The Korean Journal of Applied Statistics (2010)
23(2), 367–374

Sample Based Algorithm for k-Spatial Medians Clustering

Seohoon Jin[1] · Byoung Cheol Jung[2]

[1]Department of Informational Statistics, Korea University
[2]Department of Statistics, University of Seoul

(Received January 2010; accepted March 2010)

Abstract

As an alternative to the k-means clustering the k-spatial medians clustering has many good points because of advantages of spatial median. However, it has not been used a lot since it needs heavy computation. If the number of objects and the number of variables are large the computation time problem is getting serious. In this study we propose fast algorithm for the k-spatial medians clustering. Practical applicability of the algorithm is shown with some numerical studies.

Keywords: Cluster analysis, partitioning method, k-spatial medians clustering.

1. Introduction

Cluster analysis is a branch of statistics that, in the past three decades, has been intensely studied and successively applied to many applications. There are two main categories of clustering algorithms: hierarchical method and partitioning method. The most widely used partitioning method is the k-means clustering which minimizes within-cluster sum of squares. However, the k-means clustering is sensitive to outliers because an object with an extremely large value may substantially distort the distribution of data. This effect is particularly exacerbated due to the use of the square error function. To avoid the influence of outliers, actual objects can be picked to represent the clusters instead of taking the mean value of the objects in a cluster as a reference point. The k-mediods method, which uses actual objects in a cluster as a reference point, is performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding reference point (Kaufman and Rousseeuw, 1990). However, the medoid is also not fully free from effect of outliers.

The k-spatial medians clustering was proposed as an alternative to the k-means clustering in some previous studies such as Butler (1986), Butler (1988), Jhun (1986) and Jhun and Jin (2000). For verification of superiority of the k-spatial medians clustering let us consider an example. Suppose

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund)(KRF-2007-314-C00039).
[2]Corresponding author: Professor, Department of Statistics, University of Seoul, 90 Junnong-Dong, Dongdaemun-Gu, Seoul 130-743, Korea. E-mail: bcjung@uos.ac.kr

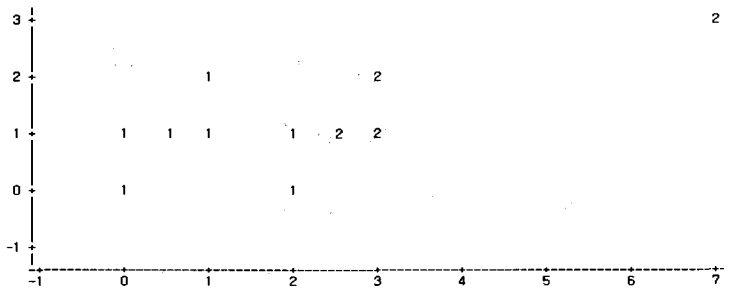

Figure 1.1. Results of k-means clustering

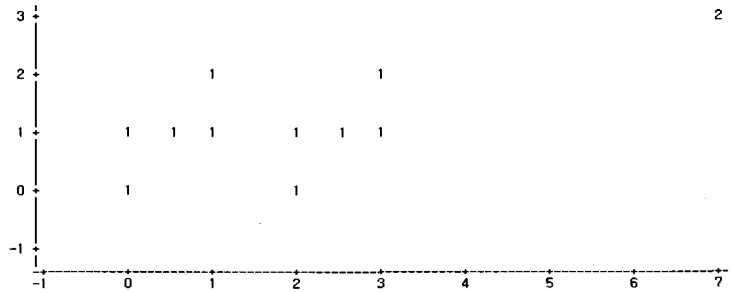

Figure 1.2. Results of k-medoids clustering

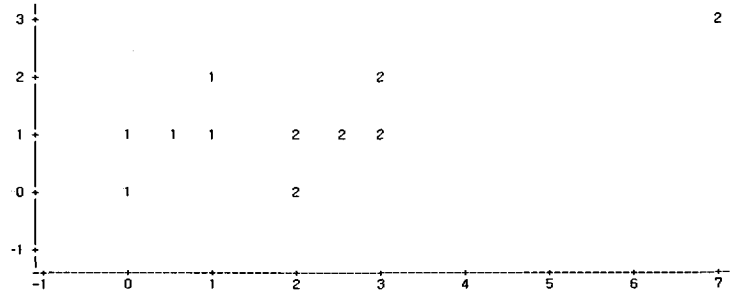

Figure 1.3. Results of k-spatial medians clustering

we measure two variables for each of 11 objects. The data consist of two the same structure clusters and one outlier(the point (7,3)). Figure 1.1, 1.2 and 1.3 show the results of k-means, k-medoids and k-spatial medians clustering sequentially. Each point is plotted by its cluster identification number. For the k-means clustering, the outlier perturbs the genuine cluster structure. The k-medoids clustering also doesn't overcome the outlier's influence. However, the k-spatial medians clustering gives the most preferable result under existence of an outlier. Since the spatial median is less sensitive to outliers, the centroid of each cluster is affected little by the outlier.

Even though the k-spatial medians clustering clearly has advantages under the situation of being outliers and special cluster structures, it has not been used a lot because of its computational difficulties. This study is about fast algorithm of the k-spatial medians clustering. We tried to improve applicability of the k-spatial medians clustering by modification of the clustering algorithm.

In Section 2, the k-spatial medians clustering was briefly reviewed with description of the k-spatial medians clustering algorithm. We have proposed a modified algorithm of the k-spatial medians clustering in Section 3. Practical applicability of the proposed algorithm has been shown with examples in Section 4. We also have applied the algorithm to real data case and examined its performance in Section 5.

2. k-Spatial Medians Clustering

The spatial median is an extended version of the ordinary univariate median to the case of multivariate. Let $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ be n points lying in R^p. We define a spatial median of the set of points $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ to any point $\underline{a} \in R^p$ which minimizes

$$\sum_{i=1}^{n} ||\underline{x}_i - \underline{a}||.$$

The k-spatial medians clustering uses the spatial median as the reference point of the cluster. Suppose that we have the objects $\underline{x}_1, \ldots, \underline{x}_n$ on R^p. The k-spatial medians clustering procedure consists of

(i) finding $\underline{a}_{n1}, \underline{a}_{n2}, \ldots, \underline{a}_{nk}$ minimizing $1/n \sum_{i=1}^{n} \min_{1 \le j \le k} ||\underline{x}_i - \underline{a}_{nj}||$.

(ii) assigning each $\underline{x}_i$ to its nearest cluster center $\underline{a}_{nj}, 1 \le j \le k$.

The obtained $\underline{a}_{n1}, \underline{a}_{n2}, \ldots, \underline{a}_{nk}$ are the spatial median vectors of each cluster.

Gower (1994) presented the algorithm for achieving spatial median. The algorithm is adopted for finding saptial medians of each cluster in the k-spatial medians clustering. By modifying the nearest centroid sorting (MacQueen, 1967) and the transfer algorithm (Banfield and Bassill, 1977), the k-spatial medians clustering can be performed. It has two distinct phases: (1) transferring an object from one cluster to another and (2) amalgamating the single member cluster with its the nearest cluster. Given a starting partition, each possible transfer is tested in turn to see if it would improve the value of clustering criterion. When no further transfers can improve the criterion value, each possible amalgamation of the single member cluster and other clusters is tested. The amalgamation of the single member cluster should be executed with the detachment of an object which is far from its cluster centroid when it is found to be beneficial. When no further amalgamations give an improvement, the transferring phase is re-entered, and continued until no more transfers or amalgamations can improve the clustering criterion value. The following algorithm proposed by Jin (1999) explains process of the k-spatial medians clustering.

Algorithm k-spatial medians clustering(conventional version)

The partition $P(n,k)$ is composed of the clusters $1, 2, \ldots, k$. Each of the n objects lies in just one of the k clusters. The number of objects in the l^{th} cluster is n_l. The distance between the i^{th} object and the centroid of the l^{th} cluster is $d(i,l)$. The error of the partition is the sum of the within cluster dispersions, $e[P(n,k)] = \sum_{i=1}^{n} d(i, l(i))$ where $l(i)$ is the cluster to which the i^{th} object belongs.

Step 1: Assume initial clusters $1, 2, \ldots, k$. Compute the cluster spatial medians $\underline{a}_i$'s and the initial error $e[P(n,k)] = \sum_{i=1}^{n} d(i, l(i))$. If desired, the initial partition could be constructed by using the k-means clustering.

Step 2: For the first object, compute $d(1,l)$ for every cluster $l, 1 \leq l \leq k$. If the minimum of this distance $d(1,l)$ over all l is not attained from the first object's parent cluster $l(1)$, transfer the first object from cluster $l(1)$ to this minimal l, adjust the cluster spatial medians and the within cluster dispersions of $l(1)$ and the minimal l, and then add the increase in error (which is negative) to $e[P(n,k)]$.

Step 3: Repeat Step 2 for the i^{th} object $(2 \leq i \leq n)$ except for the objects which form single member clusters.

Step 4: If no movement of an object from one cluster to another occurs for any object and there exist single member clusters, get into the amalgamation phase, Step 5. Else if single member clusters do not exist, stop and complete.

Step 5: For the single member cluster $l(i)$, compute minimum $d(i,l)$ for every cluster $l \neq l(i)$.

Step 6: Find the object j which has the maximum $d(j,l(j))$ for every $j \neq i$.

Step 7: If the minimum $d(i,l)$ is less than the maximum $d(j,l(j))$, the object i is assigned its nearest cluster and the object j is separated as an independent cluster from its parent cluster.

Step 8: Repeat from Step 5 to Step 7 for every other single member clusters. If no movement between clusters occurs for all objects, the k-spatial medians clustering procedure is completed. Else go to Step 2.

However, the k-spatial medians clustering process needs heavy computations. If the number of object and dimension p of data are getting larger the computation problem of the k-spatial medians clustering procedure is getting more fatal.

3. Sample Based Algorithm of k-Spatial Medians Clustering

In order to alleviate a burden of computation in k-spatial medians clustering we propose a sample based algorithm. The sample based algorithm for k-spatial medians clustering uses random samples of size m from the entire data set for performing the k-spatial medians clustering algorithm. Sample based algorithm consists of two phases which are a phase of acquiring cluster centroids and a phase of assigning entire data to acquired centroids. At the first phase we use ramdom samples instead of entire data. It can save computing difficulty coming from large amount of data. Several sets of random samples of size m are applied to the conventional algorithm of the k-spatial medians clustering repeatedly. Cluster centroids of each repetition are summarized in representitive centroids for assigning phase. The following algorithm describes sample based k-spatial medians clustering.

Algorithm Sample based k-spatial medians clustering

Step 1: Select random sample of size m from the entire data and choose the number t of repetition.

Step 2: Choose initial seeds of k-spatial medians $\underline{a}^{(0)} = (\underline{a}_1^{(0)}, \underline{a}_2^{(0)}, \ldots, \underline{a}_k^{(0)})$ from the selected sample and set $i = 1$.

Step 3: Find k-spatial medians $\underline{a}^{(i)} = (\underline{a}_1^{(i)}, \underline{a}_2^{(i)}, \ldots, \underline{a}_k^{(i)})$ from the selected sample by conventional version of the k-spatial medians clustering algorithm with the initial seeds $\underline{a}^{(i-1)} = (\underline{a}_1^{(i-1)}, \underline{a}_2^{(i-1)}, \ldots, \underline{a}_k^{(i-1)})$.

Step 4: If $i = t$ then go to Step 6 else set $i = i + 1$.

Step 5: Select random sample of size m from the entire data and go to Step 3.

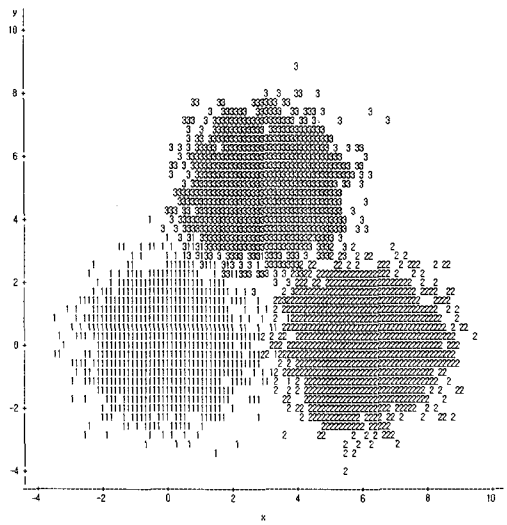

Figure 4.1. Scatter plot of the generated data (uncorrelated data)

Table 4.1. Result of k-spatial medians clustering by existing and new algorithm for the uncorrelated data

method	cluster	Dist. 1	Dist. 2	Dist. 3	Total	Computing Time (Sec.)
existing algorithm	1	4976	2	12	4990	86.45
	2	12	8	4976	4996	
	3	12	4990	12	5014	
new algorithm	1	4976	2	11	4989	5.81
	2	12	8	4977	4997	
	3	12	4990	12	5014	

Step 6: Compute $\underline{s} = (\underline{s}_1, \underline{s}_2, \ldots, \underline{s}_k)$, where $\underline{s}_j$ is the spatial median of $\underline{a}_j^{(1)}, \underline{a}_j^{(2)}, \ldots, \underline{a}_j^{(t)}$, $j = 1, 2, \ldots, k$.

Step 7: Assign entire data to the nearest cluster centroid $\underline{s}_j$, $j = 1, 2, \ldots, k$ and stop.

4. Numerical Examples

In order to verify performance of the proposed algorithm two numerical studies are proceeded. CPU and memory of the machine that we used for simulation are Phenom X4 Quad-Core Processor GP-9600 and 3.25GB RAM. The resulting clusters and CPU time are compared each other.

4.1. Normal distribution case(uncorrelated structure)

Each of 5000 data points are coming from three normal distributions $N(\underline{\mu}_i, \Sigma)$, $i = 1, 2, 3$ with $\underline{\mu}_1 = \binom{0}{0}$, $\underline{\mu}_2 = \binom{6}{0}$, $\underline{\mu}_3 = \binom{3}{5}$ and covariance structure $\Sigma = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. Figure 4.1 represents the generated data set. Each point is plotted by its underlying distribution identification number. Table 4.1 shows the results of the k-spatial medians clustering by existing algorithm and sample based algorithm.

The results of the two algorithms are almost same. However, CPU times for computation are much different. Using same machine for simulation, 86.45(sec) is needed for obtaining the result by

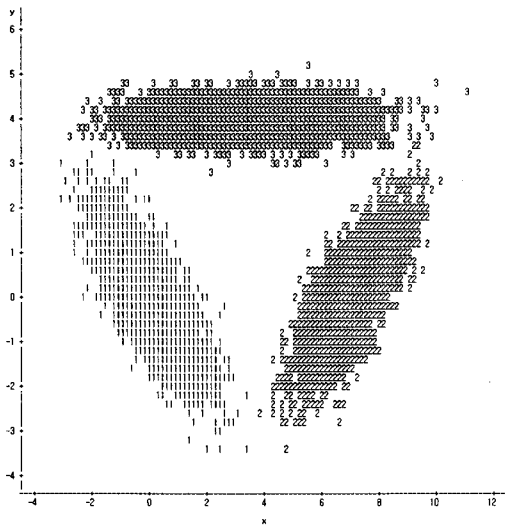

Figure 4.2. Scatter plot of the generated data (correlated data)

Table 4.2. Result of k-spatial medians clustering by existing and new algorithm for the correlated data

method	cluster	Dist. 1	Dist. 2	Dist. 3	Total	Computing Time (Sec.)
existing algorithm	1	5000	0	147	5147	17896.91
	2	0	0	4753	4753	
	3	0	5000	100	5100	
new algorithm	1	5000	0	137	5137	20.25
	2	0	0	4751	4751	
	3	0	5000	112	5112	
k-means algorithm	1	3991	0	2870	6861	
	2	0	2857	2130	4987	
	3	1009	2143	0	3152	

existing algorithm but only 5.81(sec) is necessary for sample based algorithm. We chose $m = 500$ and $t = 30$ for sample based algorithm.

4.2. Normal distribution case(correlated structure)

Each of 5000 data points are coming from three normal distributions $N(\underline{\mu}_i, \Sigma_i)$, $i = 1, 2, 3$ with $\underline{\mu}_1 = \binom{0}{0}$, $\underline{\mu}_2 = \binom{7}{0}$, $\underline{\mu}_3 = \binom{3.5}{4}$ and covariance structure $\Sigma_1 = \begin{pmatrix} 0.8 & 0.7 \\ 0.7 & 1.0 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 0.8 & -0.7 \\ -0.7 & 1.0 \end{pmatrix}$, $\Sigma_3 = \begin{pmatrix} 4 & 0 \\ 0 & 0.09 \end{pmatrix}$. Figure 4.2 represents the generated data set. Each point is plotted by its underlying distribution identification number. Table 4.2 shows the results of the k-spatial medians clustering by existing algorithm and sample based algorithm.

Similar to uncorrelated normal distribution case, the results by two algorithms are not much different from each other. However, CPU times for computation are severely different from each other. The existing algorithm used 17,896.91(sec) for obtaining the result but sample based algorithm needed only 20.25(sec) under $m = 500$ and $t = 30$. In addition, the result of k-means clustering is also given in Table 4.2 for comparing the performance with k-spatial medians clustering. Under correlated structure k-means clustering could not build proper clustering result.

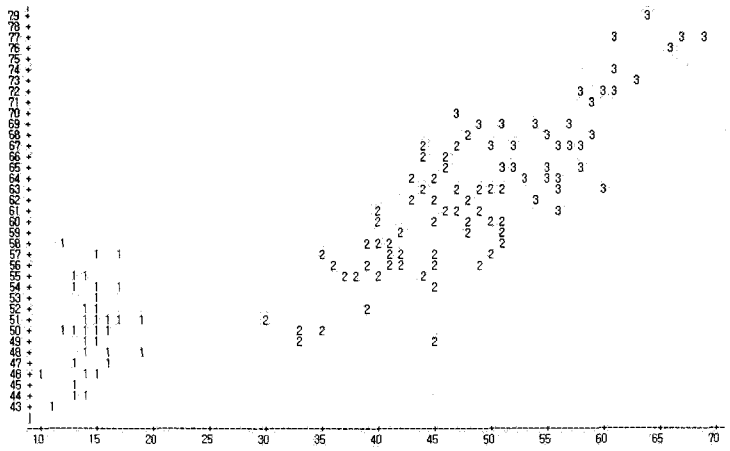

Figure 5.1. Results of k-spatial medians clustering(Sepal Length $\times$ Petal Length)

Table 5.1. mean profiles of each cluster

cluster	Sepal Length	Sepal Width	Petal Length	Petal Width
1	50.06	34.28	14.62	2.46
2	58.85	27.40	43.77	14.18
3	68.28	30.70	57.00	20.63

Table 5.2. frequencies of each cluster

cluster	Setosa	Versicolor	Virginica	Total
1	50	0	0	50
2	0	47	13	60
3	0	3	37	40

5. Case Study

In order to show the performance of the proposed algorithm, we applied the algorithm to the data of iris which was used by Fisher (1936). The data give measurements of four flower parts(sepal length, sepal with, petal length, and petal width, in centimeters) on 50 specimens of each of three species of iris.

We chose $m = 30$ and $t = 30$ for applying the proposed algorithm. Table 5.1 shows the results of the k-spatial medians clustering by the proposed algorithm. In addition, Figure 5.1 shows the resulting cluster features on the plane generated by sepal length and petal length. From Table 5.1 and Figure 5.1, we found that the three clusters are well separated each other.

The mean profiles of each cluster are given in the Table 5.2. From Table 5.1, cluster 1 has small values in petal length and petal width. The objects in cluster 3 are generally large sized objects. Four variables are well differentiated by clusters. Table 5.2 shows the frequencies of three species of iris which are divided into three groups.

6. Conclusions

The k-spatial medians clustering, one of the partitioning methods for cluster analysis, has many strong points for making clusters under existence of outliers and special cluster structures. However,

the k-spatial medians clustering has not been used a lot because its computational difficulty. The problem of computation is very serious when the number of objects and dimension of the data are increased. In order to solve the computation problem a modified algorithm of the k-spatial medians clustering was proposed. The algorithm is based on the sample of the data, so it is able to decrease computation efforts. Numerical examples and the case study showed outstanding performance of the proposed algorithm. Proper clusters were obtained with comparatively short computing time. Lastly, one thing that we have to mention is selection problem of the sample size m and the number of repetition t which are necessary for carrying out the sample based algorithm. The choice of those numbers depends on the case. We might need large number in some cases. However, according to some empirical studies we don't need very large numbers. For finding proper numbers, we can try the analysis several times independently with incresing the numbers. If the results of analysis are stable then we can finalize the analysis with those numbers.

References

Banfield, C. F. and Bassill, L. C. (1977). A transfer algorithm for non-hierarchical classification, *Applied Statistics*, **26**, 206–210.

Butler, R. W. (1986). Optimal stratification and clustering on the line using the L_1-norm, *Journal of Multivariate Analysis*, **18**, 142–155.

Butler, R. W. (1988). Optimal clustering in the real line, *Journal of Multivariate Analysis*, **24**, 88–108.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **7**, 179–184.

Gower, J. C. (1994). The mediancentre, *Applied Statistics*, **23**, 466–470.

Jhun, M. (1986). Bootstrap method for k-spatial medians, *Jouranl of the Korean Statistical Society*, **15**, 1–8.

Jhun, M. and Jin, S. (2000). On a modified k-spatial medians clustering, *Jouranl of the Korean Statistical Society*, **29**, 247–260.

Jin, S. (1999). *A Study On The Partitioning Method For Cluster Analysis,* Dissertation, Korea Unversity.

Kaufman, K. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations, *Proceeding Symposium of Mathematical Statistics and Probability*, **1**, 281–297.