Investigation on the Clusterability of Heterogeneous Dataset by Retaining the Scale of Variables

Norin Rahayu Shamsuddin^{1,*}, Nor Idayu Mahat²

¹Faculty of Computer & Mathematical Sciences, Merbok, 08400, Kedah, Malaysia

²School of Quantitative Sciences, Universiti Utara Malaysia, 06010 Changlun, Kedah, Malaysia

Received July 1, 2019; Revised September 8, 2019; Accepted September 23, 2019

Copyright ©2019 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract Clustering with heterogeneous variables in a dataset is no doubt a challenging process owing to different scales in a data. The paper introduced a *SimMultiCorrData* package in R to generate the artificial dataset for clustering. The construction of artificial dataset with various distribution helps to mimic the scenario of nature of real datasets. Our experiments shows that the clusterability of a dataset are influenced by various factors such as overlapping clusters, noise, sub-cluster, and unbalance objects within the clusters.

Keywords Gower's distance, k -medoids, Mixed Variables, *SimMultiCorrData*

Nonetheless, only a handful of packages have for mixed-type data such as *CluMix* [9], *MixClust* [10], *ClustOfVar* [11] that depends on *MixClust*, *BNPMIXcluster* [12], *clustMD* [13], and *kamila* [14].

Despite the increasing attention, the existing packages proposed employed Gaussian multivariate, which limited to generated continuous data. Therefore we introduced a package created by Fialkowski et al.[15] known as *SimMultiCorrData* to produces a heterogeneous artificial dataset. Generally, the artificial dataset from *SimMultiCorrData* package g constructed based on mean, variance, skewness, and/or kurtosis, from power method transformation (PMT):

$$y = c_0 + c_1 Z + c_2 Z^2 + \cdots + c_{r-1} Z^{r-1} \quad (1)$$

1 Introduction

The presence of mixed-type of variables is unavoidable, especially with the cheap of technologies nowadays. Clustering heterogeneous dataset is a challenging process. The outcome of the analysis gives a significant impact on the interpretation of clusters [1, 2, 3, 4]. Moreover, it demanded excessive computational skills and memory storage due to incorporation of broad categories [5]. The most common approached in treating heterogeneous data is through converting the variables into a single scale of measurement. However, this method may result in information loss [6, 7, 4]. 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.

Past studies have implemented different clustering process, namely k -means and k -prototypes, but little has used k -medoids. Subsequently, k -medoids demonstrate satisfactory results of clustering through when the measured variables are of mixed-type [8, 2]. There are varieties of programming packages introduced in R in generating artificial dataset or.

where $Z \sim iid N(0,1)$, c stands for cumulant and r represents the order method. The cumulant refers to mean, variance, skew, kurtosis, as well as standardized fifth and six cumulants with a fifth order method based on Headrick's method. Nevertheless, this package not proposed for clustering purposes.

The current paper presents further investigations on the behavior of clustering from the artificial dataset. Since we generate a heterogeneous data, the best clustering algorithm that allowed for mixed-types of variables is the k -medoids algorithm. We assessed the performance of the clustering through selected internal clustering validation (ICV). This ICVs gives the amount of goodness of a clustering structure based on compactness of clusters and separation between clusters. We also investigated the clusterability of the dataset in the existence of noise, unbalance data points in a cluster, as well as subclusters.

2 Mixed Variables in Clustering

Partitioning approach aims is to establish k cluster so that each object, x , in the dataset must belong to a particular one cluster and each cluster must have an object at least. The core process of partitioning clustering can be carried out via itera-

tion approach to minimize or to maximize an objective function. Typically, researchers tend to minimize the function to obtain a homogeneous cluster. Partitioning around medoids (PAM) [16], k -medoids uses the actual objects in the dataset as the center of the cluster rather than using the means points which known as medoids.

PAM enables multiple measures to be incorporated, such as the range of Euclidean, Manhattan, and Gower distances. On the contrary, only Gower's distance able to measure the dissimilarity of the heterogeneous dataset.

2.1 Gower's Distance

Assume a dataset with n objects with p variables $x_{ij} : (i = 1, 2, \dots, n; j = 1, 2, \dots, p)$ where p stands for dimensions of variables that comprise of c continuous, o ordinal, b binary or m discrete variables. For ordinal variables with o_t levels, the value written as $1, 2, \dots, o_t$. Generally, the difference in value denoted in the form of matrix structure attained from dissimilarity measure from a dataset.

Nonetheless, the distance for continuous variables is more extensive than the distance for mixed-type of variables [17]. Currently, several methods have been suggested to address the shortcomings of variation for mixed-type of variables (see [18, 19, 20, 21]). It was found that the most commonly

used measurement to deal with mixed-type of variables is the Gower's similarity measure. Gower introduced it in the 1970s with the features of Euclidean distance [22]. Generally, the similarity measure between i th object and k th object, s_{ik} , is calculated as:

$$s_{ik} = \frac{\sum_{l=1}^p w_{ikl} s_{ikl}}{\sum_{l=1}^p w_{ikl}} \quad (2)$$

where w_{ikl} refers to weight. $w_{ikl} = 0$ if objects i and k cannot be compared for variable $p = l$. For instance, let variable l be the criterion of chest pain after certain exercises. Thus, comparing objects i and k based on variable l would be biased in one of these two objects that do not show the pain.

Nevertheless, Gower's suggested a few methods to dismiss ordinal variables. Since many datasets have ordinal variables, Podani [23] had stretched out Gower's study by inserting the function for ordinal, as given in the following:

$$s_{ikl} = 1, \text{ if } r_{il} = r_{kl} \quad (3)$$

where r_{il} and r_{kl} are the ordinal variables for objects i and k , respectively.

If the number of objects share similar rank score for variable p , therefore:

$$s_{ik} = 1 - \frac{|r_{il} - r_{kl}| - (T_{il} - 1)/2 - (T_{kl} - 1)/2}{\max\{r_l\} - \min\{r_l\} - (T_{l,max} - 1)/2 - (T_{l,min} - 1)/2} \quad (4)$$

T_{il} refers to the number of n objects in that share similar rank score with object i for variable l , T_{kl} denotes the number of objects in n that have the same rank score with object k for variable l , $T_{l,max}$ reflects the number of objects with maximum score rank for variable l , and $T_{l,min}$ represents the number of objects with minimum score rank for variable l . An extension of Gower's distance proposed by [23] allowed for metric or non-metric version. Currently, Gower's similarity index has been widely used to support the vector machine, machine learning, recognizing the patter, bioinformatics, molecular biology, epidemiology, and other disciplines.

2.2 Internal Clustering Validation (ICV)

The ICV measures (as in [24, 25, 26, 27]) provide better options for researchers to select the most appropriate method for clustering. The validity indices determine the quality of clustering, apart from discovering the optimal K . In order to attain effective evaluation of cluster analysis, understanding the 'definition' of internal validation is crucial. In addition, internal validation is a method to assess the fitness of the structure and data in clustering as it is based on the distance between the objects in the cluster as well as between the clusters [28].

Handl and Knowles [29] explained ICV from the stance of three aspects; (i) compactness, (ii) connectedness, and (iii) separation. Compactness is considered to attain homogeneous cluster. High homogeneity can lead to overlapping of clus-

ters. Next, connectedness examines the connection of neighbouring data points in the similar clusters. Meanwhile, separation examines the degree of partition between the clusters. Additionally, inverse compactness leads to a good quality of separation and vice versa. Based on these measurements, most of the previous studies had conducted clustering by combining the measurement of compactness and separation to dismiss the cluster's density and structure. In fact, compactness and separation were estimated to be suitable number of clusters K of a dataset. The indices of ICV can be discovered in [30, 31, 32].

Intra-cluster compactness and separation were measured using Dunn (D) index [33]. D index shares same range values with Davies-Bouldin (DB) index, yet, D index generates maximum value to show compact and well-separated clusters. In addition, DB and D had utilized the ratio-type coefficient within the different clusters and also between the different clusters which do not display any trends [25]. Meanwhile, the concept of ΓC is adopted from [34] and designed by [35]. They defined it as average of total distance between objects in a cluster over other clusters. In addition, it also defines the average to the nearest and the furthest distances between the pairs of objects in the dataset (Table 1). In the C-index, the minimum value represented the good clusters. [36] suggested that Silhouette (S) index should determine the objects that do not belong to cluster and depicts the wellness of the objects in the cluster. Thus, the index will measure the average distance between the objects within a cluster and the average distance between the

Table 1. Internal cluster validation indices (ICV)

Indices	Formula
Davies-Bouldin (<i>DB</i>)	$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i$ $R_i = \max_{k=1, \dots, n_c, i \neq k} (R_{ik}) \quad i = 1, \dots, n_c$ $R_{ik} = \frac{s_i + s_k}{d_{ik}}$ $d_{ik} = d(v_i, v_k) \quad s = \frac{1}{\ c_i\ } \sum d(x, v_i)$
C-index (ΓC)	$C_{index} = \frac{S - S_{min}}{S_{max} - S_{min}}$
Dunn (<i>D</i>)	$D = \min_{1 \leq i \leq c} \left(\min_{i+1 \leq k \leq c} \left(\frac{d(c_i, c_k)}{\max_{1 \leq j \leq c} diam(c_j)} \right) \right)$
Silhouette (<i>S</i>)	$ASW = \frac{1}{n} \sum_{i=1}^n s_i \quad s_i = \frac{b(i) - a(i)}{\max a(i), b(i)}$ $a(i) = \text{average dissimilarity of } i \text{ to all other objects of } A$ $b(i) = \min_{C \neq A} d(i, C)$ $d(i, C) = \text{average dissimilarity of } i \text{ to all other objects of } C$

n_c = no. of clusters; v_i = centre of cluster; $\|c_i\|$ = norm of c_i ; c_i = cluster; S = sum of distances over all pairs of objects; S_{min} = sum of pair with smallest distances if all pairs of objects; S_{max} = sum of pair largest distances out of all pairs; $d(\cdot, \cdot)$ = distance between clusters/centred of clusters; $x_i = i^{th}$ objects.

clusters.

3 Methodology

This package created all variables from standard and normal variables from an intermediate correlation matrix. The continuous mixture variables are drawn from more than one component distribution. Categorical and count variables are generated from inverse cumulative density function (cdf). As for ordinal variable, the data generated through a process of discretising the standard normal variables at quantiles in conformity to the targetted marginal distribution. Meanwhile, the count variables comprise of standard and normal variables of uniform distribution. The continuous mixture variables drawn from more than one component distribution is describe in term of mixture distribution.

The setting for generating the heterogenous dataset is described in the following:

- (a) count variables were set to derive from one of these 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) ordinal variables were derived from two scenarios:

- (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) the continuous variables were drawn from Gaussian ($N(0, 1)$), gamma ($\Gamma(\alpha = \beta = 10)$), and chi-square (χ_4^2) distributions to have both normal and non-normal datasets.

From the uniform distribution with $U(0.25, 0.7)$, the correlation matrix was obtained and fell within the feasible correlation bounds. The `valid_corr` function for ‘Correlation Method 1’ resolute if the matrix is in the boundaries. Subsequently, the variables were generated using the function of `rcorrvar` along with the addition of error loop to reduce the errors in correlation.

Generally, ten parameters along with 300 objects were generated as follows:

- (a) five count variables – three variables with different mean values for Poisson distribution and two variables for negative binomial with different parameters r and p ,
- (b) two ordinal variables with two groups of different categories, and
- (c) three continuous variables – chi-square, gamma, and Gaussian distribution.

We employed the *k-medoids* clustering algorithm, and the outcome of clustering validated based on the selected ICVs. The overall method describe in Algorithm 1.

Algorithm 1 Algorithm in *k-medoids* and ICV

1. Selection of initial medoids
 - (a) Determine the number of *k*-clusters
 - (b) Arbitrarily choose the objects from dataset as initial medoids. The number of initial medoids is based on Step (1a) – BUILD phase.
 - (c) Compute the dissimilarity between the medoids and the objects by employing Gower's distance.
 2. Assign objects to medoids
 - (a) Assign the objects to the nearest medoids.
 - (b) Other objects (non-medoids) are selected as medoids, if the objective function is not achieved – SWAP phase.
 3. Update medoids

Steps in (2a) and (2b) are repeated until no change is noted in medoids *location*.
 4. Selection of *k*-clusters

For every *k* selected for PAM, ICV is conducted to look at the best performance of *k*.
-

The computed indices in the current paper presented in Table 1. For review purpose, the ratio-type coefficient between-within cluster variation and between cluster variations of *DB* and *D* indices were implemented. *DB* offered details regarding the separation of cluster as all clusters were measured to attain the mean values.

The score of *DB* and *D* indexes is between $[0, \infty]$. However score value in indicating a clustering differs, where *DB* should be minimized, while and *D* should produce maximized value. The range of *ΓC* is between $[0,1]$, and a score that is nearest to zero (0) signifies good clustering. *S* index takes the value of $[-1,1]$, and the average silhouette width approach value 1 indicates a strong structure of clustering has been found.

We observed the structure of clustering at difference *k* to identify the effect of diameter of a clusters and separation between clusters. We notice that unbalanced objects within cluster, overlapping between clusters, effect of noise on the diameters of a clusters as well as sub-clustering give large impact in determine the quality of clusters.

4 Findings

Based on the 300 generated observations, it was found that the setup simulation creates few data points with greater distance from the others, as shown in Figure 1. However, we do

not intend to exclude those data points to look at how the clustering performed. Generally, mixed variables dataset is known to acquire overlapping clustering. Therefore it is hard to determine the optimal number of clusters. *K*. Relying only on the scatterplot neither guarantees us where and which objects belong to which clusters. Nor in determining the number of clusters. Hence, *k-medoids* algorithm with Gower's distance were employed for further evaluations.

Figure 1. Scatterplot of heterogeneous artificial dataset.

4.1 Number of Clusters

Figure 2. Assessment of clustering results are based on the ICVs.

The dendrogram illustration in Figure 3 is a compact visualization of the dissimilarity matrix. The y-axis indicates how similar the data points in its clusters and how dissimilar it is between clusters. The heights represented at y-axis helped us to identify which data points belong to which clusters (or groups) at specific heights values based on its similarity. Note that the

dendrogram illustration does not let us set the K , but may only yield an appropriate number of clusters.

Therefore, the assessment of the clustering structure of an algorithm is evaluated based on the ICVs values. Looking at Figure 2, it is apparent that DB , ΓC , and S shows a point at which a significant change in values of the index for $k = [2, 10]$. However, D index hardly indicates the trend of changes in the index values for the heterogeneous dataset. The indexes values of the DB , D , ΓC , and S are 1.1163, 0.0501, 0.0529 and 0.3381, respectively. DB and S indices suggest the construction of clusters at $K = 2$. On the hand, D index suggested the suitable clusters formations be $K = 3$, while ΓC indicates $K = 10$ as the best number of clusters. Judging from the value obtained and the constraint of the four indexes, the suitable K for this artificial dataset should be $K = 10$ since only ΓC index able to meet its constraints.

4.2 Cluster Structure Assessment

Further investigations on the behavior of clusters formations based on the selection of K from DB , D , ΓC , and S indexes presented in Figure 4. From this figure, it can be seen that there is no distinct separation on the formation of clusters for $K = 2$ and $K = 3$. Interestingly, there occur a pattern of clusters for $K = 10$ in Figure 4 (c), where it suggested that the number of clusters for this dataset is equivalent to three. However, it only can be managed with some conditions. For example, data points allocated to cluster 5, 7, and 8 should be classified as one class, while data points in cluster 2, 4, and 9 as another class. Hence, a precise clustering formation could have formed.

From the average silhouette width (figure not shows for $K = 2$ and $K = 3$), overall average silhouette width for $K = 2$ yield 0.28. The average silhouette width of the first cluster is only 0.2, while the second cluster recorded an average of 0.45. The overall average silhouette width for $K = 3$, we obtained 0.24, where the average of the first cluster is 0.3, second clusters with 0.35 while only 0.4 in the third cluster. In $K = 10$, the average silhouette width obtained is 0.27, and the average value of each cluster shown in Figure 5.

We observed the structure at difference K to identify the effect of diameter of a cluster and separation between clusters. We notice that unbalanced objects within cluster, overlapping between clusters, effect of noise on the diameters of a clusters as well as sub-clustering as shown in Figure 6, give large impact in clusterability of this dataset.

5 Discussion

Various results in the number of clusters obtained from different ICVs. This phenomenon is common for a dataset to have more than one K that represent good clusters. This is because an unsupervised clustering makes it more challenging to discover the best K . Nevertheless, careful validation of clusters and discovering the best K as particular indices are sensitive to the outliers, noise, and even subclusters. A wise selection of ICV used in a study is essential since the numerator and denominator of the indices is much influence on the structure of

the dataset.

Halkidi, Batistakis, and Vazirgiannis [25] independent reviews define that validation assessment performance is the best when clusters are thoroughly compact (well clustered), but performance badly in data with mixed variables. Having a mixed type of variables in clustering may create some discrimination in term of dissimilarity. Bijuraj [37] mentioned that it is more appropriate if a dataset has a similar of variables.

Intersection, insertion or deletion at certain magnitude of objects within a cluster significantly influence the of dispersion (width) as well as density of a cluster. As the K increase, the intra-cluster compactness tends to decrease an inter-cluster separation increase simultaneously. k -medoids assigned objects to it nearest medoids. Medoids is based on points of dataset that represent minimal averaging dissimilarity measures. The concepts similar to the k -means. Due to this objective function, it explained the implication of the factors.

Since this dataset consists of overlapping data points in clustering, k -medoids algorithm is not suitable to carry out the cluster analysis. It is known that k -medoids algorithm hard clustering, in which each data points should belong to one and only one cluster. Therefore other clustering approaches, such as density-based method should be employed. The differences in the decision of appropriate K between ICVs were influenced by various factors that affected the nominator and denominator of ICVs.

The problems in identifying the appropriate K have created an interest among researchers in the discipline of clustering. Moreover, it becomes as essential issues as its effects on the performance of the internal validity of clustering. Few research works of this matter have been carried out and reported in [38, 39, 40] and few others cited elsewhere.

6 Concluding Remarks

In this paper, we investigate the clusterability of mixed-types of variables of a dataset through artificial dataset consist of mixed types of data from various distributions from *SimMultiCorrData* package. This package allows users to choose to work on empirical or theoretical approaches based on one's objective in generating simulation dataset. We have conducted both approaches, and we discovered that the empirical method is time-consuming, and the correlation matrix always hard to achieve. For this paper, we opt for theory methods using `cal` theory function with mixed types of variables. Furthermore, it is happened a simulation dataset from this package generate some noises and outliers which give benefits to identify the effect on clustering. Since the simulation data not purposely developed for clustering, therefore is unable to measure the level of overlap between clusters.

We have demonstrated that by 'retaining' the scale of variables through Gower's distance, the formation of clusters does exist. However, we obtained poor results of clustering using k -medoids algorithm. Results from the average silhouette width, it can be said indicated that no substantial structure had been found. As for the effect of various factors in clusterability of dataset, we have performed a weighted clustering approach to

Figure 3. Complete linkage dendrogram of artificial dataset.

Figure 4. Position of data points in its define clusters based on suggested K from ICVs.

Figure 5. Silhouettes of a clustering with $K = 10$ of the artificial dataset.

reduce the overlapping between clusters and assigning the objects to it nearest cluster. We also deleted data points that were identified as outliers/noise (results are not display) to obtained a better clustering results. Yet, there is no improvement in the validity scores but further worsen the clustering performance. Besides, the process of clustering mixed-type of variables and retaining their scale throughout the analysis was indeed challenging.

k -medoids

Acknowledgements

The authors would like to express their gratitude to the editor and reviewers for their insights to help improving this paper. We would also gratefully acknowledge the financial support by Universiti Teknologi MARA (UiTM) and the Ministry of Higher Education (MOHE) Malaysia.

REFERENCES

- [1] Z. Huang. Extensions to the k -means algorithm for clustering large data sets with categorical values, *Data Mining and Knowledge Discovery*, Vol.2, No.3, 283-304, 1998.
- [2] C. Hennig, T. F. Liao. 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)*, Vol.62, No.3, 309-369, 2013.
- [3] G. Szepannek. 'clustMixType': k -Prototypes clustering for mixed variable-type data, R package version 0.1-16, 2017.
- [4] A. H. Foss, M. Markatou. *kamila*: Clustering mixed-type data in R and Hadoop, *Journal of Statistical Software*, Vol.83, No.1, 44, 2018.
- [5] H. Ralambondrainy. A conceptual version of the K-means algorithm, *Pattern Recognition Letters*, Vol.16, No.11, 1147-1157, 1995.
- [6] W. J. Krzanowski. The location model for mixtures of categorical and continuous variables, *Journal of Classification*, Vol.10, 25-49, 1993.
- [7] G. Celeux, G. Govaert. Latent class models for categorical data, Chapman & Hall/CRC, 2015.
- [8] E. C. de Assis, R. M. C. R. de Souza. A k -medoids clustering algorithm for mixed feature-type symbolic data, *EEE International Conference on Systems, Man and Cybernetics*, 527-531, 2011.
- [9] M. Hummel, D. Edelmann, A. Kopp-Schneider. *CluMix*: Clustering and visualization of mixed-type data", *Comprehensive R Archive Network (CRAN)*, 2017.
- [10] M. Marbac, C. Biernacki, V. Vandewalle. Model-based clustering of Gaussian copulas for mixed data, *Communications in Statistics-Theory and Methods*, Vol.46, No.23, 11635-11656, 2017.
- [11] M. Chavent, V. Kuentz, B. Liquet, L. Saracco, *ClustOfVar*: An R package for the clustering of variables, *Journal of Statistical Software*, 2011.
- [12] C. Carmona, L. Nieto-Barajas, A. Canale. Model-based approach for household clustering with mixed scale variables, *Advances in Data Analysis and Classification*, Vo.13, No.2, 559, 2019.
- [13] D. McParland, I. C. Gormley. *clustMD*: Model based clustering for mixed data, *Comprehensive R Archive Network (CRAN)*, 2017.
- [14] A. H. Foss, M. Markatou. *kamila*: Methods for clustering mixed-type data. *Comprehensive R Archive Network (CRAN)*, 2018.
- [15] A. C. Fialkowski, *SimMultiCorrData*: Simulation of correlated data with multiple variable types, *Comprehensive R Archive Network (CRAN)*, 2017.
- [16] L. Kaufman, P. J. Rousseeuw. *Clustering by Means of Medoids*, Faculty of Mathematics and Informatics, 1987.
- [17] A. S. Shirshorshidi, S. Aghabozorgi, T. Y. Wah. A comparison study on similarity and dissimilarity measures in clustering continuous data, *PLoS ONE*, Vol.10, No.12, e0144059, 2015

- [18] Z.Huang. Clustering Large Data Sets with Mixed Numeric and Categorical Values, Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining,(PAKDD), 21-34, 1997, 1997.
- [19] M.S. Yang, P. Y. Hwang, D. H. Chen. Fuzzy clustering algorithms for mixed feature variables, Fuzzy Sets and Systems, Vol.141, No.2, 301-317, 2004.
- [20] A. Ahmad, L. Dey. A k -mean clustering algorithm for mixed numeric and categorical data, Data and Knowledge Engineering, Vol.63, No.2, 503-527, 2007.
- [21] J. Ji, T. Bai, C. Zhou, C. Ma, Z. Wang. An improved k -prototypes clustering algorithm for mixed numeric and categorical data, Neurocomputing, Vol.120, 590-596, 2013.
- [22] J.C. Gower. A general coefficient of similarity and some of its properties, Biometrics, Vol.27, No.4, 857-871, 1971.
- [23] J. Podani. Extending Gower's general coefficient of similarity to ordinal characters, Taxon, 331-340, 1999.
- [24] R. Tibshirani, G. Walther, T. Hastie. Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol.63, No.2, 411-423, 2001.
- [25] M. Halkidi, Y. Batistakis, M. Vazirgiannis,. Clustering validity checking methods: Part II, ACM Sigmod Record, Vol.31, No.3, 19-27, 2002.
- [26] M. Kim, R. S. Ramakrishna. New indices for cluster validity assessment, Pattern Recognition Letters, Vol.26, No.15, 2353-2363, 2005.
- [27] J. C. Rojas-Thomas, M. Santos, M. Mora. New internal index for clustering validation based on graphs, Expert Systems with Applications, Vol.86, No.15, 334-349, 2017.
- [28] A. K. Jain, R. C. Dubes. Algorithms for Clustering Data, Prentice Hall, 1988.
- [29] J. Handl, J. Knowles. Exploiting the Trade-off - The Benefits of Multiple Objectives in Data Clustering, In: Coello Coello C.A., Hernández Aguirre A., Zitzler E. (eds) Evolutionary Multi-Criterion Optimization. EMO 2005. Lecture Notes in Computer Science, Vol. 3410, Springer, Berlin, Heidelberg, 547-560, 2005.
- [30] B. Desgraupes. *clusterCrit*: clustering indices. R package version 1.2. 3", Comprehensive R Archive Network (CRAN), 2015.
- [31] M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs,. *NbClust*: An R package for determining the relevant number of clusters in a data set", Journal of Statistical Software, Vol.61, No.6 1548-7660, 2014.
- [32] M. Hassani, T. Seidl. Internal clustering evaluation of data streams, Trends and Applications in Knowledge Discovery and Data Mining, 198-209, 2015.
- [33] D. L. Davies, D. W. Bouldin. A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.2, 36, 1979.
- [34] E. C. Dalrymple-Alford. Measurement of clustering in free recall, Psychological Bulletin, Vol.74, No.1, 32-34, 1970.
- [35] L. J. Hubert, J. R. Levin. A general statistical framework for assessing categorical clustering in free recall, Psychological Bulletin, Vol.83, No.6, 1072, 1976.
- [36] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, Vol.20, 53-65, 1987.
- [37] L. V. Bijuraj. Clustering and its Applications, Proceedings of National Conference on New Horizons in IT-NCNHIT, 169-172, 2013.
- [38] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu. Understanding of Internal Clustering Validation Measures, Proceeding IEEE International Conference of Data Mining, 911-916, 2010.
- [39] J. C. Bezdek, N. R. Pal. Some new indexes of cluster validity, Part B (Cybernetics) IEEE Transactions on Systems, Man, and Cybernetics, Vol.28, No.3, 301-315, 1998.
- [40] S. Saitta, B. Raphael, I. F.C. Smith. A bounded index for cluster validity, Machine Learning and Data Mining in Pattern Recognition, Vol.30, 174-187, 2007.

Figure 6. Csusterability of dataset based various unbalanced data points in a cluster, overlapping of data points between clusters and sub-clusters.