取的聚类中心在保证分散性的条件下,更具备代表性,因此减少了后期的迭代次数,提高了运算效率,降低了聚类算法总耗时。

各个算法的聚类性能评价指标分别采用 Rand 指数、Jaccard 系数、Adjusted Rand Index 指数、F-measure 和聚类准确率来衡量。前三个指标是在正确分类信息已明确的条件下对聚类性能评价的有效指标。其中,Rand 指数能够衡量聚类结果和原始数据集样本分布的同一性,Jaccard 系数则可以衡量实现正确聚类的样本占聚类前或聚类后在同一个类簇样本的比例,Adjusted Rand Index 指数越大,则说明实现正确聚类的样本对越多,而聚类效果也越好,它的上界为 1。由图 3 可知,文中算法在各方面表现较优于传统 K-medoids 算法、快速 K-medoids 算法和文献[10]算法,在不需要主观参数的情况下取得了较好的实验效果,提高了聚类的性能。





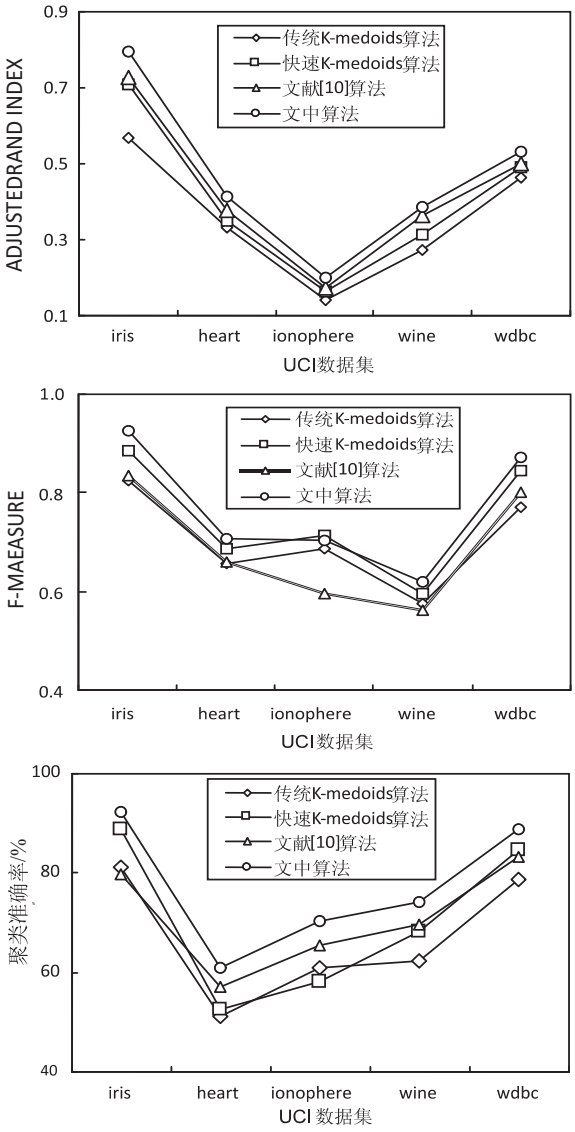


图 3　UCI 数据集的聚类结果对比

4　结束语

综上所述,在现有 K-medoids 算法不足的条件下,提出了一种改良的 K-medoids 聚类的优化算法。通过引入全局方差,并与最大距离乘积法相结合,优化了 K-medoids 聚类算法的初始聚类中心的选取,有效避免了初始化敏感的问题;同时改进了聚类时间和聚类性能。通过将该算法在数据集上进行测试,验证了算法的有效性。

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] ARORA P,DR D L,VARSHNEY S. Analysis of K-Means and K-Medoids algorithm for big data☆[J]. Procedia Computer Science,2016,78:507-512.

[3] 陈小雪,尉永清,任　敏,等. 基于萤火虫优化的加权 K-

(下转第 134 页)