〈Research Paper〉

Journal of the Korean
Statistical Society
Vol. 15, No. 1, 1986

Bootstrap Method for k-Spatial Medians

Myoungshic Jhun*

ABSTRACT

The k-medians clustering method is considered to partition observations into k clusters. Consistency and advantage of bootstrap confidence sets of k optimal cluster centers are discussed. The k-medians and k-means clustering methods are compared by using actual data sets.

1. Introduction

Independent multi-dimensional observations $X_1, X_2, \cdots, X_n$ are made on a distribution F on $R^d (d>1)$. Let F_n be an empirical distribution function of the observations. It is desired to partition these observations into k clusters so that observations within clusters are close, in some sense, and observations in different clusters are distant. Here we can consider some possible distance functions. The k-medians clustering procedure consists of

(i) finding $m_n=(m_{n1}, \cdots, m_{nk})$ minimizing $\frac{1}{n}\sum_{i=1}^{n}\min_{1\le j\le k}\|X_i-m_j\|=\int\phi(\cdot, m)\,dF_n$ where $\|\cdot\|$ is a Euclidean norm and $\phi(\cdot, m)=\min_{1\le j\le k}\|\cdot-m_j\|$, and

(ii) assigning each X_i to its nearest cluster center.

When $k=1$, m_n is the spatial median. Brown(1983) discussed statistical uses of the spatial median.

Pollard(1981) obtained strong consistency of such procedure in a more general setting. He showed almost sure convergence of m_n to m where m_n and m minimize

* Department of Statistics, The University of Michigan, Ann Arbor, Michigan 48109, U.S.A.

$\int \phi(\cdot, m) dF_n = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \le j \le k} \eta(\|X_i - m_j\|)$ and $\int \phi(\cdot, m) dF = \int \min_{1 \le j \le k} \eta(\|\cdot - m_j\|) dF$ respectively with $\eta(\cdot)$ an increasing function. If $\eta(x) = x^2 (x \ge 0)$, we call the procedure k-means clustering. Consistency of bootstrap method (i.e. correct asymptotic levels for the bootstrap confidence sets) for k-means clustering procedure was obtained (Jhun, 1985) and the performance of the bootstrap method was demonstrated for testing clusters (Jhun, 1986).

In this paper, consistency of the bootstrap method for k-medians clustering procedure will be considered. A clear advantage of using the bootstrap method in this case is that we can avoid the estimation of the complicated covariance matrix of the asymptotic distribution. Arguments are analogous to Jhun (1985). Comparison between k-medians and k-means clustering procedures will be made by using actual data sets.

2. Theory

Let $X_1, X_2, \cdots, X_n$ be independent R^d-valued $(d > 1)$ random vectors with common distribution F, and F_n be the empirical distribution function of the observations X_1, X_2, $\cdots$, X_n.

To divide the observations X_1, $X_2, \cdots, X_n$ in R^d into k clusters, first choose a vector of k optimal cluster centers $m_n = (m_{n1}, \cdots, m_{nk})$ by minizing within cluster sum of deviations $W_n(m) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \le j \le k} \|X_i - m_j\| = \int \phi(\cdot, m) \, dF_n$, then assign each X_i to its nearest cluster center. Associated with each cluster center m_{nj} is the convex polyhedral region A_{nj} of all points in R^d closer to m_{nj} than any other cluster center. Each cluster center m_{nj} is the spatial median of the observations in its cluster. We have strong consistency of $m_n = (m_{n1}, \cdots, m_{nk})$ from Pollard (1981) as follows;

Theorem 1 : Let $m_0 = (m_{01}, \cdots, m_{0k})$ minimizing $\int \phi(\cdot, m) dF < \infty$ be unique. Then $m_n \to m_0$ almost surely where $m_n = (m_{n1}, \cdots m_{nk})$ minimizes $\int \phi(\cdot, m) dF_n$.

Now we will consider bootstrap method to estimate the sampling distribution of $n^{1/2}(m_n - m_0)$. Given X_1, $X_2, \cdots, X_n$ we have an empirical distribution F_n of these observations.

Draw n independent random observations X_1^*, $X_2^*, \cdots, X_n^*$ (bootstrap sample) with replacement from F_n. Let F_n^* be the empirical distribution function of the bootstrap

sample. We can obtain $m_n^*=(m^*_{n1},\cdots,m^*_{nk})$ which minimizes $\frac{1}{n}\sum_{i=1}^{n}\min_{1\le j\le k}\|X_i^*-m_j\|=\int\phi(\cdot,\ m)dF_n^*$. Then we have a bootstrap quantity $n^{1/2}(m^*_n-m_n)$, which has the same limiting distribution as the actual distribution, under some conditions, according to the following theorem.

Theorem 2 : Let $m_n^*=(m^*_{n1},\cdots,m^*_{nk})$ be the vector of optimal k-medians for independent bootstrap sampling from an empirical distributin function F_n of the i.i.d. observations $X_1, X_2,\cdots, X_n$ on R^d.

Suppose

(i) the vector m_n that minimizes $\int\phi(\cdot,\ m)\ dF_n$ is unique(up to relabeling its co-ordinates),

(ii) the map $a\to\int\phi(\cdot,\ a)dF_n$ has positive definite second order derivative Γ_n at $a=m_n$,

(iii) $\int\|x\|^2dF(x)<\infty$.

Then

$n^{1/2}(m_n^*-m_n)$ converges weakly to $N(0,\Gamma^{-1}V\Gamma^{-1})$ where Γ and V are defined at (4) and (6) of the Appendix respectively.

Proof : See Appendix. ▨

Theorem 2 justifies the use of bootstrap method, in the sense that bootstrap confidence set of optimal 'k-medians' $m_0=(m_{01},\cdots,m_{0k})$ has the correct asymptotic level(cf. Beran, 1985).

The main reason for using the bootstrap method instead of direct asymptotic method is that estimation of the covariance matrix $\Gamma^{-1}V\Gamma^{-1}$ of the limiting distribution is too complicated. We can avoid the estimation problem by using the bootstrap distribution, which can be used as a practical approximation of the sampling distribution. Notice that the existence of the same limiting covariance matrix is good enough for our justification. When $k=1$, m_n is the spatial median (cf. Brown, 1983), and the bootstrap method is also consistent in this case from Theorem 2.

3. Applications and Examples

From Theorem 2, bootstrap confidence sets for the unknown optimal 'k-medians'

$m_0=(m_{01},\cdots,m_{0k})$ is valid. We can use the quantity $|n^{1/2}(m_n-m_0)|$ to obtain the confidence sets of m_0 where $|\cdot|$ is any metric on R^d. Notice that it is equally easy to use the bootstrap method for the different metrics on R^d. Bootstrap confidence sets for m_0 can be used to test the hypothesis about the values of $m_0=(m_{01},\cdots,m_{0k})$ by inverting the bootstrap confidence sets in the usual way. In particular, when $k=1$ we call m_0 by spatial median(ref. Brown (1983)). We can use the bootstrap distribution for angle tests (tests of spatial location) in the usual way.

In getting 'k-medians', we use the following algorithm, which is modified from one in Hartigan(1975).

Preliminaries : Let $X_i=(X_{i1},X_{i2},\cdots,X_{id})$, $i=1,2,\cdots,n$. The ith case of the jth variable has value x_{ij}. The partition $P(n,k)$ is composed of the clusters $1,2,\cdots,k$. Each of the n observations lies in just one of the k clusters. The spatial median of the jth variable over the cases in the lth cluster is denoted by b_{lj}. The number of observations in lth cluster is $n(l)$. The distance between the ith observation and lth cluster is

$$D(i,l)=\sqrt{\sum_j(x_{ij}-b_{lj})^2}$$

The error of the partition is

$$e[P(n,k)]=\sum_i D(i,l(i)),$$

where $l(i)$ is the cluster containing the ith case.

Step 1) : Assume initial clusters $1,2,\cdots,k$. Compute the cluster spatial medians b_{lj} and the initial error

$$e[P(n,k)]=\sum_i D(i,l(i)).$$

Step 2) : For the first case, compute for every cluster l

$$\frac{n(l)D(1,l)}{n(l)+1}-\frac{n[l(1)]D[1,l(1)]}{n[l(1)]-1}.$$

The increase in error in transferring the first case from cluster $l(1)$, to which it belongs at present, to cluster l. If the minimum of this quantity over all $l\neq l(1)$ is negative, transfer the first case from cluster $l(1)$ to this minimal l, adjust the cluster spatial medians of $l(1)$ and the minimal l, and add the increase in error(which is negative) to $e[P(n,k)]$.

Step 3) : Repeat Step 2 for the ith case $(2\leq i\leq n)$.

Step 4) : If no movement of a case from one cluster to another occurs for any case, stop. Otherwise, return to Step 2.

For the computation of the spatial medians in Step 2 and Step 3, we use a subroutine

written by Gower(1974). Unfortunately it took a very long time to compute an approximation of the bootstrap distribution with the facilities available to the author, and wedecided to postpone it for a while.

The k-medians clustering method is now compared with k-means clustering by using two examples with real data. The first example shows significant difference between two methods, but the second one shows almost the same result for both methods.

Example 1 : As an illustration of comparison between k-medians and k-means clustering methods the data of Table 1 [taken from page 175 of (Hand, 1981)] which shows the distibution of rhesus genes in different populations was analyzed.

Table 1. Distribution of rhesus genes in different populations

Population	Genes							
	CDE	CDe	CdE	Cde	cDE	cdE	cDe	cde
Caucasoid								
1. Danes	0.1	42.2	0.0	1.3	15.1	0.7	1.8	38.8
2. Italians	0.4	47.6	0.3	0.7	10.8	0.7	1.6	38.0
3. Spaniards	0.1	43.2	0.0	1.9	12.0	0.0	3.7	38.0
4. Australian Aborigines (Early Caucasoid) Mongoloid(Recent)	2.1	56.4	0.0	12.9	20.1	0.0	8.5	0.0
5. South Chinese	0.5	75.9	0.0	0.0	19.5	0.0	4.1	0.0
6. Japanese Mongoloid(Early)	0.4	60.2	0.0	0.0	30.8	3.3	0.0	5.3
7. Eskimos(Greenland)	3.4	72.5	0.0	0.0	22.0	0.0	2.1	0.0
8. Navaho	1.3	43.1	0.0	0.0	27.7	0.0	28.0	0.0
9. Blood	4.1	47.8	0.0	0.0	34.8	3.4	0.0	9.9
10. Chippewa	2.0	33.7	0.0	0.0	53.0	3.2	0.0	8.0
Negroid								
11. Bushman(Early Negroid)	0.0	9.0	0.0	0.0	2.0	0.0	89.0	0.0
12. Shona(Rhodesia) (Mixed Negroid-Caucasoid)	0.0	6.9	0.0	0.0	6.4	0.0	62.7	23.9

Hand (1981) obtained the following result by using ($k=4$) k-means method.

Points 1, 2, and 3 to cluster 1.

Points 4, 5, 7, and 8 to cluster 2.

Points 6, 9, and 10 to cluster 3.

Points 11 and 12 to cluster 4.

On the other hand, by using ($k=4$) k-medians method the following result can be

obtained:

Points 1, 2, and 3 to cluster 1.

Points 4 and 6 to cluster 2.

Points 5, 7, 8, 9, and 10 to cluster 3.

Points 11 and 12 to cluster 4.

In terms of allocation of points, cluster 1 and cluster 4 don't have any difference between k-medians clustering and k-means clustering. But, cluster 2 and cluster 3 have a big difference between two clustering methods. It can be explained by noticing the differences between spatial medians and means of cluster 2 and cluster 3 (See, Table 2 and Table 3).

Table 2. Spatial Median of Clusters for k-Median Clustering

	CDE	CDe	CdE	Cde	cDE	cdE	cDe	cde
Cluster 1	0.14	43.63	0.04	1.61	12.43	0.23	3.03	38.16
Cluster 2	1.25	54.32	0.0	0.0	31.26	1.68	5.32	4.72
Cluster 3	2.70	58.30	0.0	6.45	25.45	1.65	4.25	2.65
Cluster 4	0.0	7.95	0.0	0.0	4.20	0.0	75.85	11.95

Table 3. Mean of Cluster for k-Means Clustering

	CDE	CDe	CdE	Cde	cDE	cdE	cDe	cde
Cluster 1	0.20	43.33	0.1	1.3	12.63	0.47	2.37	38.27
Cluster 2	1.83	61.98	0.0	3.23	22.33	0.0	10.68	0.0
Cluster 3	2.17	47.23	0.0	0.0	39.53	3.3	0.0	7.73
Cluster 4	0.0	7.95	0.0	0.0	4.2	0.0	75.85	11.95

Example 2 : Fisher(1936) Iris data is used to compare k-medians and k-means clustering methods. The data consist of four characteristics (Sepal Length, Sepal Width, Petal Length, Petal Width) for three species of Iris. There are 50 observations from each species. Applying both k-medians and k-means clustering ($k=3$), we have Table 4 and Table 5 for 'optimal' cluster centers. Actually, we have only one data point change between cluster 2 and cluster 3 for the two different clustering methods. Overall, we don't see any serious difference between the two clustering methods.

Table 4. Spatial Median of Clusters for k-Median Clustering

	Sepal Length	Sepal Width	Petal Length	Petal Width
Cluster 1 (50)*	5.015	3.418	1.468	0.238
Cluster 2 (60)*	5.916	2.791	4.404	1.414
Cluster 3 (40)*	6.736	3.068	5.614	2.096

*Cluster size

Table 5. Mean of Clusters for k-Means Clustering

	Sepal Length	Sepal Width	Petal Length	Petal Width
Cluster 1 (50)*	5.006	3.428	1.462	0.246
Cluster 2 (61)*	5.884	2.741	4.384	1.434
Cluster 3 (39)*	6.954	3.077	5.715	2.054

*Cluster size

Acknowledgement

The author thanks the editor and the referees for their helpful comments. The author also thanks Mr. Zoonky Lee for helping out with the computer programming.

Appendix

Proof of Theorem 2

(1) $\phi(x, m_n+h)=\phi(x, m_n)+\sum_{i=1}^{k} h_i'\eta(x, m_{ni})\ I\{x\varepsilon A_{ni}\}$

$$+\|h\|\cdot R(x, m_n, h) \text{ where } \eta(x, m)=\frac{x-m}{\|x-m\|}$$

(2) $R(x, m_n, h)=\sum_{i=1}^{k}\frac{\|x-m_{ni}-h_i\|-\|x-m_{ni}\|-h_i'\eta(x, m_n)}{\|h\|}\ I\{x\varepsilon A_{ni}\}$

$$\leq\sum_{i=1}^{k}(\|h\|^{-1}(\|x-m_{ni}-h_i\|^2-\|x-m_{ni}\|^2)/(\|x-m_{ni}-h_i\|$$

$$+\|x-m_{ni}\|)+1)I\{x\varepsilon A_{ni}\}$$

$$\leq\sum_{i=1}^{k}((2\|x-m_{ni}\|+\|h_i\|)/(\|x-m_{ni}-h_i\|+\|x-m_{ni}\|)+1)I\{x\varepsilon A_{ni}\}$$

$$\leq 4k\ \varepsilon L^2(F_n)$$

and $\int R^2(x, m_n, h)dF_n\to 0$ as $h\to 0$.

Let $B_n^*=n^{1/2}(F_n^*-F_n)$ where F_n^* is an empirical distribution of independent random observations $X_1^*, \cdots, X_n^*$ with replacement from F_n. Then, by using asymptotic behavior of empirical processes on Vapnik-Chervonankis classes in triangular array setting, we have

(3) $\int\phi(\cdot,\ m_n^*)dB_n^*=\int\phi(\cdot,\ m_n)dB_n^*+(m_n^*-m_n)'\int\eta(\cdot,\ m_n)dB_n^*+o_p(r_n^*)$

where $r_n^*=\|m_n^*-m_n\|=o_p(1)$.

Also

(4) $\Gamma_n \to \Gamma$ where Γ is a positive definite second order derivative of a map $m \to \int \phi(\cdot, m)\ dF$ at $m=m_0$.

Since

$$\int \phi(\cdot, m_n^*) dF_n = \int \phi(\cdot, m_n) dF_n + \frac{1}{2}(m_n^* - m_n)' \Gamma_n (m_n^* - m_n) + o_p(r_n^*)^2,$$

we have

$$(5)\quad \int \phi(\cdot, m_n^*) dB_n^* = \int \phi(\cdot, m_n) dB_n^* - n^{1/2} Z_n'(m_n^* - m_n) + \left(\frac{1}{2}\right)(m_n^* - m_n)' \cdot F_n(m_n^* - m_n) + o_p(n^{-1/2} r_n^*) + o_p(r_n^{*2})$$

where $Z_n = -\int \eta(\cdot, m_n) dB_n^*$ converges weakly to $N(0, V)$ with

$$(6)\quad V = \int \eta(\cdot, m)\eta(\cdot, m)' dF.$$

$\int \phi(\cdot, m_n^*) dB_n^* \le \int \phi(\cdot, m_n) dB_n^*$ and (5) implies $r_n^* = O_p(n^{-1/2})$.

Let $n^{1/2}(m_n^* - m_n) = \theta_n^*$.

Then

$$\int \phi(\cdot, m_n^*) dB_n^* = \int \phi(\cdot, m_n + n^{1/2}\Gamma_n^{-1} Z_n) dB_n^* + \frac{1}{2} n^{-1} \| \Gamma_n^{1/2} \theta_n^* - \Gamma_n^{-1/2} Z_n \|^2 + o_p(n^{-1}).$$

$\int \phi(\cdot, m_n^*) dB_n^* \le \int \phi(\cdot, m_n + n^{-1/2}\Gamma_n^{-1} Z_n) dB_n^*$ forces $\theta_n^* = \Gamma_n^{-1} Z_n + o_p(1)$. ▨

References

(1) Beran, R. (1985). Bootstrap Methods in Statistics. *Jber. D. Dt. Math. Verein.*, **86**, 14~30.

(2) Brown, B. (1983). Statistical Uses of the Spatial Median. *J.R. Statist. Soc. B*, **45**, No. 1, 25~30.

(3) Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, **7**, Part 2, 179~188.

(4) Gower, J. (1974). The Mediancentre. *Appl. Statistics*, **23**, 466~470.

(5) Hand, D.J. (1981). *Discrimination and Classification.* John Wiley and Sons.

(6) Hartigan, J.A. (1975). *Clustering Algorithms.* Wiley, New York.

(7) Jhun, M. (1985). Bootstrapping k-Means Clustering. Tech. Report #135, Statistics Department, The University of Michigan.

(8) Jhun, M. (1985). Comparison of Distributions for Testing Clusters. Tech. Report #133, Statistics Department, The University of Michigan.

(9) Pollard, D. (1981). Strong Consistency of k-Means Clustering. *Ann. Statist.*, **9**, 1, 135~140.