IOP Conf. Series: Journal of Physics: Conf. Series **1487** (2020) 012011 doi:10.1088/1742-6596/1487/1/012011

Optimization of K-medoids Algorithm for Initial Clustering Center

Wang Yan E[1,*],An Jian[2], Liang Yan[1] and Wang HongGang[1]

[1] School of Technology, Xi'an Si Yuan University, Xi'an Shannxi 710038 China
[2] Shenzhen Research Institute of Xi'an Jiao tong University, Shenzhen Guangdong 518057 China

[*]Corresponding author: wye0712@snnu.edu.cn

Abstract,This paper studies the k-medoids of the partitioning clustering algorithm. A variance-based density optimization algorithm is proposed to solve the problem of random selection of initial clustering centers, 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. Under the same density radius, the samples in dense regions have high density. By dynamically selecting the samples are selected as initial clustering centers from different dense regions, in the clustering process local optimization is used to accelerate the convergence speed. These operations solve the shortcomings of K-medoids algorithm. In order to test the clustering effect, this algorithm is applied to data set of UCI machine learning. The experimental results show 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 data sets.

1. Introduction

Clustering is one of the effective means to analyze data in data mining. The advantage of clustering is that no prior knowledge of the data is needed. According to the internal characteristics of the data, the algorithm will analyze the data and use a certain correlation measurement method to obtain the hidden relationship of the datasets. As an effective method to analyze data, clustering has been widely used in many fields such as image processing[1], big data[2] and artificial intelligence[3].Now, With the rapid development of medical treatment, medical data is more and more huge. Medical images and pathological records are stored on the computer in the form of data. In order to effectively assist medical staff to predict and diagnose diseases, clustering algorithm is increasingly applied in medical data[4-5].Common clustering algorithms include partitioning based, hierarchy based, density based, model based and mesh based clustering[6].

K-medoids algorithm is one of the classical partition clustering algorithms, which is easy to implement because of its simple principle. Although the traditional k-medoids algorithm is not sensitive to noise data, it randomly selects the initial clustering center, resulting in unstable clustering results. Therefore, many scholars have optimized k-medoids algorithm to improve its clustering accuracy and stability. The classical representative of the traditional k-medoids algorithm is PAM algorithm[7]. PAM algorithm randomly selects the initial clustering center, and updates the clustering center through global search, which makes the clustering time consuming too much. Optimization PAM algorithm is the

classic algorithms of fast algorithm (Park) K-medoids algorithm[8], this algorithm to select the smallest distance samples as the initial clustering center, overcoming the random uncertainty caused by the clustering center, update the initial clustering center at the same time using iterating within class, greatly reducing the time of clustering, and the Park algorithm to choose the clustering center is in the same class, often does not conform to the actual distribution of the sample set. In order to overcome the disadvantages of Park algorithm, the selection of initial clustering center is still one of the hot topics in Park algorithm research.

2. PAM algorithm and Park algorithm

The two kinds of algorithm evaluation goal is to make the similar samples as similar as possible, different kinds of samples as different as possible. The clustering error sum of squares[9] as the evaluation index of clustering results is smaller , the higher the similarity of the same variety, the lower the similarity is not the same, the clustering result is in line with the actual distribution of samples, the better clustering effect.

3. Improved k-medoids algorithm

Let the sample set to be clustered be X, $X=\{x1,x2,\cdots,xn\}$,Where x_i is a sample containing p-dimensional data. The sample set is divided into K classes and the central set of clustering is C , $C=\{c_1,c_2,\cdots,c_k\}$,where c_i represents the cluster center of class *i.* Algorithm in this paper to measure the distance between sample similarity, with clustering error sum of squares as the objective function, the highest density of dynamic selection sample K samples and in different regions as the initial clustering center, used in the process of updating the clustering center in the class and class's foreign minister, in the process of clustering optimization clustering center choice, reduce the clustering iterations, speed up the clustering process.

3.1. Related Conception

The Euclidean distance of Samples x_i, x_j is

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

The density of sample x_i is

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

Where, $f(x_i,r)$ represents the number of samples in a sphere with radius r centered on x_i . Under the condition that r value is determined, the larger the value is, the denser x_i is in the sample set, $r=var*tp$, and $0\prec tp\le 1$, tp is the empirical value. $bwd(x_i)$ is equal to the sum of the distance from all the samples in the sphere to the other samples in the non-sphere, $wid(x_i)$ is the distance between all the samples in the sphere and, the larger the $\frac{bwd(x_i)}{wid(x_i)}$ value is, the greater the similarity between the samples in the sphere and the lower the similarity between the samples in different spheres. Therefore, the greater the value of $density(x_i)$, the greater the probability that x will be the initial cluster center.

The mean of the distances to the sample x_i is

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

The mean variance of the sample set is var :

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

Sample aggregate sum of squared errors is E :

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

Where c_i is the center of the class I, x is't the sample of class i. the smaller the value of E ,the higher the similarity between samples in the class, and the better the clustering effect.

3.2 Optimize the Implementation of K-medoids Algorithm.

Input,The sample set: $X = \{x_1, x_2, \cdots, x_n\}$,class numbers of clustering: K

Output, K partition of the sample set : $w_1, w_2, \cdots, w_k$, K the clustering centes: $c_1, c_2, \cdots, c_k$ .Where $w_1 \cup w_2 \cup \cdots \cup w_k = X$, $w_i \cap w_j = \phi, i \neq j$.

Algorithm implementation steps,

（1）Calculate the density of each sample according to definition 2, and arrange the samples according to the density from large to small.

（2）Select the sample x_{i1} with the highest density as the first initial clustering center c_1 ;The sample x_{i2} with the highest density was selected as the second initial clustering center c_2 ,and x_{i2} met $d(x_{i2}, c_1) \leq var$ .Select x_{i3} , the sample with the highest density, as the third initial clustering center c_3 , and x_{i3} satisfies $d(x_{i3}, c_1) < \frac{var}{2} \&\& d(x_{i3}, c_2) < \frac{var}{2}$, and so on, until k initial clustering centers are selected. According to definition 1, the remaining samples and the nearest cluster center are classified into one class.

（3）Calculate the $\frac{bwd(x_i)}{wid(x_i)}$ of each sample and select the smallest sample as the newest cluster center.

（4）Divide the remaining samples and the nearest clustering center into one class.

（5）According to definition 5, calculate the sum of squares of clustering errors and judge whether the condition is satisfied. If it's not satisfied it goes to step (3), if it's satisfied it goes to step (6).

（6）Output K partition of sample set, complete clustering.

3.3 Algorithm analysis

According to the above algorithm steps, when calculating the sample density, the distance between the samples and the variance of the sample set should be calculated. The size of the sample set is n, and the distance between the samples is $O(n^2)$ according to definition 1. When calculating the variance of the samples, the distance is known and the time complexity is O(n).When updating the clustering center, the time complexity within and between classes of each sample was calculated as $O(n^2)$.Therefore, the time complexity of this paper is $O(n^2)+ O(n)+ O(n^2)$.

4. Analysis of experimental results

In order to verify the effectiveness of the clustering algorithm in this paper on medical data, three breast cancer sample sets in UCI machine learning database[10] were selected for testing, and compared with PAM algorithm and Park algorithm. Experimental simulation environment: windows7, 64-bit operating system, Intel CORE i5-4200, cpu1.6ghz, 2.3ghz, 8G memory; Programming environment matlab R2012a.

The evaluation methods of the clustering results in this paper are the sum of squared clustering errors, clustering time, clustering accuracy and Rand Index[11].

4.1 Medical data set analysis

In order to test this paper algorithm, we use breast cancer datasets including Wdbc[12] ,breast-cancer–Wisconsin[13] and breast cancer Coinmbra[14]. WDBC and breast-cancer-Wisconsin data information were completed in 1995, breast cancer Coimbra data information were completed in 2018. Information about these three data sets is shown in Table 1.

Table 1. breast cancer datasets

datasets	number	attributes	k
wdbc	569	32	2
breast-cancer-wisconsin	699	11	2
Breast Cancer Coimbra	116	10	2

In order to verify the validity of the proposed algorithm in selecting the initial clustering center, artificial sample sets were generated. The sample set contains 75 samples, each of which contains 2 attributes, divided into 3 categories, and conforms to the positive distribution. The parameters to generate the sample set are shown in Table 2.

Table 2. Manual data set generation parameters

parameters	1	2	3
Mean value	$\mu_x^1 = 3.8$	$\mu_x^2 = 2.3$	$\mu_x^3 = 1.6$
	$\mu_y^1 = 3.4$	$\mu_y^2 = 3$	$\mu_y^3 = 1.4$
variance	$\delta^1 = 0.1$		

4.2 Analysis of experimental results

4.2.1 Artificial sample set result analysis. The actual distribution of the sample set of manual mode is shown in Figure 1. In order to better test the effectiveness of the algorithm in this paper, 75 samples in the artificial sample set were numbered in ascending order from top to bottom. The number of samples in the first position was no. 1, and the number of samples in the 75th position was no. 75.The sample set is divided into three categories, among which the sample serial number 1-30 is the first category, the sample serial number 31-55 is the second category, and the sample serial number 56-75 is the third category. In Figure 1, the first category is represented by "+";The second kind is expressed by "●";The third type is denoted by "○".

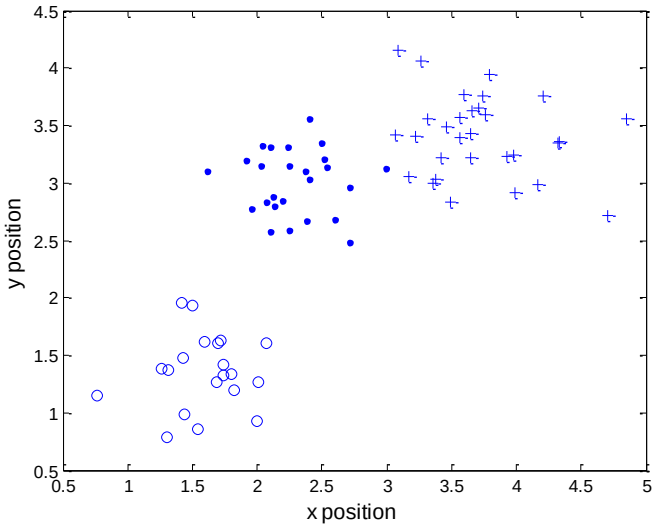

Figure 1 distribution of artificial data sets

The sequence Numbers of PAM algorithm, Park algorithm and the algorithm in this paper in the sample set where the initial clustering center is selected in the artificial sample set are shown in Table 3 below.

Table 3. the initial cluster center number

algorithm	The sample number of the initial clustering centers		
PAM	The serial number of random		
Park	45	34	35
The paper	55	4	66

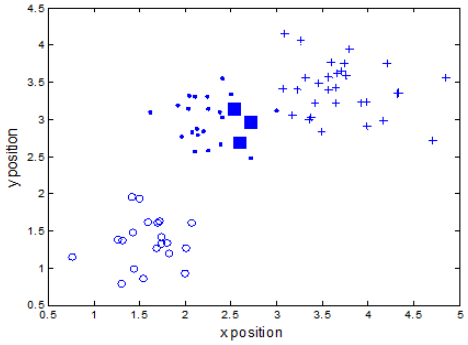

Figure 2. the initial cluster centers of Park algorithm

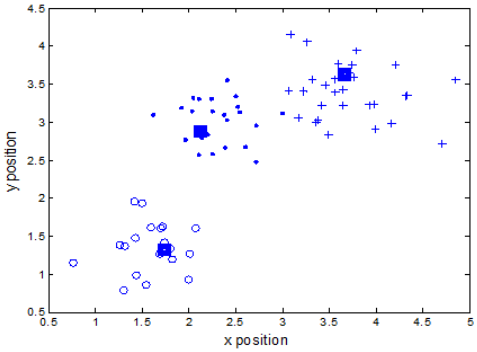

Figure 3. the initial cluster centers of this paper

Table 3 shows that the initial clustering center is randomly selected by PAM algorithm in the manual data set, so the clustering center is uncertain every time. The sequence Numbers of the initial clustering centers selected by Park algorithm are 45, 34 and 35, respectively. These three serial Numbers belong to the second type of cluster, so the first K ones with the highest density are selected by Park algorithm as the samples of the initial clustering center in the same class cluster. The specific distribution of the first K samples selected by Park algorithm is shown in Figure 2. The samples with the serial Numbers 55, 4 and 66 as the initial clustering centers are selected by the algorithm in this paper. The samples with these three serial Numbers are the samples of the second class cluster, the first class cluster and the third class cluster, respectively. The selected initial clustering centers are distributed in different classes and are in the regions with higher density of different classes, which conforms to the actual distribution of the sample set. The initial clustering center distribution selected by the algorithm in this paper is shown in Figure 3.

4.2.2 Analysis of breast cancer data set trial results. The clustering error square and clustering time of PAM algorithm, Park algorithm and the algorithm in this paper are shown in Table 4 and Table 5. For convenience, breast-cancer-wisconsin is abbreviated to BCW, and breast cancer Coimbra is abbreviated to BCC.

Table 4. The sum of squared clustering errors

datasets	PAM	Park	The paper
wdbc	3.1433e+08	7.8148e+07	7.4765e+07
bcw	53925.2	22474	20238
BCC	9465675	5.6198e+06	3.3957e+03

In Table 4, the sum of squares of clustering errors of the algorithm in this paper is smaller than that of PAM algorithm and Park algorithm. Therefore, the clustering results in this paper have the highest similarity of samples of the same class. The sum of squares of clustering errors of Park algorithm is better than PAM algorithm, and PAM algorithm has the worst sum of squares of clustering errors.

The algorithm with better clustering time in bold part in Table 5, the clustering time of the proposed

algorithm in WDBC and breast-cancer-wisconsin is obviously better than that of the other two algorithms, but the clustering time of Park algorithm in the breast cancer Coimbra sample set is better than that of the proposed algorithm, and the clustering time of PAM algorithm is the worst.

Table 5. Clustering time (ms)

datasets	PAM	Park	This paper
wdbc	396.0506	0.1154	**0.0830**
bcw	272.2016	0.1763	**0.0741**
BCC	68.2949	**0.0109**	0.0607

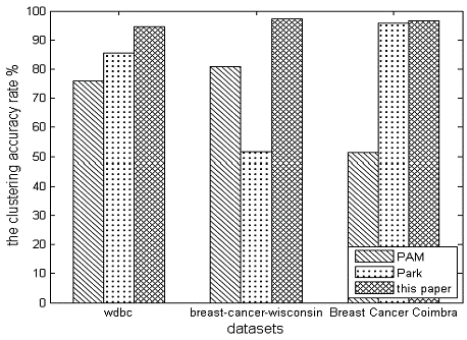

Figure 4. The clustering accuracy

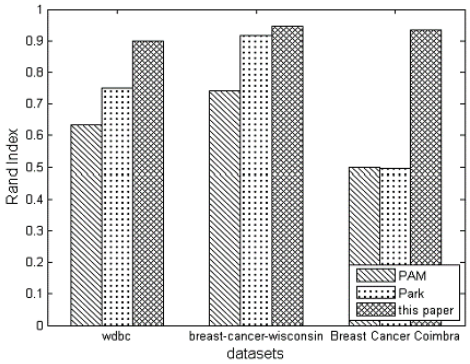

Figure 5. The Rand Index

The clustering accuracy is shown in Figure 4, Rand Index of the three algorithms is Figure 5. The bar filling lines in Figure 4 and Figure 5 are slash, dot and cross, respectively, representing PAM algorithm, Park algorithm and the algorithm in this paper.

Figure 4 shows the clustering accuracy of three clustering algorithms on different Breast Cancer sample sets. The results in Figure 4 show that the clustering accuracy of Park algorithm is better than that of PAM algorithm in WDBC and Breast Cancer Coimbra sample sets. The clustering accuracy of PAM algorithm in breast-cancer-wisconsin is better than that of Park algorithm, but the clustering accuracy of this algorithm is the best. Figure 5 shows the Rand Index values of three clustering algorithms The results show that PAM algorithm has the worst results, Park algorithm is better than PAM algorithm, but the Index of text algorithm is the best.

5. Summary

The optimized k-medoids algorithm was verified by experiments that the selected initial clustering centers were more consistent with the actual distribution of samples, the clustering error sum of squares and the clustering accuracy were better than the other two algorithms. Although the algorithm in this paper has a good clustering effect in the breast cancer data set, with the generation of medical big data, applying the optimized clustering algorithm to medical big data is also the future research direction of this paper.

Acknowledgments

We would like to thank Xi'an University for the financial support of this project (18JK1100 project, XGH19236 project) and Xi'an Jiao tong University for the financial support of this research (JCYJ20170816100939373 project).

References:

[1] TANG Tao, QIN Xiao, YI Zongjian, HAN Dongyue. Image Binarization Processing Method Using k-Medoids Clustering[J].Journal of Frontiers of Computer Science and Technology, 2015,9(2):234-242.

[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.[doi:10.1016/j.procs.2016.02.095].

[3] Khatami, Amin,Mirghasemi, Saeed,Khosravi. A new PSO-based approach to fire flame detection using K-Medoids clustering[J]. Expert Systems with Applications, 2017,68(1):69-80.
[4] Li xiaoxue, zheng jingchen, li Ming, hao yuwen. Attribute reduction cluster analysis algorithm based on medical data [J], journal of medical informatics, 2016, 37(4):59-63
[5] Huang Chen, pan yongcai, li kewei. Design of intelligent medical model of Internet of things based on sensor cluster data mining [J]. Sensor and microsystem,2014,33 (4):76-80
[6] Han Jiawei, Kamber M. Data mining, Southeast Asia edition:concepts and techniques[M]. San Francisco, CA, USA: Morgan kaufmann, 2006: 383-464.
[7] Sun Jigui, Liu Jie, Zhao Lianyu. Clustering alogrithm research[J]. Journal of Software, 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] HAN J,KAMBER M,PEI J．Date mining: concepts and techniques［M］. FAN M,translated．Beijing: China Machine Press,2012: 293-297.(HAN J, KAMBE R M, PEI J.
[10] Frank A, Asuncion A. UC Irvine Machine Learning Repository [Online], available: http://archive.ics.uci.edu/ml, January 18, 2013.
[11] Rand W M. Objective criteria for the evaluation of clustering methods[J] Jouranl of the American Statistical Association 1971,66(336):846-85
[12] Osareh A,Shadgar B.Machine learning techniques to diagnose breast cancer [C].//Proceedings of the 2010 5th International Symposium on Health Informatics and Bioinformatics. Piscataway NJ:IEEE,2010:114-121.
[13] Bennett,K. and Mangasarian,O.L. Robust linear programming discrimination of two linearly inseparable sets[J].Optim. Meth.Software,1992,1,23–34.
[14] Miguel Patrício, José Pereira, Joana Crisóstomo, Paulo Matafome, Manuel Gomes. Using Resistin, glucose, age and BMI to predict the presence of breast cancer[J], BMC Cancer 2018,18(1):123-130