Decreasing Iteration Number of K-medoids Algorithm with IFART

Sevinc Ilhan Omurca[1], Nevcihan Duru[2]

[1] Computer Engineering Department, Kocaeli University, Umuttepe Campus, Turkey
silhan@kocaeli.edu.tr
[2] Computer Engineering Department, Kocaeli University, Umuttepe Campus, Turkey
nduru@kocaeli.edu.tr

Abstract

K-medoids is a well known and widely used algorithm in data clustering. Performance of the algorithm depends on the initialization of cluster centers as in other centered based clustering techniques. In this article we used an initialization method based on fuzzy art to initialize clusters. The algorithm has been applied to different datasets. The experiments show that our approach can achieve higher or comparable performance when it is compared with conventional k-medoids.

1. Introduction

Clustering is used in many fields, such as data mining, pattern recognition, image segmentation, machine learning, and outlier detection. Data clustering tries to group the objects that the objects within the same cluster have a high degree of similarity, while data objects belonging to remaining clusters have a high degree of dissimilarity.

Typical clustering methods in the literature are partition based, hierarchical, density based and grid based clustering. Partitioning methods partitions the sample dataset into k clusters where each cluster represented by a data object which is a centroid or a cluster representative in some way. K-means [1] and k-medoids [2] are the most well known and widely used techniques for partition based clustering. They both starts by initializing the k cluster centers and assigning every data object to the nearest center then iteratively go on by finding the new k centers. By iterative partitioning they attempt to minimize the squared error.

K-means algorithm is efficient at clustering huge datasets in terms of computational time but it is sensitive to the outliers. K-medoids is less sensitive to outliers in contrast to the k-means. But k-medoids is not suitable for huge datasets due to its time complexity. Partitioning Around Medoids (PAM) [2], Clustering LARge Applications (CLARA) [2] and Clustering Large Applications based on RANdomized Search (CLARANS) [3] are three popular algorithms of k-medoids. PAM as the most powerful of them has the advantage of robustness to noisy data or outliers but at the same time it has the disadvantage of time complexity [3].

The main issue of the study is reduce computational time of the algorithm with the IFART (Improved Fuzzy Art) initialization method.

The remaining parts of this paper are organized as follows: section 2 and 3 provides brief background knowledge about k-medoids and initialization method respectively. The performance comparison and the clustering error rates are presented for three real datasets and four synthetic datasets in section 4. The final section discusses the findings and concludes with a summary of this study.

2. K-medoids

K-medoids algorithm is extends k-means algorithm, it requires the number of clusters and the initial medoids as inputs. K-medoid algorithm is similar to k-means except the cluster centers or medoids have to belong to the dataset being clustered.

A medoid is the member of a cluster whose average dissimilarity to all members is minimal within a cluster. The k-medoid method is a representative object based technique. The medoids are representing objects of the clusters as cluster centers. They have to be one of the real objects in the data set.

The algorithm works briefly as follows. A random representative data object selected for each cluster and remaining objects are attached to the nearest cluster considering their distance or similarity degree to the representative data object of that cluster. In order to improve the quality of the cluster representative object is iteratively replaced with the other nonrepresentative objects. The clustering quality is estimated by an objective function [4]. The objective function of k-medoids which is also an absolute-error criterion is as follows.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, c_i)^2 \qquad (1)$$

Here k is the cluster number; c_i is the medoid of cluster i; $d(p, c_i)$ is the distance between the data object (p) and the medoid of cluster i; E is the sum of the absolute error for all objects in the dataset.

The k-medoid algorithm is as follows.

Input:

k: the number of clusters,
D: the dataset containing objects.

Output:

A set of k clusters.

Algorithm:

Step1. Arbitrarily choose k objects in D as the initial representative objects or seeds;

Step2. For iteration 1 to max iteration {

Step3. Assign each remaining object to the cluster with the nearest representative object

Step4. Randomly select a nonrepresentative object, o_{random}

Step5. Compute the total cost (S) of swapping representative object O_j with nonrepresentative O_{random}

Step6. If S<0 then swap O_j with O_{random} to form the new set of k representative objects

Step7. Save the total cost as the iteration error} [6]

The k-medoids method is more robust than k-means in the presence of noise and outliers, because a medoid is less influenced by outliers or other extreme values than a mean. However, its processing is more costly than the k-means method [6]. The performance of the k-medoids depends on the choice of the initial representative objects.

3. Initialization Method

Fuzzy Art is an unsupervised category learning technique and it was introduced by Carpenter et al. [7]. It's capable of rapid stable clustering of continuous input patterns and very effective for large scaled data clustering. It requires less processing time considering the other clustering algorithms.

In this paper, an Improved Fuzzy Art [8, 9] algorithm is used as an initialization method for k-medoids. When the clustering results obtained by Fuzzy Art are examined, it is realized that valid and effective clustering cannot be achieved and in some cases clusters were even overlapping. A better quality clusters are achieved by moving the data object to the right clusters according to maximum membership degrees.

The steps of IFART carried out as follows. Membership degree of each data object to each cluster formed with Fuzzy Art is calculated based on the cluster centers. The calculated values are stored in a membership matrix in which the rows represent the input data and the columns represent the clusters created. For each cluster, the data object with maximum membership degree is defined as its cluster center.

The membership matrix is a data structure that summarizes the membership degrees of the data objects to the formed clusters. According to these membership degrees, the cluster representatives are determined by different approaches. In this study the data objects with maximum membership degree are defined as cluster representatives for clusters and they constitute the k initial points for k-medoids algorithm

In this initialization scheme, because of IFART is based on Fuzzy Art, it requires less processing time and the clusters obtained by IFART are more accurate than the clusters obtained Fuzzy Art.

4. Experimental Results

In this study, the IFART initialization scheme and the classic random initialization scheme are tested on 3 real and 4 synthetic datasets. Real datasets: Iris, Pima Indian Diabetes (PInd) and Yeast are taken from the UCI Machine Learning Repository [10]. The synthetic datasets: Web Logs (WL), Documents_Sim (DS), Mars, Image Extraction (IE) and a Synthetic dataset with 5000 rows (S1-5000) are taken from the databases prepared by Pei and Zaiane [11] at Canada's Alberta University, Department of Computer Sciences.

Randomly initialized k-medoid algorithm is executed for 80 times by different initial points to consider the clustering error with different initializations. In each execution, the algorithm is also runs for specified maximum iteration times (100) iterations. Squared Euclidean distance measure is used for the experiments. The minimum, maximum and average clustering errors of these executions are shown in Table 1.

The proposed method is also executed on the same datasets by the initial cluster representatives determined by IFART. Unlike the conventional method, in this method, algorithm does not have to execute 80 times due to the initial points are determined. Beyond that, the maximum iteration time is not determined as 100 iterations for k-medoids, algorithm runs for only 1 iteration. The clustering error of the proposed method is also shown in Table 1.

Table 1. Clustering Errors

Dataset	Randomly Initialized k-medoids			IFART initialized k-medoids
	Min	Max	Avg.	
WL	0.0439	0.2015	0.0790	0.0613
Mars	0.0076	0.0376	0.0222	0.0190
IE	0.0474	0.3292	0.0849	0.0695
S1-5000	0.0700	0.1786	0.1182	0.1024
Iris	0.0373	0.1641	0.0675	0.0478
PInd	0.1774	2.3379	0.5361	0.3385
Yeast	0.0061	0.0652	0.0199	0.0152

The minimum, maximum and average clustering errors of the conventional k-medoids are shown in Table 1 with columns 2, 3 and 4 respectively. K-medoids clusters the WL dataset with an error rate between [0.0439-0.2015]. The average clustering error rate due to this dataset is determined as 0.0790 for 80 different executions. It can cluster the datasets with small or high error margin according to selected initial medoids.

The last column of the table shows the clustering error belonging to the k-medoids algorithm initialized with IFART. As seen from the error estimations in Table 1, the proposed k-medoids method performs clustering with a smaller error margin in all datasets compared to the randomly initialized k-medoids.

As noted previously, the k-medoids algorithm works well for small datasets but not for large datasets that it is computationally expensive. With conventional scheme, algorithm runs for several iterations attempting to minimize the squared error, however with the proposed initialization scheme, algorithm runs for only an iteration attempting to minimize the squared error. Thus the algorithm has become faster and applicable for large datasets. To give an example, for iris dataset, randomly initialized k-medoids requires 0.0414 running time (in seconds) to complete 100 iterations and so 3.312 running time to complete 80 different executions; k-medoids algorithm initialized with IFART requires 0.0035 running time for only one iteration; the IFART algorithm requires 0.0194 running time.

Another contribution extracted from Table 1 is, the clustering error rate of the proposed method is better than average clustering error rate of the conventional k-medoids and also much better than the maximum clustering error rate which is also one of the probable clustering schemes.

With the proposed method not only the clustering error rate but also the maximum iteration number of the algorithm is decreased. While the proposed method executed only one time for only one iteration, the conventional algorithm have to be executed for several times for several iterations.

5. Conclusions

Although the k-medoids method is robust in the presence of noise and outliers its processing time is costly. Besides, the performance of the k-medoids depends on the choice of the initial representative objects. So making accurate selection of initial points improves the performance and makes the algorithm more scalable and robust.

In the conventional k-medoids method, initial points are selected randomly among the dataset objects. Due to the clustering results are changed according to random initial points, the algorithm have to be execute for many times and the clustering results have to be analyzed at the end of every execution time to find the optimum clustering. Instead of executing only one time, the algorithm is executed in many times. This is a critical waste of time. As long as the dataset size increases the wasting time becomes more critical. The main purpose of this paper is to select the accurate initial centers for k-medoids algorithm. Consequently the IFART initialization method is an effective solution for k-medoids algorithm as demonstrated by the experiments.

In IFART method, first the input data clustered with Fuzzy Art, than the membership degree to the clusters are calculated. The data objects with maximum membership degrees are determined as the initial medoids of each cluster. K-medoid initialized with IFART and conventional k-medoid executed with real and synthetic datasets.

While conventional k-medoid runs 80 times for 100 iterations in each, k-medoid initialized by IFART runs only for 1 iteration. Conventional algorithm is able to perform the clustering with a very low or very high error depending on the initial centroids selected. The proposed k-medoids method performs more stable and robust clustering as the initial centroids are predetermined. It is executed only for one time and at this execution the maximum iteration number of the algorithm is 1. In this case, it is completed in far fewer steps compared to the conventional k-medoids. Furthermore the proposed method has a lower clustering error rate according to the average error of the conventional k-medoids.

As a conclusion, the proposed initialization method makes the k-medoids a more stable and faster algorithm runs only for 1 iteration with reasonable error margins.

7. References

[1] J.B. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations”, Proc. Symp. Math. Statist. and Probability (5th), 1967, pp. 281–297.

[2] L. Kaufman, P. Russeeuw, “Finding groups in data: an introduction to cluster analysis”, John Wiley and Sons, 1990.

[3] R. Ng, J. Han, “Efficient and effective clustering methods for spatial data mining”, in:J.B. Bocca, M. Jarke, C. Zaniolo (Eds.), Proceedings of the Twentieth International Conference on Very Large Data Bases, Morgan Kaufmann, Santiago, Chile, pp. 144–155, 1994.

[4] X. Zhang, J. Wang, F. Wu, Z. Fan, X. Li, “A Novel Spatial Clustering with Obstacles Constraints Based on Genetic Algorithms and k-Medoids”, Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications, 2006.

[5] H-S. Park, C-H. Jun, “A simple and fast algorithm for K-medoids clustering”, Expert Systems with Applications, 36, pp. 3336–3341,2009.

[6] J. Han, M. Kamber, “Data Mining: Concepts and Techniques”, 2nd ed., J Gray, Ed. Morgan Kauffmann, 2006.

[7] G.A. Carpenter, S. Grossberg and D.B. Rosen, “Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System” Neural Networks, Vol. 4, pp. 759-771, 1991.

[8] S.Ilhan, N. Duru, E.Adali “Improved Fuzzy Art Method for Initializing K-means”, International Journal of Computational Intelligence Systems, Vol.3, No. 3, pp. 274-279, 2010.

[9] S. Ilhan, N. Duru, “An Improved Method for Fuzzy Clustering”, IEEE Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, Famagusta, North Cyprus, 2009.

[10] C. J. Merz, P. Murphy and D. Aha, UCI repository of machine learning databases, 1996.

[11] Y. Pei and O. Zaiane, A synthetic data generator for clustering and outlier analysis, Department of Computing Science, University of Alberta, Edmonton, AB, Canada, 2006.