© 2018 IJSRSET | Volume 4 | Issue 9 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Comparison of Cluster Ensemble and Two Step Cluster Methods on Clustering with Mixed Type Data

Fera Hermawati*, Budi Susetyo, Agus Mohamad Soleh
Department of Statistics, Bogor Agricultural University, Bogor, West Java, Indonesia

ABSTRACT

Health development is supported by the availability of adequate health facilities and personnel. To facilitate the government in determining the policies taken, it is necessary to group the region to know which areas that need improvement in health facilities and personnel. Cluster analysis is used to group objects based on certain characteristic similarities. Cluster analysis is generally applied to objects with numerical data types. Health facility and health personnel data have categorical and numerical types or also called mixed data type, so it is necessary to use clustering for mixed types data. This study aims to compare cluster ensemble method and two step cluster method in clustering mixed type data. The comparative criterion used is the ratio between diversity within cluster (S_w) and the diversity between cluster (S_b). Smaller ratio values indicate a better method. The research results showed that cluster ensemble method is a better method than the two step cluster method in clustering mixed type data.

Keywords: Clustering, Cluster Ensemble, Two Step Cluster, Mixed Type Data

I. INTRODUCTION

Health is a basic need of every human being and is the capital of every citizen in achieving prosperity. Health development is held to realize the highest degree of public health. To support this, the availability of health facilities / infrastructure and the availability of Human Resources (HR) in the implementation of health services plays an important role. Accurate health and data information support is crucial in making decisions toward the right health policy and development strategy. To facilitate the government in determining the policies taken, it is necessary to group the region to know which areas that need improvement in health facilities and health personnel.

The commonly used statistical method for grouping objects is cluster analysis. Cluster analysis is generally applied to objects with numerical data types. Health facility and health personnel data have categorical and numerical types are also called mixed data, so it is necessary to apply a clustering method to mixed data type. There are several methods developed to perform cluster analysis on mixed data. [1] presents the k-prototype algorithm, which is the development of k-means. [2] presents a Similarity Based Agglomerative Clustering (SBAC) algorithm. Another method developed for clustering objects with mixed data is two step cluster method [3]. [4] developed cluster ensemble method, while [5] used fuzzy clustering to clustering mixed data objects.

In this research, the sub districts clustering is conducted in Malang, East Java province based on on health facilities and health personnel using cluster ensemble method and two step cluster method, then compare the result of the clustering of both methods using the ratio between diversity within cluster and the diversity between cluster (S_w/S_b).

II. METHODS AND MATERIAL

A. Data Sources

The data used in this research is Potential Village data (PODES) of Malang Regency 2014 from the Central Bureau of Statistics (BPS) and Health Human Resources data of Malang Regency 2017 from the Ministry of Health (Kemenkes) of the Republic of Indonesia [6]. The categorical variables used were the availability of community health centers (puskesmas) with inpatient (X_1), availability of inpatient community health center (puskesmas) (X_2), availability of polyclinic (X_3), availability of doctor's practice (X_4), and availability of village health post (poskesdes) (X_5). The numerical variables used were the ratio of general practitioners per 100,000 population (Y_1), dentist ratio per 100,000 population (Y_2), midwife ratio per 100,000 population (Y_3), and ratio of other health personnel per 100,000 population (Y_4).

B. Procedure of Analysis

Data were analyzed using R studio and SPSS software for Windows 15.0. Here are the steps taken in this research:

1. Clustering mixed data using cluster ensemble method.
   a. Divide data into two, the data with numerical variables and data with categorical variables.
   b. Clustering categorical data using the squeezer algorithm.
   c. Clustering numerical data using the agglomerative hierarchy method with Euclidean distance, using complete linkage method and ward linkagex.
   d. Combining the results of categorical data clustering and the result of numerical data clustering by bunching them using squeezer algorithm.
2. Clustering mixed data using two step cluster method.
3. Comparing the results between cluster ensemble method and two step cluster method by comparing ratio between diversity within cluster and the diversity between cluster (S_w/S_b).

C. Cluster Ensemble

Cluster ensemble is a method that combines a bunch of different cluster solutions from each different method into one final cluster solution. Cluster ensemble consists of 2 (two) stages. The first stage is to cluster with several algorithms and store the clustering results. The second stage determines the final cluster of the first-stage results using the consensus function [7]. The advantage of this method is to improve the quality and robustness of the cluster solution [8]. In a mixed data clustering, the cluster ensemble steps can be shown in Figure 1.

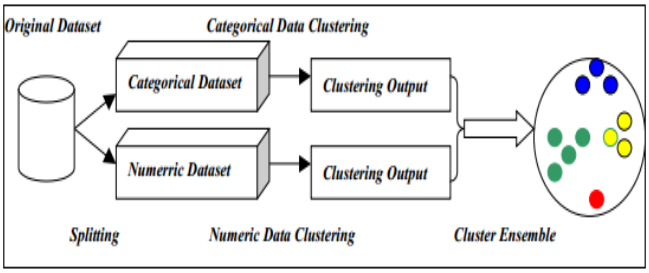

Figure 1. The cluster ensemble stages of the mixed data clustering

1. Categorical Data Clustering (Algoritma Squeezer)

This algorithm starts with setting the threshold. boundary value to determine whether an object enters to the cluster or forms a new cluster. The magnitude of the threshold can affect the result of clustering and speed of the algorithm work. The similarity between cluster (Cl) dan and object {tid} is calculated using the following formula [9]:

$$\text{Sim}(\text{Cl}, \text{tid}) = \sum_{l=1}^{m_{kat}} \left(\frac{\text{Sup}(a_l)}{\sum_j \text{Sup}(a_j)} \right) \quad (1)$$

where a_l is the value at the l variable (tid.A_l), m_{kat} = number of categorical variables, $\text{Sup}(a_l)$ is the number of objects on Cl containing the a_l values.

In the squeezer algorithm, the data is read sequentially. The first object is read and initialized as

the first cluster, then continued by reading the second object. The similarity of the second object and the first cluster is calculated (sim (1,2)). If sim (1,2) ≥ threshold, then the object is inserted in the first cluster. If sim (1,2) < threshold, then the object forms a new cluster. Perform that step until all the objects are read. When calculating the similarity between the object and the cluster, there will be a similarity value as much as the number of clusters that have been previously established. Determine the maximum of similarity (sim_max). If sim_max ≥ threshold, , then the object is inserted as a cluster member with maximum similarity. If sim_max < threshold, then the object form a new cluster.

2. Numerical Data Clustering

The numerical data clustering uses the agglomerative hierarchy method with Euclidean distance of the formula [10]:

$$d_{ij} = \sqrt{\sum_{l=1}^{m}\left(x_{il} - x_{jl}\right)^2} \qquad (2)$$

where: d_{ij} = distance between the i object and j object, i, j = 1,2, ... , x_{il} = the value of i object on the l -th variable, l = 1,2, ... ,m , m = number of numerical variables.

The agglomerative method used is complete linkage and ward linkage. The optimum cluster determination is determined based on the 30 validity indices issued by the NbClust packages at each linkage method [11]. Optimum cluster is determined by by the number of cluster assigned by the majority of the validity indices. In each linkage taken is 2 clustering result namely the cluster with the first of the majority validity index and the cluster with the second of the majority validity index, so there are 4 results of the fourth numerical data clustering used in the final clustering method of cluster ensemble.

D. Two Step Cluster

The two step cluster method is another clustering method that can handle objects that have mixed data types [12]. The procedure of clustering objects in two step cluster method is carried out through two stages [3, 12], i.e. pre-clustering stage and the optimal clustering stage. Initial stages in pre-clustering is to standardize all numerical variables, then perform the initial clustering by forming Cluster Feature (CF) Tree sequentially to form subcluster-subcluster. The distance calculation is using Log-Likelihood distance formula. The distance between j and s cluster is defined as follows :

$$d(j,s) = -N_j\left(\sum_{k=1}^{K^A}\frac{1}{2}\log\left(\hat{\sigma}_k^{\,2} + \hat{\sigma}_{jk}^{\,2}\right) + \sum_{k=1}^{K^B}\hat{E}_{jk}\right) - N_s\left(\sum_{k=1}^{K^A}\frac{1}{2}\log\left(\hat{\sigma}_k^{\,2} + \hat{\sigma}_{sk}^{\,2}\right) + \sum_{k=1}^{K^B}\hat{E}_{sk}\right) + N_{(i,s)}\left(\sum_{k=1}^{K^A}\frac{1}{2}\log\left(\hat{\sigma}_k^{\,2} + \hat{\sigma}_{(i,s)k}^{\,2}\right) + \sum_{k=1}^{K^B}\hat{E}_{(i,s)k}\right) \qquad (3)$$

where : $\hat{E}_{jk} = -\sum_{l=1}^{L_k}\frac{N_{jkl}}{N_j}\log\frac{N_{jkl}}{N_j}$, N_{jkl} = the amount of data in the j cluster of the k-th categorical variables with the l-th category, $\hat{\sigma}_{jk}^{\,2}$ = the variance estimator for the k-th numeric variable in the j-th cluster , K^A = total number of numerical variables, K^B = total number of categorical variables, L_k = the number of categories in k-th categorical variables.

The next stage is to cluster the result of subcluster formed using agglomerative hierarchy method. The determination of the optimal number of clusters is calculated using the Bayesian Information Criterion (BIC) value. The BIC value for the j cluster is as follows :

$$BIC_j = -2\sum_{j=1}^{J}\left(-N\left(\sum_{k=1}^{K^A}\frac{1}{2}\log\left(\hat{\sigma}_k^{\,2} + \hat{\sigma}_{jk}^{\,2}\right) + \sum_{k=1}^{K^B}\hat{E}_{jk}\right)\right) + r_k\log(N) \qquad (4)$$

where : $r_k = J\left\{2K^A + \sum_{k=1}^{K^B}(L_k - 1)\right\}$

At each clustering stage, calculations are performed for BIC, Log-Likelihood distance, BIC change value, BIC change ratio and distance ratio. Determination of the number of optimum cluster is by looking at the

smallest BIC values. However, there are some cases where the larger the cluster, the smaller the BIC value. In this case, determination of the maximum cluster using the BIC change ratio ($\frac{BIC_k}{BIC_1}$) was first valued less than 0.04. Determination of the optimal number cluster is to calculate the ratio of two largest distance changes ($\frac{R(k_1)}{R(k_2)}$). If the value is more than 1.15, then the optimum cluster number is cluster with the first largest distance ratio (k_1), If the value is equal or less than 1.15, then the number cluster equals to the maximum (k_1, k_2).

E. Cluster Diversity

A good cluster is homogeneous within the cluster and heterogeneous between clusters. In cluster analysis, homogeneity among members in the cluster and homogeneity between clusters was calculated using the diversity within cluster (S_w) and the diversity between cluster (S_b). The diversity within cluster (S_w) and the diversity between the groups (S_b) are formulated by [13] as follows:

$$S_w = \frac{1}{C}\sum_{i=1}^{C} S_i \quad (5)$$

$$S_b = \sqrt{\frac{1}{C-1}\sum_{i=1}^{C}(\bar{x}_i - \bar{x})^2} \quad (6)$$

where : $S_i = \sqrt{\frac{1}{n_i-1}\sum_{i=1}^{n_i}(x_i - \bar{x})^2}$ = standard deviation of the i-th cluster , C = number of cluster, n_i = number of objects in the i-th cluster , x_i = object in the i-th cluster , $\bar{x}_i$ = mean of the i-th cluster, $\bar{x}$ = mean of objects on all clusters.

Adopting from a diversity measurement in numerical data, it is developed a diversity measurement for categorical data as in [14]. The diversity within cluster (S_w) and the diversity between cluster (S_b) formulated is as follows:

$$S'_w = \sqrt{\frac{WSS}{(n-C)}} \quad (7)$$

$$S'_b = \sqrt{\frac{BSS}{(C-1)}} \quad (8)$$

where :

$WSS = \frac{n}{2} - \frac{1}{2}\sum_{c=1}^{C}\frac{1}{n_c}\sum_{k=1}^{K} {n_{kc}}^2$ = Within Sum of Squares

$BSS = \frac{1}{2}\left(\sum_{c=1}^{C}\frac{1}{n_c}\sum_{k=1}^{K} {n_{kc}}^2\right) - \frac{1}{2n}\sum_{k=1}^{K} {n_k}^2$ = Between Sum of Squares

$n = \sum_{c=1}^{C} n_c = \sum_{k=1}^{K} n_k = \sum_{k=1}^{K}\sum_{c=1}^{C} n_{kc}$ = total number of objects

$n_k = \sum_{c=1}^{C} n_{kc}$ = number of objects in the k-th category, k = 1,2,...,K

$n_c = \sum_{k=1}^{K} n_{kc}$ = the number of objects in the c-th cluster

n_{kc} = the number of objects in the k-th categories and the c-th cluster

The smaller the ratio between S_w and S_b then the clustering method used has a good performance.

III. RESULTS AND DISCUSSION

A. Clustering with Cluster Ensemble Method

1. Categorical Data Clustering

The threshold value obtained from the calculation is 4.67, so the number of cluster formed on categorical data clustering is 6 cluster. Figure 2 shows the number and percentage of sub-districts of each cluster. The result shows that there are 2 groups consisting of only 1 sub-district, namely group 3 and group 6.

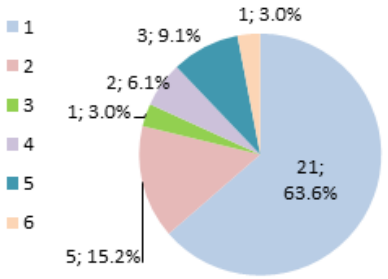

Figure 2. Number and percentage of cluster members from categorical data clustering results

2. Numerical Data Clustering

The result of hierarchical clustering by using the complete linkage and the ward linkage can be seen in the cluster dendogram in Figure 3.

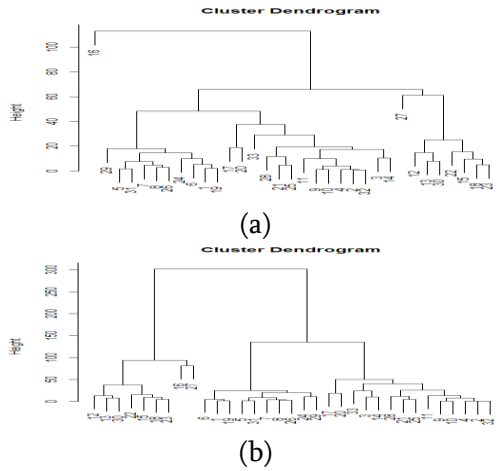

Figure 3 . The cluster dendogram of sub district clustering uses the Euclidean distance of the complete linkage (a) and the ward linkage (b)

It can be seen that the dendogram cut can be performed on the number of 2, 3 or 4 cluster. In addition to using visual techniques, determination of the number cluster can be performed through the calculation of the validity index. Calculation of the validity index in the complete linkage indicates that 2 is the optimum cluster number and 4 as the next optimum cluster number. The calculation of the validity index on the ward linkage shows 2 as the optimum cluster number and 5 is the next optimum cluster number. The four results of this clustering are used in the final clustering of cluster ensemble methods. The number of members of each cluster of numerical data clustering can be seen in Table 1.

Table 1. Number of cluster members in each cluster resulting from the clustering of numerical variables

Method	Cluster	Number of member
Euclidean-Complete (1)	1	32
	2	1
Euclidean-Complete (2)	1	24
	2	7
	3	1
	4	1
Euclidean-Ward (1)	1	24
	2	9
Euclidean-Ward (2)	1	10
	2	14
	3	7
	4	1
	5	1

3. Final Clustering of Cluster Ensemble

In the final clustering of cluster ensemble method there are 4 combinations of numerical data clustering and categorical data clustering. At this stage, the threshold used is 0.1, 0.2 to 1.9. The selection of the final cluster is seen through the smallest ratio between diversity within cluster and diversity between cluster (S_w/S_b). The number of clusters, thresholds, ratio between diversity within and between the cluster and the number of selected cluster can be seen in Tabel 2.

Table 2. Ratio between diversity within and diversity between the clusters (S_w/S_b) of cluster ensemble method in 4 combinations of numerical clustering

Clustering Method	Threshold	Number of Cluster	Sw/Sb	Selected clusters
Euclidean-Complete (1) (EC1)	0.1 – 0.9	2	**1.670**	2
	1.0 – 1.9	7	0.246	
Euclidean-Complete (2) (EC2)	0.1	2	1.435	5
	0.2	3	1.240	
	0.3 – 0.4	4	0.993	
	0.5 – 0.9	5	**0.689**	
	1.0 – 1.9	12	0.108	
Euclidean-Ward (1) (EW1)	0.2 – 0.5	2	1.284	2
	0.6 – 0.9	2	1.353	
	1.0 – 1.9	9	0.311	
Euclidean-Ward (2) (EW2)	0.1	2	1.435	5
	0.2	3	1.240	
	0.3 – 0.4	4	0.993	
	0.5 – 0.9	5	**0.689**	
	1.0 – 1.9	12	0.108	

From Table 2, it can be seen that the clustering of EC2 and EW2 produces the same result. The selection of the final cluster in addition to select a small S_w/S_b diversity ratio value, also takes into consider the more meaningful cluster characteristics. It is seen from the cluster characteristics of each cluster result, which has more meaning is on the clustering of EC2 and

EW2 on the threshold of 0.5 to 0.9 with a ratio S_w/S_b of 0.689 which produces 5 groups, so in clustering of cluster ensemble method, the number of final group is 5.

B. Clustering with Two Step Cluster

In two step cluster method, determining the number of cluster using BIC criterion. Based on the processing using SPSS, the smallest BIC value found in the 2nd group is 220.067. It also can be seen that with the increasing number of cluster, BIC value is more increasing. In this case, the optimal number of cluster is 2, because it has the smallest BIC value.

Table 3. BIC value (Bayesian Information Criterion) on two step cluster

Number of cluster	BIC value	BIC change value	BIC change ratio	ratio of distance
1	232.330			
2	**220.067**	-12.264	1.000	1.538
3	227.990	7.923	-0.646	2.320
4	257.268	29.277	-2.387	1.304
5	290.320	33.053	-2.695	1.210
6	325.522	35.201	-2.870	1.138
7	361.970	36.448	-2.972	1.298
8	400.484	38.514	-3.140	1.252
9	440.395	39.911	-3.254	1.024
10	480.438	40.043	-3.265	1.245
11	521.547	41.109	-3.352	1.003
12	562.667	41.120	-3.353	1.229
13	604.595	41.928	-3.419	1.361
14	647.458	42.863	-3.495	1.129
15	690.618	43.159	-3.519	1.575

In addition to the optimum cluster provided by the two step clusters, the researchers compared the number of cluster that equal to the result of clustering of cluster ensemble i.e. 5 cluster. The value ratio of diversity within and between the cluster (S_w / S_b) to the number of cluster 2 is 1.457, while value ratio of diversity within and between the cluster (S_w / S_b) to the number of group 5 is 0.875. Of the 2 diversity ratios, the smallest ratio S_w / S_b were at number of cluster 5. The number of selected cluster in the two step cluster method was 5 cluster.

C. Comparison of Clustering Result of Cluster Ensemble Methods and Two Step Cluster Methods

The clustering using cluster ensemble and two step cluster produces the same number of cluster, i.e. 5 cluster. The ratio of diversity within cluster and the diversity between cluster (S_w/S_b) in cluster ensemble and two step cluster methods are presented in Table 4.

Table 4. The ratio of diversity within cluster and diversity between cluster of cluster ensemble and two step cluster methods

Variable	S_w/S_b	
	Cluster Ensemble	Two Step Cluster
X_1	0.855	0.928
X_2	1.137	1.216
X_3	0.988	1.062
X_4	1.008	1.482
X_5	0.244	0.215
Y_1	0.287	0.000
Y_2	0.288	0.000
Y_3	0.698	0.934
Y_4	0.698	2.040
Rata-rata	**0.689**	0.875

From Table 4, it can be seen that the ratio of diversity within cluster and diversity between cluster in the cluster ensemble method gives a smaller value than the two step cluster method. In this case, cluster ensemble method is better used to cluster the sub-district in Malang regency based on health facilities and health personnel.

IV. CONCLUSION

In this case, the cluster ensemble method performs better in clustering mixed data than the two step cluster method based on the ratio of diversity within cluster and diversity between cluster.

V. ACKNOWLEDGEMENTS

The authors would like to acknowledge the valuable comments and suggestions of the Editors. These led to a considerable improvement in the paper. This work is financially supported by National Statistical Office (BPS) scholarship program in collaboration with Department of Statistics at Bogor Agricultural University.

VI. 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, pp. 283-304, Sept. 1998.

[2]. C. Li and G. Biswas. "Unsupervised Learning with Mixed Numeric and Nominal Data". IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 4, pp. 673-690, Jul/Aug. 2002.

[3]. SPSS Inc. White paper – technical report, "The SPSS Two Step Cluster Component". 2001.

[4]. Z. He, X. Xu, S. Deng. "Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach". High Technology Letters, vol. 9, no. 4, Oct. 2005.

[5]. R.R. Dewangan, L.K. Sharma, A.K. Akasapu. "Fuzzy Clustering Technique for Numerical and Categorical Dataset". International Journal on Computer Science on Engineering, pp. 75-80, 2010.

[6]. Kementerian Kesehatan Republik Indonesia. "Data SDM Kesehatan yang Didayagunakan di Fasiitas Pelayanan Kesehatan (Fasyankes)." Internet: http://bppsdmk.kemkes.go.id/info_sdmk/info/, May. 9, 2018].

[7]. R. Ghaemi, M.N. Sulaiman, H. Ibrahim, N. Mustapha. "A Survey: Clustering Ensembles Techniques", in Proceedings of World Academy of Science, Engineering and Technology, 2009, pp. 636-645.

[8]. A. Strehl and J. Ghosh. "Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions". Journal of Machine Learning Research, vol. 3, pp. 583-617, Feb. 2002.

[9]. Z. He, X. Xu, S. Deng. "Squeezer: An Efficient Algorithm for Clustering Categorical Data". Journal Computer Science and Technology, vol. 17, no. 5, Sept. 2002.

[10]. R.A. Johnson and D.W. Wichern. Applied Multivariate Statistical Analysis Fifth Edition. New Jersey: Prentice Hall, 2002, pp. 31.

[11]. 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, Oct. 2014.

[12]. J. Bacher, K. Wenzig, M. Vogler. "SPSS Two Step Cluster – a First Evaluation". Available at https://www.ssoar.info/ssoar/bitstream/handle/document/32715/ssoar-2004-bacher_et_al-SPSS_TwoStep_Cluster_-_a.pdf?sequence=1. 2004.

[13]. M.J. Bunkers and J.R. Miller. "Definition of Climate Regions in the Northern Plains Using an Objective Cluster Modification Technique". Journal of Climate, vol. 9, pp. 130-146, Jan. 1996.

[14]. A. Dewi. "Metode Cluster Ensemble untuk Pengelompokkan Desa Perdesaan di Provinsi Riau". M.A. thesis, Institut Teknologi Sepuluh Nopember, Surabaya, 2012.