基于医疗数据的聚类挖掘策略研究

王艳娥[1],安 健[2],王红刚[1],丁心安[1],杨 倩[1]
(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)

摘 要:基于医疗数据集,研究划分式聚类算法K-medoids。针对该算法随机选取初始聚类中心、收敛速度慢、聚类结果不稳定等问题,提出基于方差的密度优化算法。该算法以样本集的均方差和距离均值为基础,再根据样本集的大小计算样本集的密度半径,在相同密度半径下稠密区域的样本具有较高的密度,通过动态选择不同高密度区域的样本作为初始聚类中心,在进行聚类的过程中通过局部优化,加快收敛速度,解决传统K-medoids存在的缺点。将该优化算法应用在UCI机器学习的医疗数据集上测试聚类效果,实验验证该算法选择的初始聚类中心位于样本集的稠密区域,更符合数据集的原始分布,且在乳腺癌数据集上具有较高的聚类准确率,聚类结果稳定,收敛速度快。

关键词:医疗数据;K-medoids算法;聚类;密度优化;方差
中图分类号:TP311.5 **文献标识码**:A **文章编号**:1673-629X(2020)07-0066-05
doi:10.3969/j.issn.1673-629X.2020.07.015

Research on Clustering Mining Strategy Based on Medical Data Sets

WANG Yan-e[1],AN Jian[2],WANG Hong-gang[1],DING Xin-an[1],YANG Qian[1]
(1.School of Science and Technology,Xi'an Siyuan University,Xi'an 710038,China;
2.Shenzhen Research Institute of Xi'an Jiaotong University,Shenzhen 518057,China)

Abstract:Based on the medical data set,the partitioning clustering algorithm K-medoids is studied. A variance-based density optimization algorithm is proposed to solve the problems of random selection of initial clustering center,slow convergence speed and unstable clustering results in K-medoids algorithm. Based on the mean square deviation and distance mean of the sample set,the density radius of the sample set is calculated according to the size of the sample set. Samples in the dense region with the same density radius have higher density. By dynamically selecting the samples as initial clustering centers from different dense regions,local optimization is adopted in the clustering process to accelerate the convergence speed,so as to solve the shortcomings of traditional K-medoids. In order to test the clustering effect,this algorithm is applied to medical data set of UCI machine learning. The experiment shows that the initial clustering centers selected by the algorithm are located in the dense area of the sample set,which is more in line with the original distribution of the data set. The algorithm has higher clustering accuracy,more stable clustering results and faster convergence speed on breast cancer data sets.

Key words:medical data;K-medoids algorithm;clustering;density optimization;variance

0 引 言

聚类是数据挖掘中有效分析数据的手段之一,根据物以类聚的思想,将相关性较高的数据划分为一类,相关性较低的数据划分为不同类。聚类的优势是不需要数据的先验知识,算法会根据数据的内部特征,使用一定的相关性测量方法分析数据,从而得到隐藏在数据内部的数据之间的关系。聚类被广泛应用在图像处理[1]、大数据[2]、人工智能[3]等众多领域。现在随着医疗的信息化,医疗数据越来越庞大,医疗的图像、病理记录等以数据方式存储在电脑上,为有效协助医护人员对病情进行预测和诊断,聚类算法被越来越多地应用在医疗数据中[4-5]。常用的聚类算法有基于划分、基于层次、基于密度、基于模型和基于网格的聚类方法[6]。

K-medoids算法是划分式聚类算法的经典算法之一,因其原理简单,易实现得到广泛应用。传统的K-

收稿日期:2019-08-20 **修回日期**:2019-12-23
基金项目:陕西省教育科学研究计划项目(18JK1100);深圳市科技计划项目(JCYJ20170816100939373);陕西省高等教育科学研究项目(XGH19236)
作者简介:王艳娥(1979-),女,讲师,硕士研究生,研究方向为数据挖掘。

medoids 算法虽对噪声数据不敏感,但随机选取初始聚类中心,导致聚类结果不稳定。因此众多学者对 K-medoids 算法进行优化,提高算法的聚类准确率和稳定性。传统 K-medoids 算法的经典代表是 PAM 算法[7],PAM 算法随机选择初始聚类中心,且通过全局搜索更新聚类中心使得聚类时间消耗太大。优化 PAM 算法的经典算法是快速 K-medoids 算法(Park 算法)[8],该算法选择距离最小的样本作为初始聚类中心,克服聚类中心随机选择导致的不确定性,同时更新初始聚类中心时采用类内进行迭代,大大减少聚类时间,但 Park 算法选择出的聚类中心常处于同一个类中,并不符合样本集的实际分布。为克服 Park 算法的缺点,初始聚类中心的选择仍是 Park 算法研究的热点之一。基于样本集方差优化 K-medoids 初始聚类中心选择[9],选择方差最小且处于不同区域的样本作为初始聚类中心,虽然在一定程度上选择稳定的初始聚类中心,但这些样本仍然处于同一个类簇中,并不能反映样本的实际分布。基于宽度优先搜索的 K-medoids 聚类算法[10],以粒计算初始化获取初始有效粒子,再根据二叉树宽度优先搜索原理迭代出最优聚类中心,在一定程度克服传统 K-medoids 算法存在的缺陷,但对多维的医疗数据而言聚类时间消耗过大。

文中主要基于医疗数据研究 K-medoids 算法,在 PAM 和 Park 算法的基础上优化 K-medoids,以期优化后的算法能够在医疗数据上有较好的聚类效果。

1　PAM 算法和 Park 算法

PAM 算法的思想是在样本集中随机选取 K 个样本作为初始聚类中心,剩余样本按照与 K 个聚类中心距离的远近分配到最近的类簇中;然后反复选取任意非类中心样本取代聚类中心,再重新分配非类中心样本,直到满足一定的评价标准,聚类结束。

Park 算法针对 PAM 算法随机选取初始聚类中心和更新聚类中心时使用的是剩余全部样本进行搜索迭代这两个缺点进行优化。基本思想是以距离为基础计算每个样本的密度,选择前 K 个密度最大的样本作为初始聚类中心,在更新聚类中心时选取同一类簇中距离其他样本距离和最小的样本作为新的聚类中心,直到满足一定的评价标准,聚类结束。Park 算法虽然有效克服了 PAM 算法存在的问题,但是 Park 算法选择的 K 个初始聚类中心并不符合样本的实际分布,K 个样本常处于同一个类中,使得聚类的结果并不理想。

这两种算法评价标准的目标是使同类的样本尽可能相似,不同类的样本尽可能相异,采取常用的聚类误差平方和[11]作为评价结果的指标,聚类误差平方和越小说明同类簇的相似性越高,不同类相似性越低,聚类的结果符合样本的实际分布,聚类效果越好。

2　改进的 K-medoids 算法

设待聚类的样本集为 X,$X=\{x_1,x_2,\cdots,x_n\}$,其中 x_i 是一个包含 p 维数据的样本。将样本集划分为 K 类,聚类中心集为 C,$C=\{c_1,c_2,\cdots,c_k\}$,其中 c_i 表示第 i 类的聚类中心。文中算法以样本之间的距离度量样本的相似度,以聚类误差平方和作为目标函数,动态选择样本密度最高的 K 个样本且处于不同区域作为初始聚类中心,在更新聚类中心的过程中采用类内和类外相结合的方法,优化聚类中心的选择,减少聚类迭代次数,加快聚类过程。

2.1　相关概念

定义 1:样本 x_i,x_j 的欧氏距离为 $d(x_i,x_j)$:

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

定义 2:样本 x_i 的密度为 $\mathrm{density}(x_i)$:

$$\mathrm{density}(x_i)=f(x_i,r)+\frac{\mathrm{bwd}(x_i)}{\mathrm{wid}(x_i)} \tag{2}$$

其中,$f(x_i,r)$ 表示以 x_i 为中心、半径为 r 的球体内样本的个数,在 r 值确定的情况下,该值越大说明 x_i 处于样本集越密集区域,$r=\mathrm{var}*\mathrm{tp}$,且 $0<\mathrm{tp}\leqslant 1$,tp 为经验值。$\mathrm{bwd}(x_i)$ 等于所有球体内样本到非球体内其他样本距离之和,$\mathrm{wid}(x_i)$ 为球体内所有样本之间距离和,$\frac{\mathrm{bwd}(x_i)}{\mathrm{wid}(x_i)}$ 值越大表示同类相似性越大,不同类相似性越低。所以 $\mathrm{density}(x_i)$ 的值越大,x_i 作为初始聚类中心的几率越大。

定义 3:样本 x_i 的距离均值 $m(x_i)$:

$$m(x_i)=\frac{1}{n}\sum_{j=1}^{n}d(x_i,x_j) \tag{3}$$

定义 4:样本集的均方差 var:

$$\mathrm{var}=\frac{1}{n-1}\sum_{j=1}^{n}(d(x_i,x_j)-m(x_i))^2 \tag{4}$$

定义 5:样本集聚类误差平方和 E:

$$E=\sum_{i=1}^{k}\sum_{x\in W_j}|x-c_i|^2 \tag{5}$$

其中,c_i 是第 i 类的聚类中心,x 是非 i 类的样本。E 值越小,类内样本之间相似性越高,聚类效果越好。

2.2　优化 K-medoids 算法实现

输入:样本集 $X=\{x_1,x_2,\cdots,x_n\}$,类簇数 K;

输出:样本集的 K 个划分 $w_1,w_2,\cdots,w_k$,以及 K 个聚类中心 $c_1,c_2,\cdots,c_k$,其中 $w_1\cup w_2\cup\cdots\cup w_k=X$,$w_i\cap w_j=\varphi, i\neq j$。

算法实现步骤:

(1)根据定义 2 计算每个样本的密度,并对样本按照密度的从大到小排列。

(2)选择密度最大样本 x_{i1} 作为第一个初始聚类中心 c_1;选择密度最大的样本 x_{i2} 作为第二个初始聚类中心 c_2,且 x_{i2} 满足 $d(x_{i2},c_1) \leqslant \text{var}$;选择密度最大的样本 x_{i3} 作为第三个初始聚类中心 c_3,且 x_{i3} 满足 $d(x_{i3},c_1) < \frac{\text{var}}{2} \&\& d(x_{i3},c_2) < \frac{\text{var}}{2}$,依次类推,直到选出 k 个初始聚类中心。并根据定义 1 将剩余样本与距离最近的聚类中心划分为一类。

(3)计算每个样本的 $\frac{\text{bwd}(x_i)}{\text{wid}(x_i)}$,选择其中最小的样本最为新的聚类中心。

(4)重新将剩余样本与距离最近的聚类中心划分为一类。

(5)根据定义 5 计算聚类误差平方和,判断是否满足条件。如果不满足转到(3),如果满足转到(6)。

(6)输出样本集的 K 个划分,完成聚类。

2.3 算法分析

根据以上的算法步骤,在计算样本密度时,需计算样本之间的距离以及样本集的方差。样本集的规模为 n,样本之间的距离根据定义 1 计算时,时间复杂度为 $O(n^2)$,计算样本的方差时,距离已知,时间复杂度为 $O(n)$。更新聚类中心时,计算每个样本的类内和类间时间复杂度为 $O(n^2)$。因此文中算法的时间复杂度为 $O(n^2)+O(n)+O(n^2)$。

选择初始聚类中心时,文献[9]选择高密度区域距离固定的样本作为初始聚类中心。不管这个距离过大或过小,选择的初始聚类中心都不能反映样本集聚类中心的实际分布,如果距离过大可能导致选择离群点作为初始聚类中心,或无样本可作为初始聚类中心。文中通过不断动态调整后续待样本与已选择聚类中心之间的距离,选择满足一定距离且密度最高的样本作为初始聚类中心,保证初始聚类中心的选择始终在样本集的密集区域,且能够符合样本集的实际分布。

3 实验结果分析

为验证文中聚类算法在医疗数据上的有效性,选择 UCI 机器学习数据库[12]中三个乳腺癌样本集进行测试,并与 PAM 算法与 Park 算法进行对比。实验仿真环境:Windows7,64 位操作系统,intel CORE i5-4200,CPU1.6 GHz 2.3 GHz,内存 8 G;编程环境 matlab R2012a。

聚类结果的评价方法采用聚类误差平方和、聚类时间、聚类准确率和 Rand Index[13]。

3.1 医疗数据集分析

文中用于测试的乳腺癌数据集为 wdbc[14]、breast-cancer-wisconsin[15]和 Breast Cancer Coinmbra[16],其中 wdbc 和 breast-cancer-wisconsin 数据信息在 1995 年完成,Breast Cancer Coimbra 数据信息在 2018 年完成。关于这三个数据集的信息见表 1。

表 1 乳腺癌数据集

数据集	样本数	属性数	类数
wdbc	569	32	2
breast-cancer-wisconsin	699	11	2
Breast Cancer Coimbra	116	10	2

wdbc 数据集包含 569 个样本(实际的病例数据),分为 2 类,样本无缺失数据。数据集中每个样本包含 32 个属性,其中第 1 个属性为样本的编号,第 32 个属性为样本的真实类别,其余属性为乳腺细胞相关测量数据,如半径、纤维、光滑度等。

breast-cancer-wisconsin 数据集包含 699 个样本,分为 2 类。该数据集共有 11 个属性,第 1 个属性为样本编号,第 11 个属性为样本类别,其余属性为乳腺细胞测量数据。数据集中有 16 个样本缺失数据,文中按照数据挖掘缺失数据中取平均值方法,对缺失数据进行补充完整,保证数据集的样本数量。

Breast Cancer Coimbra 是 2018 年完成采集的一组乳腺癌数据集,包含 116 个样本,分为 2 类,无样本缺失数据值。该数据集每个样本包含 10 个属性,第 1 属性值为年龄,第 10 组为类别,其中 64 个样本为癌症,52 个为健康病例。

为验证文中算法在选择初始聚类中心的有效性,生成人工样本集。该样本集包含 75 个样本、每个样本含有 2 个属性,分为 3 类,且符合正态分布。生成样本集的参数如表 2 所示。

表 2 人工数据集生成参数

参数	第 1 类	第 2 类	第 3 类
均值	$\mu_x^1 = 3.8$	$\mu_x^2 = 2.3$	$\mu_x^3 = 1.6$
	$\mu_y^1 = 3.4$	$\mu_y^2 = 3.0$	$\mu_y^3 = 1.4$
方差		$\delta^1 = 0.1$	

3.2 实验结果分析

3.2.1 人工样本集结果分析

人工模式样本集的实际分布见图 1。为了更好地测试文中算法在选择初始聚类中心的有效性,对人工样本集中的 75 个样本按照从上到下的顺序按升序编号,处在第 1 个位置的样本编号为 1 号,第 75 个位置的样本编号为 75 号。样本集划分为三类,其中样本序号 1~30 为第一类,样本序号 31~55 为第二类,样本序号 56~75 为第三类。在图 1 中第一类用“+”表示;第二类用“●”表示;第三类用“○”表示。

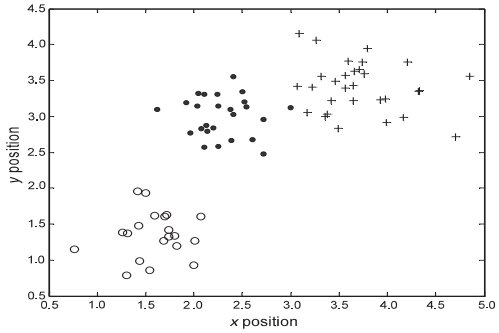

图 1　人工数据集分布

PAM 算法、Park 算法和文中算法在人工样本集中选择的初始聚类中心所在样本集中的序号见表 3。

表 3　不同算法选择的初始聚类中心序号

算法	初始聚类中心样本序号
PAM	聚类中心序号随机
Park	45　34　35
文中算法	55　4　66

由表 3 可知,在人工数据集中 PAM 算法随机选择初始聚类中心,因此聚类中心每一次都不确定;Park 算法选择的初始聚类中心序号分别为 45、34、35,这三个序号都属于第二类簇,所以 Park 算法选择密度最大的前 K 个作为初始聚类中心的样本位于同一个类簇中。Park 算法选择的前 K 个样本的具体分布见图 2。文中算法选择作为初始聚类中心的样本序号分别是 55、4、66,这三个序号的样本分别是第二类簇、第一类簇和第三类簇的样本,选择的初始聚类中心分布在不同类,且处于不同类密度较高的区域,符合样本集实际分布。文中算法选择的初始聚类中心的分布见图 3。

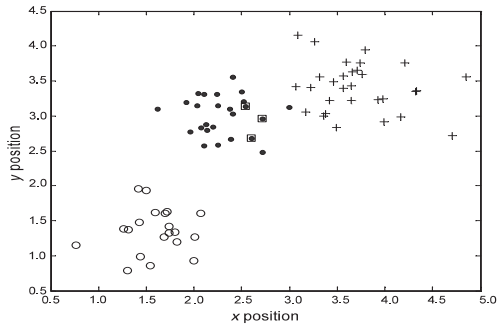

图 2　Park 算法选择的初始聚类中心

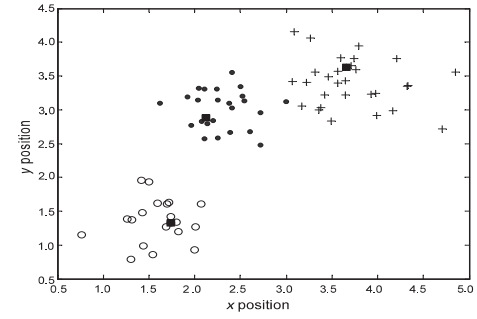

图 3　文中算法选择的初始聚类中心

3.2.2　乳腺癌数据集实验结果分析

PAM 算法、Park 算法和文中算法在三种乳腺癌样本集的聚类误差平方、聚类时间见表 4、表 5。为了方便表示,breast-cancer-wisconsin 在表中简写为 bcw,Breast Cancer Coimbra 简写为 BCC。

表 4　三种算法的聚类误差平方和

样本集	PAM	Park	文中算法
wdbc	3.143 3e+08	7.814 8e+07	7.476 5e+07
bcw	53 925.2	22 474	20 238
BCC	9 465 675	5.619 8e+06	3.395 7e+03

由表 4 可见,文中算法的聚类误差平方和均小于 PAM 算法和 Park 算法,因此文中算法的聚类结果同一类样本的相似度最高。Park 算法聚类误差平方和优于 PAM 算法,PAM 算法的聚类误差平方和最差。

表 5　三种算法的聚类时间

样本集	PAM	Park	文中算法
wdbc	396.050 6	0.1154	0.083 0
bcw	272.201 6	0.176 3	0.074 1
BCC	68.294 9	0.010 9	0.060 7

由表 5 可见,文中算法在 wdbc 和 breast-cancer-wisconsin 的聚类时间明显优于其他两种算法,但 Park 算法在 Breast Cancer Coimbra 样本集中的聚类时间优于文中算法,PAM 算法的聚类时间最差。

三种算法在乳腺癌数据集上的聚类准确率和 Rand Index 如图 4 和图 5 所示。

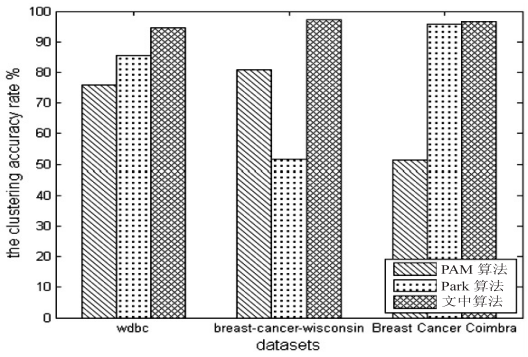

图 4　三种算法的聚类准确率

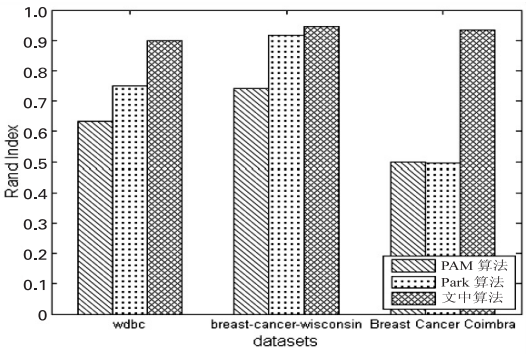

图 5　三种算法的 Rand Index 值

图 4 为三种聚类算法在不同乳腺癌样本集上的聚类准确率,结果显示 Park 算法的聚类准确率在 wdbc 和 Breast Cancer Coimbra 两个样本集上优于 PAM 算法,PAM 算法在 breast-cancer-wisconsin 中的聚类准确率优于 Park 算法,而文中算法的聚类准确率最优。图 5 为三种聚类算法在不同乳腺癌样本集上的 Rand Index 值,结果显示 PAM 算法的结果最差,Park 算法较优于 PAM 算法,而文中算法的指标最优。

4 结束语

基于医疗数据,针对 K-medoids 算法存在的不足进行优化,优化后的 K-medoids 算法通过实验验证,选取的初始聚类中心更符合样本的实际分布,聚类误差平方和、聚类准确率等指标均优于其他两种算法。该算法虽然在乳腺癌数据集中具有较好的聚类效果,但随着医疗大数据的产生,将优化的聚类算法应用于医疗大数据也是未来研究的方向。

参考文献:

[1] 唐 涛,覃 晓,易宗剑,等. 基于 k 中心点聚类的图像二值化方法[J]. 计算机科学与探索,2015,9(2):234-241.

[2] ARORA P, DEEPALI D, VARSHNEY S. Analysis of K-means and K-medoids algorithm for big data[J]. Procedia Computer Science,2016,78:507-512.

[3] KHATAMI A, MIRGHASEMI S, KHOSRAVI A, et al. A new PSO-based approach to fire flame detection using K-Medoids clustering[J]. Expert Systems with Applications, 2017,68:69-80.

[4] 李晓雪,郑静晨,李 明,等. 基于医疗数据的属性约简聚类分析算法[J]. 医学信息学杂志,2016,37(4):59-63.

[5] 黄 辰,潘永才,李可维,等. 基于传感器聚类数据挖掘的物联网智慧医疗模型设计[J]. 传感器与微系统,2014,33(4):76-79.

[6] HAN J, KAMBER M. Data mining, Southeast Asia edition: concepts and techniques[M]. San Francisco, CA, USA: Morgan Kaufmann,2006:383-464.

[7] 孙吉贵,刘 杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):48-61.

[8] 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.

[9] 谢娟英,高 瑞. 方差优化初始中心的 K-medoids 聚类算法[J]. 计算机科学与探索,2015,9(8):973-984.

[10] 颜宏文,周雅梅,潘 楚. 基于宽度优先搜索的 K-medoids 聚类算法[J]. 计算机应用,2015,35(5):1302-1305.

[11] HAN J, KAMBER M, PEI J. Date mining: concepts and techniques[M]. Beijing: China Machine Press,2012:293-297.

[12] FRANK A, ASUNCION A. UC irvine machine learning repository[EB/OL]. 2013-01-18. http://archive.ics.uci.edu/ml.

[13] RAND W M. Objective criteria for the evaluation of clustering methods[J]. Jouranl of the American Statistical Association,1971,66(336):846-850.

[14] OSAREH A, SHADGAR B. Machine learning techniques to diagnose breast cancer[C]//Proceedings of the 2010 5th international symposium on health informatics and bioinformatics. Antalya, Turkey: IEEE,2010:114-121.

[15] BENNETT K, MANGASARIAN O L. Robust linear programming discrimination of two linearly inseparable sets[J]. Optimization Methods and Software,1992,1:23-34.

[16] PATRÍCIO M, PEREIRA J, CRISÓSTOMO J, et al. Using Resistin, glucose, age and BMI to predict the presence of breast cancer[J]. BMC Cancer,2018,18(1):123-130.