Multidisciplinary Informatics Journal
Vol. 2, No. 2, December 2019, Pg. 98 – 108

INVESTIGATION ON THE CLUSTERABILITY OF MIXED VARIABLES BASED ON SIMULATED DATASET FROM *SimMultiCorrData* PACKAGE

Norin Rahayu Shamsuddin[1,2] and Nor Idayu Mahat[2]

[1]Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA,
Merbok, 08400, Kedah, Malaysia
[2]School of Quantitative Sciences, Universiti Utara Malaysia, Universiti Utara Malaysia,
06010 Changlun, Kedah, Malaysia
Author Correspondence, e-mail: norinrahayu@uitm.edu.my

Received: 6 September 2019 / Accepted: 5 November 2019 / Published online: 15 December 2019

ABSTRACT

Mixed-type variables gain significant attention in clustering as widely observed in real datasets. The construction of clustering or clusterability of mixed dataset seems hard to form owing to a different scale in data. The current paper assessed the impact of clusters structures that developed from simulation data of mixed data types derived from different distributions through the R package *SimMultiCorrData* that mimics the real-world scenario. This particular package was therefore adopted to acquire mixed variables simulation data and investigated the clusterability of mixed variables in a dataset.

Keywords: Clusterability; *k*-medoids; Mixed-variables; *SimMultiCorrData*.

1. INTRODUCTION

Due to the existence of standard technologies that can be found abundantly these days, it is inevitable to encounter mixed variables. Researchers often find that there are mixed variables in the data matrix such as ordinal, categorical and/or metric variables. It claimed that the mixed-type of variables discovered in biological experiments, clinical reports, customer profiling, gene expression, and market survey.

The mixed form of variables is difficult to manage as the impacts of these variables are significant to the output interpretation, which turns out to be an interesting matter in the analysis of clustering [1, 2]. Furthermore, inaccurate conclusions can be drawn due to the lack

of an appropriate method for the treatment of such variability of variables. As a result, cluster analysis can lead to a faulty process of classification of objects [3]. The best clustering technique for performing cluster analysis, according to [4] is to preserve the original structure of the measured variables [5, 1]. Previous studies had implemented various clustering process, known explicitly as *k*-means and *k*-prototypes, while only a few studies had looked into *k*-medoids. As a result, *k*-medoids proved capable of achieving satisfactory results during object clustering, even though the measured variables were mixed-type [6, 1].

Numerous methods in measuring clusterability have been proposed, and extensive analysis regarding this matter has been discussed in detail by [7]. Sometimes, it is hard to uncover the clusterability since not all dataset consists of meaningful partitions. In this paper, we investigated the presence of clusterability of the artificial dataset generated from *SimMultiCorrData* [8] package from the R programming language through a *k*-medoids algorithm. However, this package is not constructed for clustering purposes.

2. MIXED VARIABLES IN CLUSTERING

It is permitted to change the variables into a single type as it is commonly used. This technique, however, can lead to a loss of information [9, 2]. Meanwhile, conducting a separate cluster analysis can abandon the connection between the variables which can be inappropriate. As for constructing the cluster analysis, it involves mixed variables, and it requires a more significant effort to build the mathematical model that suitable to the problem.

Indeed, these complications have given rise to some interest among the researchers in the discipline of clusters. Moreover, it has become an essential matter as it brings implications to the performance of the internal validity of clustering. Several studies concerning this problem has dealt with, for instance by [10], and several other cited works. The conventional methods used to deal with mixed-type of variables have been carried out by;

(i) changing all variables into a single-type of scale [11, 2],

(ii) conducting distinct cluster analyses for various types of variables, and later combining with the outcomes [12], and

(iii) creating a cluster analysis that associated with mixed-type of variables [13].

The widely used algorithms used by most researchers to cluster the objects with mixed-type of variables are *k*-mean, *k*-prototypes, and *k*-medoids, where *k* indicates the number of clusters desired. Because of its simplicity and convenience to implement and efficiency in the

clustering of massive data sets, *k*-means tend to be popular [14] clustering algorithm. However, *k*-means constricted to metric variable and prone to impacts of outliers. Although attempts to implement k-means on mixed variables have been made, the process of discretization is required. [15].

[16] referred to *k*-prototypes as a combination of *k*-means and *k*-mode that deals with the mixed-type of variables for large dataset that deal with: (a) square Euclidean distance for a metric variable, and (b) the number of mismatches for categorical variables between objects. Nonetheless, the method demands an actual weight to disfavour either categorical or metric variables.

Under the partitioning around medoids (PAM) designed by [17], *k*-medoids is referred to as a variant of *k*-means. It uses the actual objects in the dataset as the center of cluster rather than using the means points as implemented in *k*-means. In addition, k-medoids are less sensitive to outliers than other clustering algorithms in the partitioning cluster. PAM allows the incorporation of various measures such as Euclidean, Manhattan, and Gower's distance which is more suitable for mixed-type of datasets.

2.1 Data Generation Package for Clustering with Mixed Variables in R

Varieties of programming packages are implemented in R for the purpose of generating the clustering simulated data or model. For mixed data, only a handful of packages were made. such as *clustMD* [18], *BNPMIXcluster* [19], and *kamila* [2]. [18] introduced the *clustMD* package – a model-based approach that depends on the combination of latent variables that adhere to the distribution of the Gaussian mixture. *BNPMIXcluster* [19] is associated with the Bayesian non-parametric clustering approach that able to generate a dataset in various sampling. The data of the combination of *k*-means algorithm with both Gaussian and the multinomial mixture had been generated through the *kamila* package [2]. *kamila* does not apply a dummy coding for variables that has more than two levels $g > 2$, and this package offers simple separation of cluster independently between categorical and continuous variables. However, the drawbacks of this package are that it only allows two clusters to overlap.

2.2 *SimMultiCorrData* Package

Since all the packages in Section 2.1 carried out their proposed method using Gaussian multivariate, which limited to generated continuous data, an alternative is required produces simulated data of mixed-type variables. *SimMultiCorrData* [8] package generated dataset with

combination of continuous, count and categorical (ordinal, nominal or binary) variables. This package simulated the continuous variables by implementing the method of moment proposed by Headrick [20] from power method transformation (PMT):

$$y = c_0 + c_1 Z + c_1 Z^2 + \cdots + c_{r-1} Z^{r-1} \quad (1)$$

where $Z \sim iid$ N(0,1), c stands for cumulant and r represents the order method. The continuous mixture variables are drawn from more than one component distribution, as described in terms of the mixture distribution.

Categorical and count variables are generated from inverse cumulative density function (cdf). The ordinal variable data generated through a process of discretizing the standard normal variables at quantiles according to desired marginal distribution [21]. While the count variables have a standard normal variable of uniform distribution, in general, all variables generated from standard normal variables with an imposed of intermediate correlation matrix and the details of these methods are discussed in [8].

3. METHODOLOGY

The *SimMultiCorrData* is used to generate artificial datasets of a mixture of continuous, count, and categorical (ordinal, nominal or binary) variables. The correlation matrix between each variable was generated from the uniform distribution, U(0.25,0.7). The valid_corr function for 'Correlation Method 1' resolute if the matrix is in the boundaries. The dataset of mixed-type variables are as followed:

(a) generate count variables from three distributions:

   (i) Poisson distribution ($\lambda = 2, 6$ and 11); (ii) negative binomial distribution $NB(2, 0.2)$, and (iii) negative binomial $NB(6, 0.8)$ distribution.

(b) constructed the ordinal variables from:

   (i) group 1 with three categories $(p(x) = 0.35; 0.75; 1)$, and (ii) group 2 with four categories $(p(x) = 0.25; 0.5; 0.8; 1)$

(c) formulate continuous variables for normal and nonnormal datapoints from Gaussian $\mathcal{N}(0,1)$, gamma $\Gamma(\alpha = \beta = 10)$, and chi-square (χ_4^2) distributions.

Overall, the artificial dataset consist of ten parameters along with 300 datapoints.

In the current paper, the Gower's dissimilarity from [22] was applied to calculate the distance between objects *i*th and *k*th as given in the following:

$$d_{(x_i{}'x_k)} = \frac{\sum_{j=1}^{p} w_{ikj} d_{ikj} \delta_{ikj}}{\sum_{j=1}^{p} w_{ikj} \delta_{ikj}} \quad (2)$$

An extension to the Gower's dissimilarity index proposed in [4] using the concept of a mean of square distance as in Euclidean distance been applied the paper. The purposes of the extension are to allow for more variability in finding the dissimilarity between objects. We employed *k*-medoids algorithm because of its capability to measure various distance measures. From Equation 2, the extended version of Gower dissimilarity is as follows:

$$d_G = \sqrt{d_{(x_i{}'x_k)}} \quad (4)$$

The *clusterCrit* package construct by Desgraupes [23] is employed to obtain the ICV's values. This package consists of comprehensive internal and external clustering indices. The intCriteria function enabled simultaneous generation results ffrom various indicies. In this paper, we opt for four types of ICVs in determining the appropriate k for the dataset which are Davies-Bouldin (DB), Dunn (D), C-index (ΓC), and Silhouette (S). DB and ΓC should produce a minimum value of index while D and S generate maximum value to show compact and well-separated clusters.

4. FINDINGS AND DISCUSSION

4.1 Preamble Analysis

Based on the 300 generated observations, it was found that the setup simulation had produced outliers and noise for the dataset. In previous study by [24] has carried out the analysis with the existance of outliers. The results shows that the outliers gave significant effect on the outcome of clustering. Hence, the cleaning process was conducted by discarding the outliers using simple-box plot which resulted in only 274 clean observations. In order for the data to be generated, the principal components analysis (PCA) was conducted to visualize the data and to reckon the possible number of clusters as shown in Figure 1.

4.2 Cluster Analysis

The tendency of the clustering was assessed from the graphical information and the statistical approaches such as Hopkins statistics to identify the presence of clustering. Based on Figure

1, a scatterplot from the PCA methods exhibited some significant cluster. The contradicting result of the dataset was suggested from the statistical methods to generate important and meaningful clusters (H=0.3259).

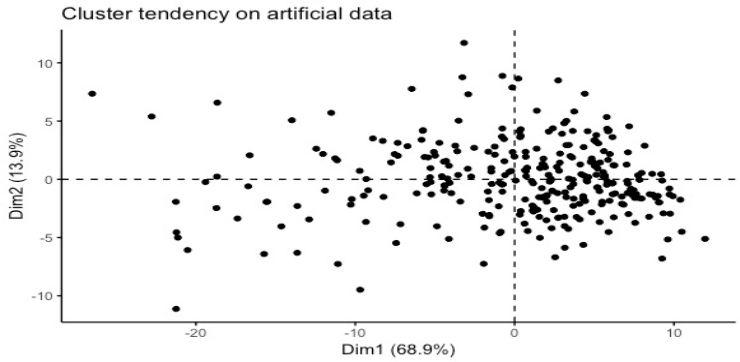

Fig.1. Assessing clustering tendency through PCA approach

By conducting a clustering tendency assessment, it determines the significance of the cluster's structure. Even though there are differences between the graphical and statistical assessments for clustering, we believed that the simulation dataset contains meaningful clusters. Each object in each variable brings up information that constructs clusters. Based on the dissimilarity measures of $d_{(x_i,x_k)}$ and d_G Table 1 tabulates the ICVs values of four indices. The values of ICV are in bold to represent the minimum or maximum value based on the indices.

Table 1. ICVs' value of artificial dataset of $d_{(x_i,x_k)}$ for n = 274

	$d_{(x_i,x_k)}$			
k	DB	ΓC	D	S
2	1.1648	0.2177	0.0730	**0.3293**
3	1.2049	0.1246	0.0578	0.2843
4	1.1740	0.1112	0.0445	0.2462
5	1.1353	0.0857	0.0424	0.2763
6	1.0483	0.0870	0.0764	0.2562
7	1.0722	0.0861	0.0989	0.2604
8	1.0091	0.0568	0.0617	0.2780
9	0.8153	**0.0514**	**0.1006**	0.3018
10	**0.7952**	0.0543	0.0890	0.2795

The Gower's dissimilarity square root d_G helps to decide the dissimilarity between objects. [25], however sound an alarm note about the impact of d_G. While this method helps to show low similarity, the clustering is not well represented. From Table 1 and 2, we have identified *C* at which the indices give its maximum or minimum values. The results from each indice is highlighted in bold for both $d_{(x_i,x_k)}$ and d_G. From the table, D and ΓC indices indicate a suitable *C* for this artificial dataset to be *C* = 9. On contrarily, DB suggested *C* = 9 is the best formation of clusters, while S index indicated *C* = 2 as the best number of clusters.

Theoretically, the mixture component of distribution should be produced equivalent to the number of distribution model if they are well separated. Nonetheless, the mixed types of variables tend to produce overlapping datapoints which tend to give different output of *C*.

Table 2. ICVs' value of artificial dataset d_G for $n = 274$

	d_G			
k	DB	ΓC	D	S
2	1.2810	0.1971	0.1041	**0.3032**
3	1.3047	0.1116	0.0738	0.2784
4	1.2651	0.0956	0.0800	0.2745
5	1.1685	0.0763	0.0847	0.2658
6	1.0857	0.0767	0.1189	0.2513
7	1.1463	0.0750	0.1427	0.2530
8	1.0819	0.0540	**0.1492**	0.2643
9	1.5207	0.0580	0.1492	0.2367
10	**0.8973**	**0.0502**	0.1492	0.2601

Since it is hard to identify the appropriate *C* for a dataset based on index value, one can opt to look at the dendrogram illustration in Figure 2. When there exist an overlapping of objects between clusters especially in mixed variables, the possibility of discovering the clusters becomes difficult, and this could lead to the poor performance of clustering. It was indeed challenging to maintain their scale throughout the analysis, as it presented misleading results in the validation evaluation. Overall, clustering is the task of categorising a set of objects in a way that the objects are in same group are similar in some sense than to those in other groups.

In other words, datasets with similar variable scaleable to delivery more informative information on clustering [26].

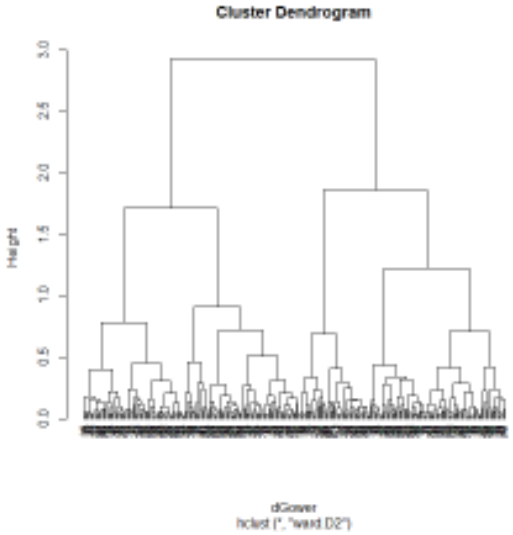

Fig.2. Dendogram of artificial datasets

6. CONCLUSION

In this paper, we investigate the clusterability of mixed types of variables of a dataset through *SimMultiCorrData* package. The artificial dataset from this package created some noise and outliers which contribute on clustering effect. However, the level of overlap between clusters was unable to be measured as due to the simulation data was not purposely created for clustering. We have demonstrated that by 'retaining' the scale of variables through Gower's distance, the formation of clusters does exist. The results on the number of clusters vary for different indices. This indicate that the selection of ICVs play significant role in identifying *C*.

7. REFERENCES

[1] C. Hennig and T. F. Liao (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3), pp. 309–369.

[2] A. H. Foss and M. Markatou (2018). kamila: Clustering Mixed-Type Data in R and Hadoop. *Journal of Statistical Software*, 83(13). Retrieved from https://www.jstatsoft.org/article/view/v083i13

[3] S. Joško (2014). Two Aspects of Bias in Multivariate Studies: Mixing Specific with General Concepts and "Comparing Apples and Oranges." *Montenegrin Journal of Sports Science and Medicine*, 3(1), pp. 23–29.

[4] S. S Pavoine, J. Vallet, A. B. Dufour, S. Gachet and H. Daniel (2009). On the challenge of treating various types of variables: application for improving the measurement of functional diversity. *Oikos*, 118(3), pp. 391–402.p

[5] I. Irigoien and C. Arenas (2016). Diagnosis using clinical/pathological and molecular information. Statistical methods in medical research, *Statistical methods in medical research*, 25(6), pp. 2878–2894.

[6] E. C. De Assis and R. M. C. R. de Souza (2011). A K-medoids clustering algorithm for mixed feature-type symbolic data. *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 527-531.

[7] A. Dolfsson, M. Ackerman and N. C. Brownstein (2019). To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88, pp.13–26.

[8] A. Allison, C. Fialkowski and M. Allison (2018). Package "SimMultiCorrData" Type Package Title Simulation of Correlated Data with Multiple Variable Types Version 0.2.2. Retrieved from https://cran.r-project.org/web/packages/SimMultiCorrData/SimMultiCorrData.pdf

[9] W. J. Krzanowski (1993). The location model for mixtures of categorical and continuous variables. *Journal of Classification*, 10(1), pp.25–49.

[10] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu (2010). Understanding of Internal Clustering Validation Measures. *2010 IEEE International Conference on Data Mining*.

[11] A. R. De Leon, A. Soo, and T. Williamson (2011). Classification with discrete and continuous variables via general mixed-data models. *Journal of Applied Statistics*, 38(5), 1021–1032. https://doi.org/10.1080/02664761003758976.

[12] M. J. Reddy, and B. Kavitha (2012). Clustering the mixed numerical and categorical dataset using similarity weight and filter method. International Journal of Database Theory and Application, 5(1), pp.121-134.

[13] A. Foss, M. Markatou, B. Ray, and A. Heching (2016). A semiparametric method for clustering mixed data. *Machine Learning*, 105(3), pp.419-458.

[14] J. A. Hartigan and M. A. Wong (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), pp.100-108.

[15] H. Ralambondrainy (1995). A conceptual version of the K-means algorithm. *Pattern Recognition Letters*, 16(11), pp.1147-1157.

[16] J. Z. Huang, M. K. Ng, H. Rong and Z. Li (2005). Automated variable weighting in *k*-means type clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.27, No.5, 657-668.

[17] J. Z. Huang, M. K. Ng, R. Hongqiang and Z. Li (2005). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5), pp.657–668.

[18] D. McParland and I. C. Gormley (2016). Model based clustering for mixed data: clustMD. *Advances in Data Analysis and Classification*, 10(2), pp.155-169.

[19] C. Carmona, L. Nieto-Barajas and A. Canale (2019). Model-based approach for household clustering with mixed scale variables. *Advances in Data Analysis and Classification*, 13(2), pp.559-583.

[20] T. C. Headrick (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics & Data Analysis*, 40(4), pp.685-711.

[21] A. Barbiero and P. A. Ferrari (2016). An R package for the simulation of correlated discrete variables. *Communications in Statistics - Simulation and Computation*, 46(7), pp.5123–5140.

[22] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert and K. Hornik (2017). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.6. Retrieved from https://cran.r-project.org/web/packages/cluster/cluster.pdf

[23] B. Desgraupes (2016). Package 'clusterCrit.'. Retrieved from https://cran.r-project.org/web/packages/clusterCrit/clusterCrit.pdf

[24] N. R. Shamsuddin and N. I. Mahat (2019). Investigation on the Clusterability of Heterogeneous Dataset by Retaining the Scale of Variables. *Mathematics and Statistics*, 7(4), pp.49-57. https://doi.org/10.13189/ms.2019.070707.

[25] L. Kaufman and P. J. Rousseeuw (1990). Finding Groups in Data. *Wiley Series in Probability and Statistics*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.2307/2532178.

[26] L. V. Bijuraj (2013). Clustering and its Applications. In Proceedings of National Conference on New Horizons in IT-NCNHIT ,*1*, pp. 169-172.