Communications for Statistical Applications and Methods
2015, Vol. 22, No. 1, 55–67

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.1.055
Print ISSN 2287-7843 / Online ISSN 2383-4757

Variable Selection and Outlier Detection for Automated K-means Clustering

Sung-Soo Kim[1,a]

[a]Department of Information Statistics, Korea National Open University, Korea

Abstract

An important problem in cluster analysis is the selection of variables that define cluster structure that also eliminate noisy variables that mask cluster structure; in addition, outlier detection is a fundamental task for cluster analysis. Here we provide an automated K-means clustering process combined with variable selection and outlier identification. The Automated K-means clustering procedure consists of three processes: (i) automatically calculating the cluster number and initial cluster center whenever a new variable is added, (ii) identifying outliers for each cluster depending on used variables, (iii) selecting variables defining cluster structure in a forward manner. To select variables, we applied VS-KM (variable-selection heuristic for K-means clustering) procedure (Brusco and Cradit, 2001). To identify outliers, we used a hybrid approach combining a clustering based approach and distance based approach. Simulation results indicate that the proposed automated K-means clustering procedure is effective to select variables and identify outliers. The implemented R program can be obtained at http://www.knou.ac.kr/~sskim/SVOKmeans.r.

Keywords: Automated K-means clustering, variable selection, outlier detecting, VS-KM, adjusted rand index, Mahalanobis distance.

1. Introduction

The K-means clustering method assigns a case to the cluster for which the distance to the smallest cluster mean. It starts with an initial partition with user-given k clusters, and repeatedly reassigns cases to the closest cluster's center and updates partitions. The K-means clustering does not require computation of all possible pairwise distances of cases and only requires looping steps of calculating centroids of new clusters and reassigning cases to closest clusters; therefore, it is easily applicable to very large data sets and is widely used in data mining. However, K-means clustering has two crucial problems - the number of clusters and initial centroids of clusters. The number of clusters should be provided before clustering and the K-means cluster solution is dependent on initial centroids. To be applicable to very large data sets, one of the possible solutions for these problems is to apply the results of prior hierarchical clustering methods based on random small samples. This process is consisted of the following two-stage clustering procedure (Kim, 2009). The first stage is to run hierarchical clusters to obtain the number of clusters and cluster centroids based on random samples, and the second stage is to run nonhierarchical K-means clustering using first stage results.

In clustering analysis, it has been frequently observed that only a limited subset of variables is valuable to defined the cluster structure (Brusco and Cradit, 2001). Furthermore, the incorporation

This paper was supported by Korea National Open University Research Fund in 2012.
[1] Department of Information Statistics, Dongsung-dong, Jongno-gu, Seoul 110-791, Korea. E-mail: sskim@knou.ac.kr

Published 31 January 2015 / journal homepage: http://csam.or.kr

of masking variables which do not define cluster structure may complicate or obscure the recovery of cluster structure during hierarchical or nonhierarchical cluster analysis (Milligan, 1980, 1989; Fowlkes and Mallows, 1983; Brusco and Cradit, 2001). For the general approaches to identify masking variables in cluster analysis, see Gnanadesikan *et al.* (1995) and Brusco and Cradit (2001). For the variable selection in K-means clustering, Carmone *et al.* (1999) proposed a graphical variable-selection procedure, named HINoV (heuristic identification of noisy variables) based on the adjusted Rand (1971) index of Hubert and Arabie (1985). Brusco and Cradit (2001) proposed a heuristic variable-selection procedure, VS-KM (variable-selection heuristic for K-means clustering). This procedure utilizes the adjusted Rand index like HINoV, and adds variables in a forward manner as well as uses between-cluster and total sum-of-squares information.

K-means clustering is sensitive to outliers. Outliers are the set of objects that are considerably dissimilar from the remainder of the data (Jayakumar and Thomas, 2013) and can be considered as data points that do not conform to normal points that characterize the data set (Pamula *et al.*, 2011). As a clustering point of view, outliers can be defined as small clusters that are far from most of points (Jiang *et al.*, 2001). Detecting outliers is an important task with a direct application in a wide variety of application domains such as fraud detection, stock market analysis, intrusion detection, and marketing (Pamula *et al.*, 2011; Jayakumar and Thomas, 2013). When doing K-means clustering, the task of outlier detection should be performed for the results to be stable as well as to detect outliers. In this paper we will provide automated K-means clustering procedure combined with variable selection and outlier detection. Automated K-means clustering consists of the following functions.

1) It automatically determines the number of clusters and initial centroids of clusters whenever a new variable is added.

2) It automatically selects a subset of variables which are valuable to define cluster structure and effective to reduce the influence of variables with minimal contribution to the cluster structure.

3) It identifies outliers whenever a new variable is added in a forward manner.

When terminating automated K-means clustering, we can select subset of variables to find cluster structure and detect outliers. Also, through 3), we can reveal the relationship between the variables and outliers. We will review some approaches to select variables and to detect outliers in Section 2, and provide the detailed automated K-means clustering process in Section 3. R implementation and simulation results are provided in Section 4, and concluding remarks are provided in Section 5.

2. Review of Variable Selection and Outlier Detection for K-means Clustering

Here we will provide a description of VS-KM (variable-selection heuristic for K-means clustering) and we will describe the outlier detection method for automated K-means clustering.

2.1. Variable-Selection heuristic

Carmone *et al.* (1999) proposed a variable selection method, HINoV (heuristic identification of noisy variables), based on the principle that a good measure of actual recovery might be useful to guide the selection of cluster variables to include in the analysis. They used the adjusted Rands (1971) index by Hubert and Arabie (1985) to measure the agreement of partitions.

Brusco and Cradit (2001) developed a heuristic variable-selection procedure, VS-KM(variable-selection heuristic for K-means clustering) that builds on the strengths of HINoV and adds variables

in a forward manner as well as uses information about the between-cluster and total-sum-of-squares, similar to the Fowlkes *et al.* (1988) method. This procedure begins by selecting first two variables considering the adjusted Rand index and the ratio of between cluster sum-of squares to the total sum-of-squares, and then adds a variable in a forward manner. For detailed description of VS-KM, see Brusco and Cradit (2001).

In VS-KM process, the number of fixed clusters for a K-means partition should be given in advance and the process of adding variables is applied for fixed number of clusters. However, the number of clusters is generally unknown and can vary depending on used input variables. Hence, it is recommended that the number of cluster be decided automatically for each variable instead of fixed number of clusters. In the VS-KM method, the first two selected variables play an important role since variables with a similar shape of clusters of the first two variables are added independently systematically. VS-KM computes the ratio of the between cluster sum-of-squares to the total sum-of squares for all the possible pairs of partition to avoid the initial selection of masking two variables which have a large Rand index due to a high correlation. However, it is recommended that we choose several pairs of variables which have the highest adjusted index and only compute the ratio of these selected variable since the main index of selecting variables is an adjusted Rand index.

The VS-KM method is effective to select significant variables. However, it still has two crucial problems of K-means clustering-the number of clusters and initial centroids. The number of clusters and initial centroids can vary according to selected variables; therefore, it is recommended that we combine the variable-selection process with the automatic decision of the number of clusters whenever a new variable is added to the previously selected variables instead of continuing the variable-selection process with a fixed number of clusters.

2.2. Outlier detection for K-means clustering

An outlier is an observation that deviates from other observations as to arouse suspicions that it was generated by a different mechanism (Hawkins, 1980) and is also defined as a noisy observation that does not fit to the assumed model that generated the data (Hautamaki *et al.*, 2005). Outlier detection is an important task in a wide variety of application domains such as credit card fraud detection, medical anomaly detection, and industrial damage detection. There have been many approaches to detect outliers, which are categorized as statistical tests based on Mahalanobis distance, depth-based approaches, deviation-based approaches, distance-based approaches, density-based approaches and clustering-based approaches. For a brief review of these approaches, refer to Pamula *et al.* (2011), Jayakumar and Thomas (2013) and Kriegel *et al.* (2010). However, all these approaches to detect outliers are based on fixed user-input variables. Outliers can be dependent on the input variables; consequently, it is helpful to show the relationships between outliers and selected variables if we can detect outliers while we systematically add variables.

Outliers can be categorized as two parts, global and local outliers. Global outlier means that it is far isolated from the center of data set, *i.e.*, observation inconsistent with rest of the data set. Local outlier is an observation inconsistent with its neighborhoods. It is noted that global outliers are not always outliers since they can be considered as another cluster if there are some cases over some threshold compared to the total number of cases. Our focus is to find local outliers. To detect local outliers, we adopt hybrid approaches that combine clustering-based approaches and distance-based approaches similar to the approaches used by Pachgade and Dhande (2012). The data set is partitioned into Kclusters using K-means clustering in hybrid approaches; subsequently, the (robust) Mahalanobis distance is calculated with each instance for each cluster. Since squared (robust) Mahalanobis distance is asymptotically distributed as χ^2-distribution and all points whose distance is larger than threshold,

for example $\chi^2(0.975)$ or $\chi^2(0.99)$, will be declared as "potential outlier". For the detailed use of Mahalanobis distance and Robust Mahalanobis distance to detect outliers, see Rousseeuw and Leroy (1987), Rousseeuw and van Zomeren (1990), Rocke and Woodruff (1996) and Bartkowiak (2005). Also small clusters (*i.e.*, clusters containing significantly less points than other clusters) are considered potential outliers (Jayakumar and Thomas, 2013).

3. Variable Selection and Outlier Detection for Automated K-means Clustering

The crucial problems of K-means clustering decide the number of clusters and initial centroids of clusters. A variety of suggested methods may be helpful in particular situations. The results of K-means clustering against the number of K values (or previous application of the hierarchical clustering methods) can be used to solve problems. For details, see Everitt *et al.* (2001). It is very helpful for users to explore the data in the data mining approach if the number of clusters can be provided without the previous user-handling task.

Many approaches for the selection of the number of clusters in K-means clustering have been tried. Recently Kim (2009) proposed a semi-automated K-means clustering procedure to determine the number of clusters and initial centroids and Kim (2012) also combined a VS-KM procedure to select variables. A semi-automated K-means clustering procedure can be described as: This procedure selects random sample from a large data sets and applies Ward's (1963) hierarchical method and Mojena's (1977) rule to determine the number of clusters. After repeating this step several times, the number of clusters and initial centroids are determined and K-means clustering is proceeded using full data sets. When this procedure is combined with the VS-KM method, we can see the role of variables in a forward manner in K-means clustering since a new result of K-means clustering is obtained whenever a new variables is added to the existing input variables and we can then get the selected variables in the final step. It is noted that K-means clustering is sensitive to outliers, and outliers can also be dependent on the selected variables in K-means clustering. Hence if outliers can be identified while in the process of variable selection, automated K-means clustering can provide information about variable selection and outlier detection.

Our focus is on variable selection and outlier detection in K-means clustering; therefore, we use the following two-stage K-means clustering procedure instead of semi-automated K-means clustering. Two-stage K-means clustering procedure (we refer to this procedure as TStep-KM) can be stated as:

(Two- Stage K-means clustering procedure: TStep-KM)

Step 1. Select random sample from full data sets using simple random sampling or full data sets.

Step 2. Apply Ward's clustering method and decide the number of clusters using Mojena's Rule and obtain cluster centers

Step 3. Run K-means clustering using full data sets.

Step 1 of this procedure can be effective when data sets are large, as used by many approaches (Banfield and Raftery, 1993; Brusco and Credit, 2001; Wehrens *et al.*, 2004). In Step 2, we applied Ward's clustering method and used Mojena's Rule (Mojena, 1977; Mojena *et al.*, 1980) to decide the number of clusters. Mojena's Rule can be stated as:

(Mojena's Rule)

Generate $\underline{h} = (h_1, h_2, \ldots, h_{n-1})$ where h_j is the minimum Euclidean error sum of squares at which fusion takes place in stage j and n is the number of objects in the data matrix. We determine the number of clusters in stage satisfying $h_{j+1} > \bar{h} + ks_h$ where $\bar{h}$ and s_h are, respectively, the mean and standard deviation of $n-1$ values, and k is the standard deviate value.

To decide the number of clusters, we can use the other decision rules. For general studies of choosing the number of clusters, see Milligan and Cooper (1985). Now we describe detailed procedure of variable selection and outlier detection for K-means clustering as:

(Variable selection and outlier detection for K-means clustering)

Step 0. Initialize - choose the options of data transformation, sampling, Mojena's constant value

Step 1. Run TStep-KM procedure and develop a partition, $\boldsymbol{p}_j$, of clusters using only variable j, for $j = 1, 2, \ldots, D$ where D is the number of variables.

Step 2. Compute the adjusted Rand index for all $D(D-1)/2$ pairs of partitions $\boldsymbol{p}_j$ and $\boldsymbol{p}_k$ ($j = 1, 2, \ldots, D-1$, $k = j+1, \ldots, D$) and find two variables showing highest top 5 adjusted Rand index.

Step 3. Compute the ratio of between cluster sum-of-squares to the total sum-of-squares, q_{jk}, only of two variables showing highest top 5 adjusted Rand index, where $q_{jk} = q_{kj}$ for $j = 1, 2, \ldots, D-1$ and $k = j+1, \ldots, D$.

Step 4. Select two variables which have the highest ratio of between cluster sum-of-squares to the total sum-of-squares. Let j' and k' denote two variables and set $\boldsymbol{S} = \boldsymbol{S} \bigcup \{j', k'\}$ and $\boldsymbol{U} = \boldsymbol{U} - \{j', k'\}$, where is the set of variables selected for inclusion in the cluster analysis and is the set of unselected variables.

Step 5. Using selected variables, run TStep-KM procedure and develop a partition $\boldsymbol{y}$ that defines a partition developed using variables $j \in \boldsymbol{S}$ and apply outlier detection procedure.

Step 6. For each unselected variable $j \in \boldsymbol{U}$, run K-means clustering and compute the adjusted Rand index between selected variables in Step 5 and unselected variables.

Step 7. Let $\lambda = \mathrm{Max}_{j \in U}(G_j)$. If $\lambda < G_{\min}$, or $\lambda < \eta \cdot G_{fac}$, then go to Step 8. Otherwise, let j' denote the variable for which $G_{j'} = \lambda$, set $\eta = \lambda$, and set $\boldsymbol{S} = \boldsymbol{S} \bigcup \{j'\}$, $\boldsymbol{U} = \boldsymbol{U} - \{j'\}$. If $\boldsymbol{U} = \varnothing$, then go to Step 8. Otherwise go to Step 5.

Step 8. Variables in $\boldsymbol{S}$ are selected for inclusion and variables in $\boldsymbol{U}$ are discarded. Run TStep-KM procedure and apply outlier detection procedure using only the variables in $\boldsymbol{S}$.

The main difference between VS-KM method and our proposed procedure is that VS-KM starts with user-supplied fixed cluster number while the cluster number in our procedure is determined automatically. Therefore the cluster number can be changed when a new variable is added in a forward manner. Through Step 5, whenever a new variable is added, we can see the results of new K-means clustering, and potential outliers are supplied. Hence, we can explore relationships between input variables and outliers. In the variable selection process, we have a question on how to deal with potential outliers. We process the variable selection process without removing potential outliers detected

in earlier steps. Finding outliers is an exploratory work; therefore, we should carefully check the outliers again and decide how to deal with them after finding potential outliers. The effect of outliers in the process of selecting variables can be another work and we hope to provide some further research at a later time.

4. R Implementation and Experimental Results

We implemented R program (http://www.knou.ac.kr/~sskim/SVOKmeans.r). We use well known Iris data sets and simulation data sets to show the effectiveness of our procedure for variable selection and outlier detection. Simulation data sets are generated using the R package "clusterGeneration".

4.1. Iris data sets

We provide the R implementation process using the Iris data sets to show the performance of our proposed procedure since it is known to many users and easy to understand (Arai and Barakbah, 2007). Iris data sets have three classes of Iris flowers (Setosa, Versicolor and Virginica) with 4 variables (Sepal Length, Sepal Width, Petal Length and Petal Width). For simplicity the three classes are named as (1, 2, 3) and 4 variables are ordered as (1, 2, 3, 4) in the results.

<R Console 1> Initial step of selecting options

```
> iris.data = iris[,-5]
> iris.mem = iris[,5]
> write(iris.mem, "c:/data/vsod/iris.mem", ncolumns=10)
> source("c:/vskm/SVOKmeans.r")
> SVOKmeans(data=iris.data)
------------------------------------------------------------
 Step 0-1: Standardize Variables ?
    1. 0-1 Transform    2. Z-Score    3. Raw Data
 ------------------------------------------------------------
    Select(default:1) : 3
 ------------------------------------------------------------
 Step 0-2: 1. Sampling 2. Full Data : # of data= 150
 ------------------------------------------------------------
    Select(default=1): 2
-------------------------------------------------------------------------------
 Step 0-3: Mojena's k for deciding the number of clusters(def=1.25): 1.25
-------------------------------------------------------------------------------
 Step 0-4: Parameter for Outlier
          - Mahal. distance(1=default) Robust Mahal. distance(2) : 1
          - Chisq-quantile for outlier(def.=0.975): 0.975
          - Ratio of cluster small size for outlier(def=0.03): 0.03
-------------------------------------------------------------------------------
```

<R Console 1> shows the initial step of selecting options. We can choose options of data transformation in Step 0-1, Sampling or Full data in Step 0-2, Mojena's K value in Step 0-3 and Parameters for outliers in Step 0-4. If we choose large Mojena's value, then it tends to choose a small number of clusters. Mojena *et al.* (1980) recommended the value of $k \leq 2.5$, and Milligan and Cooper (1985) recommended the value of $k = 1.25$. As parameters for to detect detecting outliers, we provide options of choosing Mahalanobis or Robust Mahalanobis distance, chi-square quantile for the threshold of outliers and the ratio of small cluster size for determining cluster-based outliers. When

we choose Robust Mahalanobis distance, we provide an option to select robust method to compute robust multivariate location and scale estimate(<R Console 3>).

<R Console 2> shows the repeating process of variable selection and outlier detection. The stopping criteria for variable selection in Step 7 were based on parameter values of $G_{\min} = 0.05$ and $G_{fac} = 0.5$. Step 6 shows the adjusted Rand indexes between pre-selected variables $(3, 4)$ and unselected variables (1), (2). We can see that variables 1 and 2 are added in a forward manner along with the clustering results and potential outliers whenever a variable is added systematically consequently, we can find some relationships between variables and outliers. SSB/SST stands for the ratio of between-cluster sum of squares to the total sum of squares. High value of SSB/SST means that the performance of clustering results for classification is good. From the results of SSB/SST, we can find that the performance of the first two selected variables $(3, 4)$ is higher than the $(3, 4, 1)$ and $(3, 4, 1, 2)$ variables. Last, the repeating result shows that selected variables are $(3, 4, 1, 2)$, and local potential outliers are 4 cases, and global outliers are 6 cases. Our focus is on the local outliers since they are identified after K-means clustering. Global outliers provide insight into the identified outliers.

<R Console 2> Procedure of variable selection and outlier detection : Step 5–7

```
Step 5 : Results of first selected var's
        Selected Var's = ( 3 4 )
        UnSelected Vars = ( 1 2 )
        Number of Cluster =  3
        Cluster Sizes =  50 52 48
        Potential Outliers(Local) =  44 99 25
        Mahal. Distance(Local) =  3.376 3.144 2.86
        Number of Outliers(Local) =  3
        Potential Outliers(Global) =  115 135 142
        Mahal. Distance(Global) =  3.214 2.938 2.745
        Outlier Cutoff =  2.716
        Number of Outliers(Global) =  3
        SSB/SST     =  0.9431

 Step 6 : Adjusted Rand Index between Y and Unselected var's
[1] 0.3968 0.1594 0.0000 0.0000

 Step 7 : Repeating procedure for adding var's
        Selected Var's = ( 3 4 1 )
        UnSelected Vars = ( 2 )
        Number of Cluster =  3
        Cluster Sizes =  50 62 38
        Potential Outliers(Local) =  115 44 99 135 15
        Mahal. Distance(Local) =  3.842 3.505 3.328 3.318 3.184
        Number of Outliers(Local) =  5
        Potential Outliers(Global) =  135 115 142 107
        Mahal. Distance(Global) =  3.39 3.356 3.229 3.133
        Outlier Cutoff =  3.058
        Number of Outliers(Global) =  4
        SSB/SST     =  0.903

  Repeat(Step5-6) : adj.max =  0.2525  which =( 2 )
        Selected Var's = ( 3 4 1 2 )
        UnSelected Vars = ( - )
        Number of Cluster =  3
```

```
Cluster Sizes =  50 62 38
Potential Outliers(Local) =  115 42 44 132
Mahal. Distance(Local) =  3.919 3.511 3.509 3.34
Number of Outliers(Local) =  4
Potential Outliers(Global) =  132 135 118 142 42 115
Mahal. Distance(Global) =  3.62 3.589 3.58 3.527 3.38 3.378
Outlier Cutoff =  3.338
Number of Outliers(Global) =  6
SSB/SST  =  0.8843
```

<R Console 3> Last Results using robust Mahalanobis distance

```
Step 0-4: Parameter for Outlier
        - Mahal. distance(1=default) Robust Mahal. distance(2) : 2
          Robust Function : cov.mve(1=def) cov.rob(2)  cov.mcd(3)  covMcd(4) : 1
        - Chisq-quantile for outlier(def.=0.975): 0.975
        - Ratio of cluster small size for outlier(def=0.03): 0.03

Step 8 : Last K-Means Results Using Selected Variables
        Selected Vars = ( 3 4 1 2 )
        UnSelected Vars = ( - )
        Number of Cluster =  3
        Cluster Size =  50 62 38
        Potential Outliers(Local) =  132 118 119 123 106 136 44 108 42 131 115 126 24 110 130 33 23 99
        Robust Mahal. Distance(Local) =  8.598 8.337 7.919 7.254 6.415 5.277 5.259 5.201 4.977 4.811 4.5
        Number of Outliers(Local) =  21
        Potential Outliers(Global) =  132 135 118 115 142 123 108 42 146 130 145 126 137 101 136 106
        Robust Mahal. Distance(Global) =  4.832 4.718 4.701 4.629 4.506 4.018 3.985 3.967 3.956 3.855
        Outlier Cutoff =  3.338
        Number of Outliers(Global) =  16
        SSB/SST  =  0.8843
```

<R Console 3> shows the options of robust Mahalanobis distance and outlier results. R package chemometrics (Filzmoser and Varmuza, 2013) contain a function Moutlier to calculate the Mahalanobis distance and robust Mahalanobis distance. However, we provide a function to calculate robust Mahalanobis distance using various algorithms since the Moutlier function is unstable depending on the data sets used. From the result of outliers, we find that robust Mahalanobis distance is more sensitive than the Mahalanobis distance to detect outliers. For reference, we hope to try other real data sets such as Hawkins-Bradu-Kass data, Modified Wood Gravity data and Stackloss data which are provided in R systems (In R systems, the name of data sets are hbk, wood, stackloss respectively) and compare the results are provided by Bartkowiak (2005).

4.2. Simulation data sets

Simulation data sets are generated using the R package "clusterGeneration" developed by Qui and Joe (2006a,b). We produced 18 ($3 \times 2 \times 3$) data sets considering the following 4 factors. Number of each cluster cases ranges from 2000 to 4000.

A. Number of clusters = 3, 4, 5

B. Cluster Separation = 0.21, 0.40

C. Number of true variables (noisy variables) = 4(2), 6(2), 8(2)

D. Number of Outliers = 10

The R script for generating simulated clustering data is:

```
library(clusterGeneration)
# cluster=4, Sep=0.4, Var=6, Noisy=2, outliers=10
genRandomClust(numClust=4, sepVal=0.4,
      numNonNoisy=6, numNoisy=2, numOutlier=10,
      numReplicate=1, fileName="c:/data/vsod/cls_v462_10", clustszind=2, rangeN=c(2000, 4000) )
```

The size of generated simulation data sets in cluster ranges from 2000 to 4000; therefore, it is recommended to use sampling data to find initial cluster center for K-means and determine the number of clusters. The results for 18 simulation data sets give similar results, so we only provide the results using one of the generated data sets produced by the listed R scripts.

<R Console 4> Initial step of selecting options for simulated clustering data

```
> cls.v462 = read.table("c:/data/vsod/cls_v462_10_1.dat", header=T)
> head(cls.v462)
          x1        x2        x3        x4         x5         x6         x7         x8
1 11.849928  2.480717  4.800595  8.173042 -15.822687  0.4944058  4.6258837 -8.147684
2  2.961294 -3.036412  4.780150  7.948143   2.994635  3.1003792 -0.2787702  3.840099
3 -3.946592 -4.093422 -7.076713  7.844510  -1.806446  3.6114040 -4.9670648 13.333173
4  4.536866  3.276747  8.483979 16.190117   6.907191  3.0973701  1.9950518 -2.288880
5  3.780257  6.959286  5.364083  5.009962   5.469241 -5.3185041  8.2094895  7.256754
6 -2.916995 -2.195306 -9.761634  2.418568  -9.356987  0.1854160 -6.8012440  1.246606
> SVOKmeans(data=cls.v462)
 ----------------------------------------------------------
 Step 0-1: Standardize Variables ?
   1. 0-1 Transform   2. Z-Score   3. Raw Data
 ----------------------------------------------------------
   Select(default:1) : 1
 ----------------------------------------------------------
 Step 0-2: 1. Sampling 2. Full Data : # of data= 10777
 ----------------------------------------------------------
   Select(default=1): 1
 Type Sampling Rate (10-100%, def=10%)   : 10
-------------------------------------------------------------------------------
 Step 0-3: Mojena's k for deciding the number of clusters(def=1.25): 2.5
-------------------------------------------------------------------------------
 Step 0-4: Parameter for Outlier
         - Mahal. distance(1=default) Robust Mahal. distance(2) : 1
         - Chisq-quantile for outlier(def.=0.975): 0.995
         - Ratio of cluster small size for outlier(def=0.03): 0.003
-------------------------------------------------------------------------------
```

<R Console 4> shows the initial step of selecting 0-1 transformation, sampling rate 10%, Mojena's value 2.5 and quantile value 0.995 to detect outliers.

<R Console 5> shows the last results of simulated clustering data. Here selected variables are $(3, 5, 1, 2, 7)$ in the order of adding variables to the first two selected variables $(3, 5)$ and identified cluster-based potential outliers are 44 cases that covered 10 outliers in simulated clustering data. Identified outliers and Mahalanobis distances are provided according to the descending order. The result of Step 8 show that it interesting to find that the outliers in simulation data sets are identified as

having the largest 10 Mahalanobis distances. For reference, we provide the adjusted Rand index and confusion matrix with objects removing potential outliers from two clustering data sets. In this result, adjusted Rand index and confusion matrix shows a perfect coincidence between two clustering data.

In the variable selection procedure, we note that the inclusion of noisy variables can cause serious recovery problems even when all true variables are contained in selected variables, and as long as the first true pair variables are selected, other true variables can be omitted without significant degradation in cluster recovery (Milligan, 1985; Brusco and Cradit, 2001). We can see that noisy variables (4, 8) are not selected and that the value of adjusted Rand indexes is high (even when true variable 6 is not selected), which means that the process shows a high performance of cluster recovery. In identifying outliers, we can see that there is some tendency to detect more outliers than outliers in the simulated data sets. After finding potential outliers, it is better to check again why they are identified as outliers. Generally, finding more outliers than in the data sets is not a serious problem, since it can be adjusted by changing the cutoff value.

We have not listed the step by step results; however if we check the results as in <R Console 2>, it will be more helpful to decide which variables to select as well as to explore the relation between variables and outliers. Here, we listed only the results of potential outliers using Mahalanobis distance, but we recommend identifying outliers using robust Mahalanobis distance as well as recommend to run with other simulation data sets.

<R Console 5> Simulation results comparing original data and automated K-means clustering

```
Step 8 : Last K-Means Results Using Selected Variables
          Selected Vars = ( 3 5 1 2 7 )
          UnSelected Vars = ( 4 6 8 )
          Number of Cluster =  4
          Cluster Size =  2278 2620 2585 3294
          Potential Outliers(Local) =  10769 10770 10772 10774 10776 10777 10768 10773 10775 10771
          Mahal. Distance(Local) =  15.905 15.114 13.905 13.698 13.442 12.276 11.812 10.78 9.646 7.073
          Number of Outliers(Local) =  44
          Potential Outliers(Global) =  10777 10768 10769 10770 10772 10776 10774 10775 2638 9949 6532
          Mahal. Distance(Global) =  10.95 10.602 10.116 9.697 8.115 7.578 6.99 6.814 5.697 5.274 5.222
          Outlier Cutoff =  4.093
          Number of Outliers(Global) =  44
          SSB/SST      =  0.8257
 ---------------------------------------------------------------------------------------
   Adjusted Rand Index ? - Simulated Data(1), Real Data(2), None(3=default) : 1
 ---------------------------------------------------------------------------------------
 --- original group member file : c:/data/vsod/cls_v462_10_1
Read 10777 items
Read 2 items

 ---------- < Simulation Data by clusterGeneration > ----------
 Original True Var's = ( 1 2 3 5 6 7 )
 Original Noisy Var's = ( 4 8 )
 Original Cluster Size =  10 2618 3290 2583 2276
 Original Outliers =  10768 10769 10770 10771 10772 10773 10774 10775 10776 10777

 ---------- < Automated K-Means Clustering > ----------
 Selected Var's = ( 1 2 3 5 7 )
 UnSelected Var's = ( 4 6 8 )
 Cluster Size =  2278 2620 2585 3294
```

```
Potential Outliers(Local) =    10777 10776 10775 10774 10773 10772 10771 10770 10769 10768 10540 10498
Number of Potential Outliers(Local) =    44
Potential Outliers(Global) =    10777 10776 10775 10774 10772 10770 10769 10768 10564 10540 10127 9949
Number of Potential Outliers(Global) =    44
Adjusted Rand Index =    1

Confusion Matrix
                cluster.id
origin.cluster      1    2    3    4
             1      0 2606    0    0
             2      0    0    0 3277
             3      0    0 2575    0
             4   2275    0    0    0
```

5. Concluding Remarks

We presented the automated K-means clustering procedure combined with selecting variables and identifying outliers. For variable selection process, we applied VS-KM method proposed by Brusco and Cradit (2001). VS-KM method is a heuristic algorithm to select subsets of variables for inclusion in a K-means cluster analysis. Hence it starts with a user-given fixed number of clusters and proceeds to select variables with keeping the initial fixed number of clusters. However, it is recommended to determine the number of clusters whenever a new variable is added to pre-selected variables since a suitable number of clusters in K-means cluster analysis depends on selected variables.

The proposed automated K-means clustering procedure combines the VS-KM algorithm proposed by Brusco and Cradit (2001) with a semi-automated K-Means procedure proposed by Kim (2009, 2012), and the identification of outliers is also applied to the variable selection process. Through the automated K-means clustering process shown in <R Console 2>, we can see the results of variable selection and identification of outliers systematically whenever a new variable is added in a forward manner.

We provided the R scripts (www.knou.ac.kr/~sskim/SVOKmeans.r). In R implementation, we used the Ward's method and applied the Mojena's Rule to determine the number of clusters. However, there are many approaches to determine the number of clusters such as model-based clustering analysis (Banfield and Raftery, 1993; Fraley and Raftery, 1998), Gap approach (Tibshirani *et al.*, 2001). We hope that some researchers can implement these methods to supply results to users. To detect potential outliers, we used a hybrid approach that combines a clustering based approach and distance based approach using (robust) Mahalanobis distance. There are many other methods to detect multiple outliers. We also hope these methods can be implemented. In the variable selection process, we proceed variable selection without removing potential outliers. We hope further works show the effect of outliers in the process of selecting variables. In our implementation, we need several options as shown in <R Console 1>. The GUI approach is needed for the easier handling of the options and represents a future work for a practical application for users. If some errors in R scripts are found, we hope notification will be given to the author (sskim@knou.ac.kr).

References

Arai, K. and Barakbah, A. R. (2007). *Hierarchical K-means: an algorithm for centroids initialization for K-means*, Reports of the Faculty of Science and Engineering, Saga University, **36**, 25–31.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics*, **49**, 803–821.

Bartkowiak, A. (2005). Robust Mahalanobis distances obtained using the 'Multout'; and "Fast-mcd' Methods, *Biocybernetics and Biomedical Engineering*, **25**, 7–21.

Brusco, M. J. and Cradit, J. D. (2001). A variable-selection heuristic for K-means clustering, *Psychometrika*, **66**, 249–270.

Carmone, F. J., Kara, A. and Maxwell, S. (1999). HINoV; A new model to improve market segmentation by identifying noisy variables, *Journal of Marketing Research*, **36**, 501–509.

Everitt, B. S., Landau, S. and Leese, M. (2001). *Cluster Analysis*, Arnold.

Filzmoser, P. and Varmuza, K. (2013). Package Chemometrics. Documentation available at: http:// cran.r- project.org/web/packages/chemometrics/index.html.

Fowlkes, E. B., Gnanadesikan, R. and Kettenring, J. R. (1988). Variable selection in clustering, *Journal of Classification*, **5**, 205–228.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings (with comments and rejoinder), *Journal of the American Statistical Association*, **78**, 553–584.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering methods? Answers via model-based cluster analysis, *Computer Journal*, **41**, 578–588.

Gnanadesikan, R., Kettenring, J. R. and Tsao, S. L. (1995). Weighting and selection of variables for cluster analysis, *Journal of Classification*, **7**, 271–285.

Hautamaki, V., Cherednichenko, S., Karkkainen, I., Kinnunen, T. and Franti, P. (2005). *Improving K-Means by Outlier Removal*, LNCS Springer, Berlin / Heidelberg, may 2005, 978–987.

Hawkins, D. (1980). *Identifications of Outliers*, Chapman and Hall, London.

Hubert, L. and Arabie, P. (1985). Comparing partitions, *Journal of Classification*, **2**, 193–218.

Jayakumar, G. S. and Thomas, B. J. (2013). A new procedure of clustering based on multivariate outlier detection, *Journal of Data Science*, **11**, 69–84.

Jiang, M. F., Tseng, S. S. and Su, C. M. (2001). Two-phase clustering process for outliers detection, *Pattern Recognition Letters*, **22**, 691–700.

Kim, S. (2009). Automated K-means clustering and R implementation, *The Korean Journal of Applied Statistics*, **22**, 723–733.

Kim, S. (2012). A variable selection procedure for K-means clustering, *The Korean Journal of Applied Statistics*, **25**, 471–483.

Kriegel, H.-P., Kröger, P. and Zimek, A. (2010). Outlier detection techniques, *The 2010 SIAM International Conference on Data Mining*, Available from: https://www.siam.org/meetings/sdm10/ tutorial3.pdf.

Milligan, G. W. (1980). An examination of six types of the effects of error perturbation on fifteen clustering algorithms, *Psychometrika*, **45**, 325–342.

Milligan, G. W. (1985). An algorithm for generating artificial test clusters, *Psychometrika*, **50**, 123–127.

Milligan, G. W. (1989). A validation study of a variable-weighting algorithm for cluster analysis, *Journal of Classification*, **6**, 53–71.

Milligan, G. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, **50**, 159–179.

Mojena, R. (1977). Hierarchical grouping method and stopping rules: An evaluation, *The Computer Journal*, **20**, 359–363.

Mojena, R., Wishart, D. and Andrews, G. B. (1980). Stopping rules for Ward's clustering method, *COMPSTAT*, 426–432.

Pachgade, S. D. and Dhande, S. S. (2012). Outlier detection over data set using cluster-based and distance-based approach, *International Journal of Advanced Research in Computer Science and*

Software Engineering, **2**, 12–16.

Pamula, R., Deka, J. K. and Nandi, S. (2011). An outlier detection method based on clustering, *Second International Conference on Emerging Applications of Information Technology*, 253–256.

Qiu,W.-L. and Joe, H. (2006a). Generation of random clusters with specified degree of separation, *Journal of Classification*, **23**, 315–334.

Qiu, W.-L. and Joe, H. (2006b). Separation index and partial membership for clustering, *Computational, Statistics and Data Analysis*, **50**, 585–603.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods, *Journal of American Statistical Association*, **66**, 846–850.

Rocke, D. M. and Woodruff, D. L. (1996). Identification of outliers in multivariate data, *Journal of the American Statistical Association*, **91**, 1047–1061.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*, John Wiley and Sons, New York.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, **85**, 633–651.

Tibshirani, R., Walther, G. and Hastie, T. (2001). *Estimating the Number of Clusters in a Dataset via the Gap Statistic*, Technical report, Dept of Biostatistics, Stanford University, Available from : http://www-stat.stanford.edu/~tibs/research.html.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function, *Journal of American Statistical Association*, **58**, 236–244.

Wehrens R., Buydens L., Fraley, C. and Raftery, A. (2004). Model-based clustering for image segmentation and large datasets via sampling, *Journal of Classification*, **21**, 231–253.

Received October 28, 2014; Revised December 15, 2014; Accepted January 13, 2015