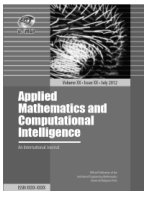
App. Math. and Comp. Intel., Vol. 2 (2) (2013) 149–155

http://amci.unimap.edu.my



PAM modification using trimmed K-median based on TCLUST cluster analysis

M. A. Md. Jedi[a,*], and R. Adnan[a]

[a]*Department of Mathematics, Faculty of Science*
Universiti Teknologi Malaysia
81310 Skudai, Johor Baharu, Malaysia

Received: 22 April 2013, Accepted: 20 July 2013

Abstract: *This paper will discuss the TCLUST algorithm using restriction of constrains to scatter matrices. We are discussing among three constrains eigenvalue, matrix determinant and same sized cluster (sigma) that affect the shape of clusters. Trimming process using TCLUST is made to detect the best proportion of contaminated data and the best number of clusters to be used in the next step. Based on prior knowledge of TCLUST we are using the PAM to determine the best mediod that shape the data. The results are discussed between the three types of constraints. At the end of this paper we compared the TLUCT based on trimmed k-means method with modified PAM based on trimmed k-median method.*

Keywords: *TCLUST, PAM, trimmed k-means*

PACS:

1 Introduction

TCLUST is based on trimmed k-means. It is not a new concept but to apply in cluster analysis it is required sensible justification to make it possible to analysis. Therefore, this analysis was referred as robust methods where it possible to handle the large amount of outlying data. Compared to the other methods (non-robust), clustering result may be heavily influenced even by small amount of contaminated data [3]. Garcia et.al. refer the outlying data through outlying model or called "spurious outlier model". Partition Around Mediod, PAM is another clustering method. It is not a robust method if there are a large amount of outlying data exists. However it will consider being robust method if there is only small amount of outlying data exits. The idea to make PAM to be possible to remains as robust method is to apply the trimming process for k-median in PAM. To implement this, justification of spurious outlier model is required as prior knowledge to PAM method.

* Corresponding Author: author@yahoo.com (M. A. Md.Jedi)

2 TCLUST and PAM

2.1 TCLUST algorithm

TCUST method simply removes outlying data and does not intent to fit them at all. The spurious outlier model is a probabilistic framework for robust clustering. [1],[2]. Let $f(\cdot;\mu,\Sigma)$ denoted the probability density function of the p-variate normal distribution with mean, μ and covariance matrix Σ. The spurious outlier model is defined through likelihoods like

$$\left[\prod_{j=1}^{k}\prod_{i\in R_j} f(x_i;\mu_j,\Sigma_j)\right]\left[\prod_{i\in R_0} g_i(x_i)\right] \tag{1}$$

g_i is probability function of outliers where R_0 are the indices of the outliers (generated by g_i). By maximizing (1) vectors μ_j and positive definite matrices Σ_j can be simplified by

$$\sum_{j=1}^{k}\sum_{i\in R_j} \log f(x_i;\mu_j,\Sigma_j) \tag{2}$$

Notice that the outlier function does not take into account in (2). This will yields the Minimum Covariance Determinant (MCD) estimator by maximizing (2) for k=1. However, for k > 1 direct maximizing is not well defined because (2) is not bound with any constraint on the scatter matrices Σ_j. Therefore by considering the clusters size or weight, the partition of the clusters can be defined through log likelihood function

$$\sum_{j=1}^{k}\sum_{i\in R_j}\left(\log \Pi_j + \log f(x_i;\mu_j,\Sigma_j)\right) \tag{3}$$

For (3), the scatter matrices Σ_j have to be constrained such that the maximizing of (3) becomes a well defined problem.

2.2 Modification of PAM

The TCLUST result will be used as a prior knowledge for PAM. For this study, the trimmed data are used to be further analyzing in PAM. The partition of $\{R_0,...,R_k\}$, vector μ_j, positive definite matrices Σ_j and weight $\pi\in[0,1]$ upon post-maximizing of (3) will be applied using the rules of thumb of PAM. For these methods we refer the procedure as trimmed k-median because of trimming part in TCLUST. Theoretically the maximum of (3) will yield the data with no outlier. For bivariate case, the initial data will be calculate based on

$$\sum_{i=1}^{n} p_{ij}\,(j=1,..,n) \text{ where } p_{ij}=\frac{d_{ij}}{\sum_{l=1}^{n} d_{il}} \quad i=1,...,n;\ j=1,...,n \tag{4}$$

and d_{ij} is euclidean distance between every pairs of all data. By sharing the same characteristics of group assignment in TCLUST, modification of PAM are bound with a constraints.

3 Methodology and Simulation

Before we analysis the data using TCLUST, there are essentially three types of constraints as proposed by Garcia et.al. [1]. These three constraints give the different graphical output of data resulting from the nature of algorithm. Notice that TCLUST's scatter matrices constraints are controlled by constant c such that

$$M_n / m_n \le c \text{ where } c \ge 1 \tag{5}$$

In this method simulation of data generated TCLUST result will be gained and had been used in PAM. This is to test if there is a different between TCLUST and modified PAM.

A. TCLUST's Constraints

Three types of constraints mentioned are scatter matrices of eigenvalues, matrix determinant and the same size of cluster. The expansion and details of constrains are

- Eigenvalues of the group covariance matrices are defined in such that

$$M_n = \max_{j=1,\ldots,k} \max_{l=1,\ldots,p} \lambda_l\left(\Sigma_j\right) \text{ and } m_n = \min_{j=1,\ldots,k} \min_{l=1,\ldots,p} \lambda_l\left(\Sigma_j\right) \tag{6}$$

- Scatter matrices determinants where

$$M_n = \max_{j=1,\ldots,k} \left|\Sigma_j\right| \text{ and } m_n = \min_{j=1,\ldots,k} \left|\Sigma_j\right| \tag{7}$$

- Equal scatter matrices where

$$\Sigma_1 = \ldots = \Sigma_k \tag{8}$$

Equation (6) and (7) will satisfy (5). The constant c will control the strength of scatter constraints. To modified PAM, the information of TCLUST's constraints is required.

B. Algorithm (modified PAM)

- Random starts: Draw k random initial mediod $c_1^0,\ldots,c_k^0$, k random initial covariance matrices $\Sigma_1^0,\ldots,\Sigma_k^0$
- Concentration steps: Assign covariance matrices $\Sigma_1,\ldots,\Sigma_k$ to the nearest mediod $c_1,\ldots,c_k$
- Keep the set H made of the $\left\lceil n(1-\alpha)\right\rceil$ observation closest to the center $c_1^l,\ldots,c_k^l$. $\left\lceil n(1-\alpha)\right\rceil$ observation x_i 's with smallest value for $d_{ij} = \min_{j=1,\ldots,k} \pi_j^l f\left(x; c_j^l, \Sigma_j^l\right)$
- Partition H onto $\{H_1,\ldots,H_k\}$ where H_j contains the observations in H closer, by using Euclidean to the center c_j^l than to other centers.
- Update the center $c_1^{l+1},\ldots,c_k^{l+1}$ and covariance matrices $\Sigma_1^{l+1},\ldots,\Sigma_k^{l+1}$ such that c_j^{l+1} is the sample mediod and Σ_j^{l+1} is a sample covariance matrix of the observations in H_j.
- Repeat the step and keep the best solution in sense to maximizing (2).

C. Simulation Study

In non hierarchical cluster analysis one of the most complex problem is the choice of the number of cluster, k. We might have an idea about the initial number of clusters, but usually k is completely

unknown. To demonstrate the TCLUST cluster analysis, simulation study is conducted to interpret the mixture of component and outlier proportion.

For simulation study, we focus on the eigenvalues of the group covariance matrices $\lambda_l(\Sigma)_j$. Since the choice of k should depend on assumptions, we will consider a data set with $n = 1000$ through three-component Gaussian mixture with mixing parameter $\pi_1 = 0.35,\ \pi_2 = 0.55,$ and $\pi_3 = 0.1$. The assumed means are $\mu_1 = (1,1)'$, $\mu_2 = (3,6)'$, and $\mu_3 = (6,9)'$. The covariance matrices are

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 10 & -2 \\ -2 & 10 \end{pmatrix},\ \text{and},\ \Sigma_3 = \begin{pmatrix} 60 & 0 \\ 0 & 60 \end{pmatrix}$$

Note that, three-component Gaussian mixture has the largest eigenvalue of 60 where we defined $c = 60$. For TCLUST analysis we will assumed the data having 5% of outliers. After we trimmed 5% of the data, all the non-outlier data will be further analysis using PAM. To demonstrate the relation between α, k and c let us consider the Gaussian mixture with eigenvalue restriction. Considering Σ_1 and Σ_2, the quotient of the largest and smallest eigenvalue is 12 and 1 respectively, whereas the maximal quotient is 60 if we consider Σ_1, Σ_2 and Σ_3. Thus $c = 12$ would allow to consider two clusters while $c = 60$ would allow to assume three clusters there.

Figure 1 are perfectly sensible and the final choice of α and k only depends on the value given to c. Fig 1(a) considers three clusters while fig. 1(b) considers two clusters. Although the proportion of outliers is assumed to be 5% we consider $k = 2$ because the third Gaussian component are too small.

In TCLUST, some additional function called 'DiscrFact' in R can be applied in order to evaluate the quality of the cluster assignments and the trimming decisions. Let $\hat{R} = \{\hat{R}_0, \hat{R}_1, ..., \hat{R}_k\}, \hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_k)$ and $\hat{\pi} = (\hat{\pi}_1, ..., \hat{\pi}_k)$ be the value obtained by maximizing (2) and (3), hence

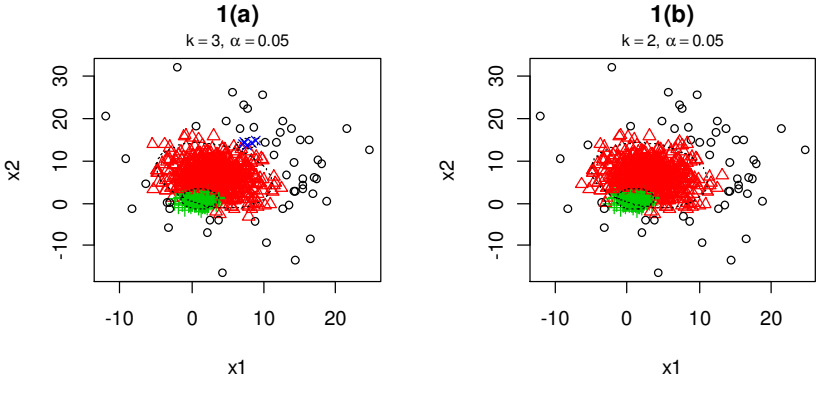


Figure 1: Clustering results for simulated data set with c= 60 (1a) and c=12 (1b).

$D_j\left(x_i;\hat{\theta},\hat{\pi}\right)=\hat{\pi}_j\phi\left(x_i,\hat{\theta}_j\right)$ is a measure of the degree of affiliation of observation x_i and j. These value can be ordered as $D_{(1)}\left(x_i;\hat{\theta},\hat{\pi}\right)\le\ldots\le D_{(k)}\left(x_i;\hat{\theta},\hat{\pi}\right)$. Thus the quality of the assignment decision of a non trimmed observation x_i to cluster j can be evaluated by comparing its degree of affiliation with cluster j to the best second possible assignment. That is, Drisciminant factor $DF_{(i)}$

$$DF_{(i)}=\log\left(D_{(k-1)}\left(x_i;\hat{\theta},\hat{\pi}\right)\Big/D_{(k)}\left(x_i;\hat{\theta},\hat{\pi}\right)\right) \tag{9}$$

Observation with large $DF_{(i)}$ indicate doubtful assignment or trimming decisions. However "Silhouette" plot can be used to summarizing the discriminant factors. Large $DF_{(i)}$ values indicate of not very well-determined clusters. Figure 2 shows the Silhouette plot have the best two cluster in data with has minimum absolute value of $DF_{(i)}$.

Figure 2 suggest the classification of data should have two clusters. Silhouette plot gives the values of mean discriminant factors to indicate the strength of group assignment. Whereas doubtful assignment resemble the location of the doubtful decisions which are located in the overlapping area.

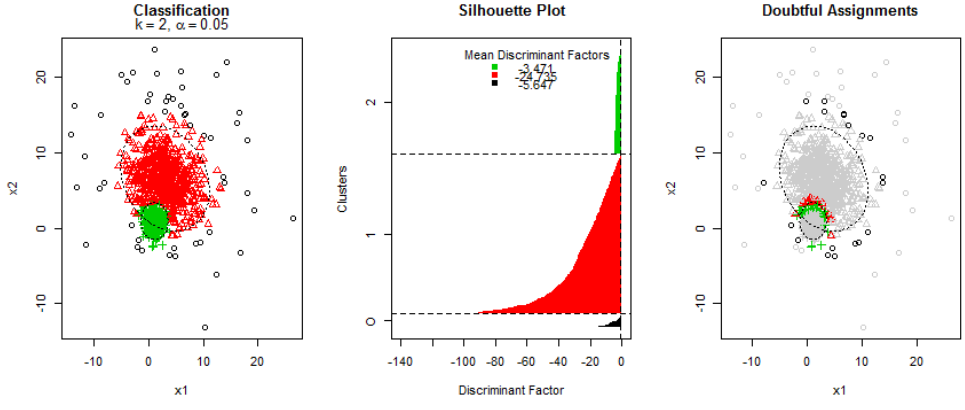


Figure 2: Graphical displays based on the discriminant factor values for TCLUCT cluster solution with k=2 and c =60.

In modified PAM the same simulation data were applied. After trimmed the outlier, PAM result shows in fig. 3. For PAM cluster plot it seems to have two distinct cluster with overall mean Silhouette 0.46. Noted that modified PAM show better result compared to TCLUST because of the lower value of mean Silhouette for second cluster that is 0.30 compared to TCLUS mean Silhouette that is 3.471.

In real data, the following example is applied to a bivariate data set based on the Old Faithful Geyser data on TCLUST's R-package. The data explain the eruption length of geyser against the

prvious eruption length. In this data set, there are 3% of the outliers present. Figure 4 explained 3 identical cluster with 6 anomalous eruption length. The absolute mean $DF_{(i)}$ for Silhouette plot for cluster 3 is 10.01. However fig.5 suggest that modified PAM shows better result for cluster 3 with having lower value of mean Silhouette 0.61 compared to TCLUST. After the trimming, cluster plot of PAM seems to be more explainable with 3 distinct cluster.

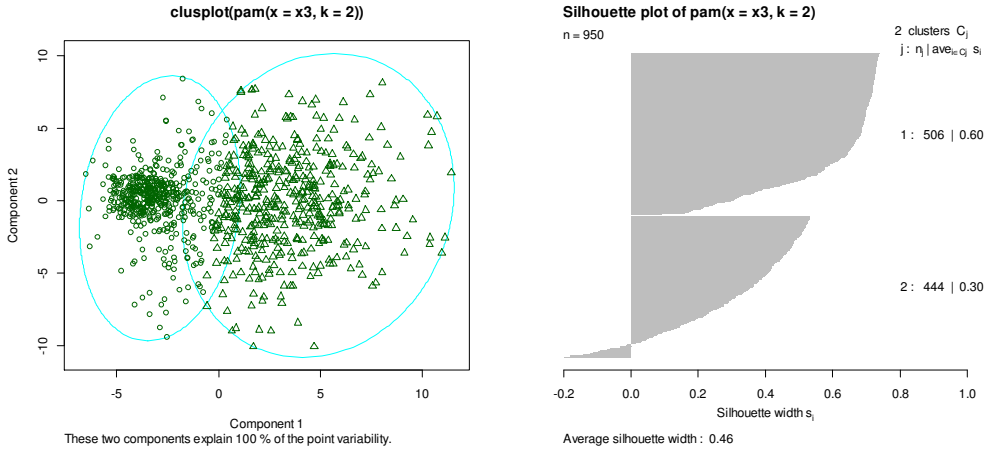


Figure 3: Graphical display of modified PAM to demonstrated the mean Silhouettee.

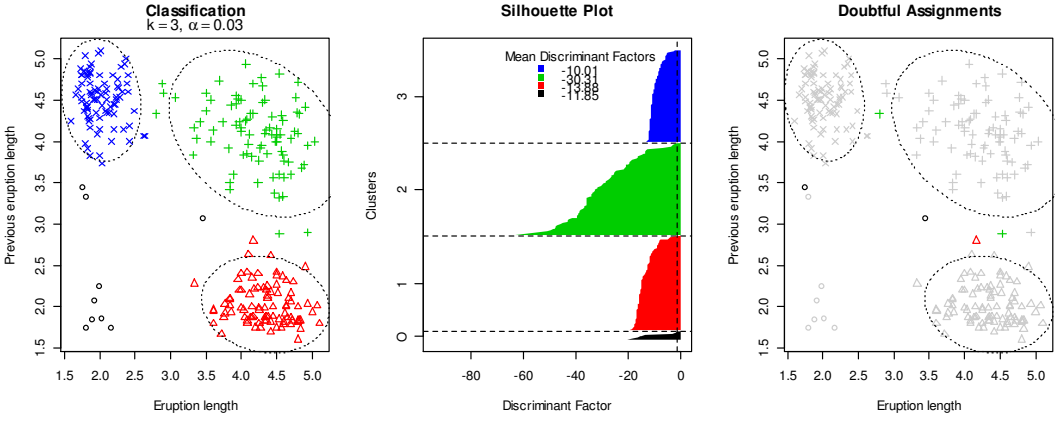


Figure 4: Graphical displays of eruption length data for TCLUCT cluster solution with k=2 and c =50.

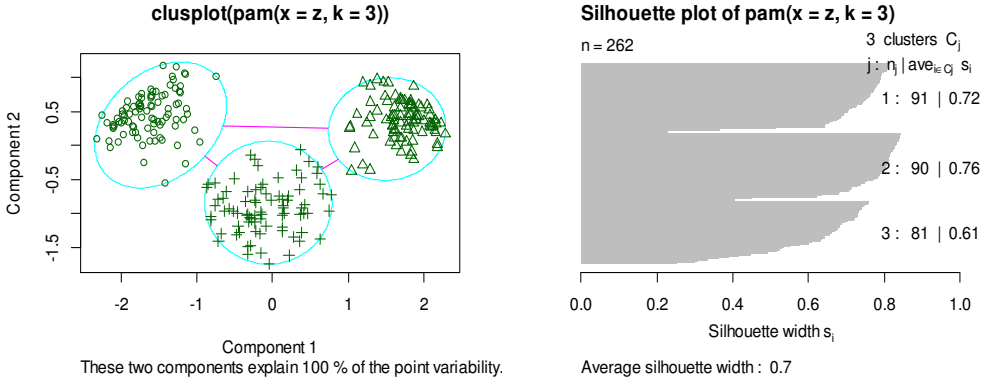


Figure 5: Graphical display of modified PAM to demonstrated the mean Silhouettee for eruption length data

4 Conclusion

Theoretically, the estimation using median is more robust than mean. The same concept is applied with TCLUST and PAM. TCLUST is based on t-kmeans whereas PAM is based on k-median. However when deal with outlier, the existences of abundant outliers give the PAM is less robust compared to TCLUST. TCLUST is better in such the trimming of the outlier playing a big role to execute the analysis. New method has been proposed with modified PAM where we also applied the trimming of the outlier to further analyses using PAM.

Result shows both the simulation studies and real data suggest that modified PAM is better compared to TCLUST. For TCLUST the function of $DF_{(i)}$ is used to calculate the mean Silhouette.

References

[1] Fritz, et.al. "A fast algorithm for robust constrained clustering", unpublish manuscript. 2011

[2] Fritz, L.A, and Mayo-Iscar. "TCLUST: An R package for a Trimming Approach to Cluster Analysis", Preprint available at http://cran.r-project.org/web/packages/tclust/vignettes/tclust.pdf, May 4, 2011

[3] Garcia et.al. "A review of robust clustering methods". *Advances in Data Analysis and Classification*, 4(2-3), 89-109. 2010

[4] Garcia et.al. "Exporing the number of groups in robust model-based clustering", University of Valladolid, Spain, preprint available at http://www.eio.uva.es/infor/personas/langel.html, 2011

[5] Garcia et.al. "Robustness properties of k-means and trimmed k-means", J.Amer.Statist.Assoc., 94, 956-969. 2010

[6] Garcia et.al. "Trimming tools in exploratory data analysis", J.Comput. Graph. Statist., 12, 434-449. 2003

[7] Hathaway, R.J. "A Constrained formulation of maximum likelihood estimator for normal mixture distributions", Ann. Statist, 13, 795-800. 1985

[8] Scott. et.al. "Custering based on likelihood ratio criteria", Biometrics, 27, 387-397. 1971