ICIC Express Letters
Volume **12**, Number **10**, October **2018**
ICIC International ©2018 ISSN 1881-803X
pp. **973–980**

COMPARISON OF PARAMETER-FREE AGGLOMERATIVE HIERARCHICAL CLUSTERING METHODS

Agung Triayudi and Iskandar Fitri

Informatic Department
Universitas Nasional
Jalan Sawo Manila, Pejaten Barat, Pasar Minggu, Jakarta 12520, Indonesia
{ agungtriayudi; iskandar.fitri }@civitas.unas.ac.id

Received March 2018; accepted June 2018

Abstract. *In this paper, new algorithms are introduced within the scope of agglomerative hierarchical clustering free parameters, i.e., ALG (Average Linkage Dissimilarity Increment Distribution-Global Cumulative Score Standart) algorithm. This algorithm incorporates the cluster isolation dissimilarity increment technique and the cumulative clustering technique between clusters globally. The result of ALG algorithm test on iris dataset, wine dataset and WDBC (Wisconsin Diagnostic Breast Cancer) dataset using cophenetic correlation validity value of 0.8693; 0.7708; 0.8337 and the value of silhouette coefficient yields a value of 0.6785; 0.6278; 0.6787, this result shows ALG algorithm outperformed LSS-GCSS (Local Standart Score-Global Cumulative Score Standart) algorithm in previous research.*
Keywords: ALG algorithm, Cluster, Dataset, Result, LSS-GCSS

1. **Introduction.** Clustering method is a fundamental problem that has been the focus of great research in machine learning. Clustering is automatically formed by groups of objects that are all interconnected. Therefore, the similarity between objects assigned to the same cluster tends to be larger than in different groups [1,2]. Clustering is widely used in many different fields such as astronomy, medicine, economics, weather, finance and others [3].

Clustering is an important approach to finding commonality in data and placing the same data into different clusters. Clustering divides the data set into multiple clusters where the similarities in a group are larger than the different clusters [6]. The idea of data clustering has a simple nature and is similar to the pattern of human thinking [4]. When we are given large amounts of data representation, we usually tend to summarize this large amount of data into a small number of groups or categories for further analysis [7]. In addition, most of the data collected in various problems will be seen to have some inherent properties built on natural clusters [8].

Clustering hierarchy builds a cluster hierarchy or, in other words, a cluster tree, also known as dendrogram. The hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) [9,10]. Agglomerative clustering begins with a single point cluster (singleton) and repeatedly combines two or more most appropriate clusters. The divisive cluster starts with one cluster of all data points and repeats the most appropriate clusters. The process continues until the termination criterion (the number of k required from the cluster) is reached.

Research on cluster isolation has been done much earlier, among which [12] is about the new criteria for cluster isolation based on the assumption of the inequality between neighboring patterns within the cluster. The proposed criteria lead to potential dendrograms that are different from those obtained by standard hierarchical procedures, based on the

DOI: 10.24507/icicel.12.10.973

effect of pruning on the dendrogram branch. However, this study did not explain the evaluation and validation of the proposed algorithm, so it has not been tested properly and correctly.

Another study of Parameter-free in clustering [18], Parameter Free Minimum Spanning Tree (MST) clustering, that is used is based on two processes with minimum user intervention; splitting the initial MST to get rough clustering, then fine tuning is done through merging the neighboring clusters. This research demonstrated that MST based clustering algorithm outperforms other clustering algorithms, including AHC and k-means on datasets from the UCI machine learning repository. A fundamental weakness in this technique is that it is not able to process datasets on a large scale and has not been tested on other validation techniques such as cophenetic correlation coefficient and silhouette coefficient.

Continued research on cluster isolation [13] that concerns different forms or different data separation paradigms can be adequately modeled by exponential distributions when analyzing differences in inequality between neighboring patterns; the average value of the parametric model is closely related to the scarcity of data, regardless of the shape of the orientation or its form. The number of clusters is intrinsically found without requiring design parameter specifications or involving optimization processes that require computation. Although succeeding in minimizing the gap increment in the dendrogram, this second generation cluster isolation algorithm still has not used the value of evaluation or validation of each cluster.

Another study of cluster isolation was also proposed [15] on a family with an agglomerative hierarchical method based on high-order dissimilarity. The advantage of this method compared to the traditional relationship algorithm is that they can automatically find the number of clusters using the minimum description criteria. By comparison, traditional algorithms require users to set the number of clusters or use some external criteria to find them. This property leads to significant algorithmic performance improvements over traditional grouping algorithms.

Recent research [16] on the issue of modeling the activities of learners in online discussion forums from a cluster-based perspective, led to highly context-dependent analysis scenarios in which the actual number of clusters is a priori unknown. In order to avoid user intervention in estimation number of clusters, which can easily lead to the emergence of undesirable biases in the model obtained. The experimental results showed that the LSS-GCSS (Local Standart Score-Global Cumulative Score Standart) algorithm is able to provide optimal clustering solutions in the face of various grouping scenarios. The LSS-GCSS algorithm is tested on UCI machine learning dataset, i.e., iris, wine and WDBC yielding validation values of cophenetic correlation 0.568; 0.726; 0.4395 and the validation value of silhouette coefficient 0.484; 0.392; 0.456.

It is observed from the study literature that the cluster isolation algorithm has evolved from minimizing the gap margin of each cluster within the dendrogram and being able to properly estimate the number of clusters, but in the validation of cophenetic correlation coefficient and silhouette coefficient still get a small value, and this means there is still an irreparable gap. This paper proposes ALG (Average Linkage Dissimilarity Increment Distribution-Global Cumulative Score Standart) combination algorithm that will improve gap in validation value of cophenetic correlation coefficient and silhouette coefficient.

2. **A New Algorithm of Hierarchical Clustering.**

In this paper, it is introduced of a new hierarchical clustering algorithm namely ALG (Average Linkage Dissimilarity Increment Distribution-Global Cumulative Score Standart).

This new algorithm is the result of a combination of AHC (Agglomerative Hierarchical Clustering) based on DID (Dissimilarity Increment Distribution) [15] and parameter-free algorithm GCSS (Global Cumulative Score Standart) [16].

Algorithm 1: ALG Algorithm

Input: dataset X and parameter H
procedure
$M_p : M_p(i, j)$
Select the most similar clusters $(C_i, C_j) minDist = \min\{d(x_i, x_j) : x_i \in C_i, x_j \in C_j\}$
if $|C_i| < H$ and $|C_j| < H$ then
    Merge clusters C_i, C_j into a new cluster C_b using ALDID (Equation (4)) and GCSS (Equation (7))
end if
if $|C_i| \geq H$ and $|C_j| < H$ then
    if $dissinc(x_i, x_j, x_k) = |d(x_i, x_j) - d(x_j, x_k)|$ of (C_j) is not in the tail then
    the $pdissinc(w; \lambda)$ (Equation (2)) then $dissinc(x_i, x_j, x_k) = |d(x_i, x_j) - d(x_j, x_k)|$ of (C_i) then
    Merge clusters C_i, C_j into a new cluster C_b using ALDID (Equation (4)) and GCSS (Equation (7))
    else
    Do not merge C_i, C_j
    end if
end if
if $|C_i| \geq H$ and $|C_j| \geq H$ then
    Compute gap $C_i(C_j)$ and gap $C_j(C_i)$
    Compute $DC(C_i)$, $DC(C_j)$ and $DC(C_i \cup C_j)$
    if gap $C_i(C_j)$ is in the tail of the $pdissinc(w; \lambda)$ (Equation (2)) then
    $disinc(x_i, x_j, x_k) = |d(x_i, x_j) - d(x_j, x_k)|$ of (C_i) then
    Freeze cluster C_i
    else if gap $C_j(C_i)$ is in the tail of the $pdissinc(w; \lambda)$ (Equation (2)) then
    $disinc(x_i, x_j, x_k) = |d(x_i, x_j) - d(x_j, x_k)|$ of (C_j) then
    Freeze cluster C_j
    else if $DC(C_i \cup C_j) \leq DC(C_i) + DC(C_j)$ then
    Merge clusters C_i, C_j into a new cluster C_b using ALDID (Equation (4)) and GCSS (Equation (7))
    else
    Do not merge C_i, C_j
    end if
end if
until all pairs of clusters should not be merged

2.1. **First step.** Determining the proximity matrix (M_p) where the AHC method starts with every single object in one cluster (single cluster M) and performs a series of merging operations ($M - 1$ merging steps) [16].

$$M_p(X) = \begin{pmatrix} 0 & d_{x_1x_2} & \cdots & d_{x_1x_2} \\ d_{x_1x_2} & 0 & \cdots & d_{x_1x_2} \\ \vdots & \vdots & \ddots & \vdots \\ d_{x_1x_2} & d_{x_1x_2} & \cdots & 0 \end{pmatrix} \tag{1}$$

2.2. **Second step.** The DID was derived, using the Euclidean distance as the dissimilarity measure $d(\cdot, \cdot)$, under the hypothesis of Gaussian distribution of data. This distribution was written as a function of the mean value of the dissimilarity increments, which is denoted as λ [15].

$$pdissinc(w;\lambda)$$
$$= \frac{\pi\beta^2}{4\lambda^2} w \exp\left(-\frac{\pi\beta^2}{4\lambda^2}w^2\right) + \frac{\pi^2\beta^3}{8\sqrt{2}\lambda^3} X \left(\frac{4\lambda^2}{\pi\beta^2} - w^2\right) \exp\left(-\frac{\pi\beta^2}{8\lambda^2}w^2\right) erfc\left(\frac{\sqrt{\pi}\beta}{2\sqrt{2}\lambda}w\right) \quad (2)$$

Specify the merging criteria based on AHC-DID [15].

- It is considered that C_j has M minus patterns and M patterns have more, if the mean of the addition of C_j is less than the average α of C_i, i.e., the increase of C_j at the tail of the DID C_i. If it does not fall on the tail, the C_i and C_j clusters are combined; if not, then it keeps separated.
- Now, suppose C_i and C_j already have M or more patterns. So, check if gap $C_i(C_j)$ is behind the DID cluster C_i. When that happens, C_i is "frozen", meaning C_i is no longer available for merging with other groups. Similarly, tests for C_j with respect to C_i are performed, but only if the preceding C_i is not "frozen". Here it only allows one cluster to be "frozen" in each algorithm iteration.
- In the end, if C_i or C_j is "not frozen", for the cluster yielded from the merging of C_i and C_j, $C_i \cup C_j$, is calculated by the same procedure, with the assumption that λ_{ij} is the parameter of DID for cluster $C_i \cup C_j$. Now, if DC $C_i \cup C_j$ has lower value of $DC(C_i) + DC(C_j)$ (the length of description to leave cluster which is separated), cluster C_i and C_j is merged and forms new cluster; if not, the group is made separated purposely.

$$DC(C_i) = \frac{1}{2}(1 - \log(12)) + \log\lambda_i + \frac{1}{2}\log(I(\lambda_i)) - \log p(w;\lambda_i) \quad (3)$$

2.3. **Third step.** The assumption of the ALDID algorithm is to consider the newly formed cluster, $C_b = C_i \cup C_j$, obtained by combining C_i and C_j, and C_a is one of the remaining groups formed in the preceding steps. Also, let us consider $|C_i|$ and $|C_j|$ as the number of patterns on the C_i and C_j clusters, respectively. We define the ALDID algorithm by characterizing the merging function, according to the size of the $d*(C_a, C_b)$ distance between the clusters [15].

$$d_A(C_a, C_b) = \frac{|C_i|}{|C_i| + |C_j|} d(C_i, C_a) + \frac{|C_j|}{|C_i| + |C_j|} d(C_j, C_a) \quad (4)$$

2.4. **Fourth step.** GCSS algorithm in essence compares the closeness level of a new cumulative hypothetical cluster (cd_k) with the closeness level of cumulative of both prospective groups (cd_i and cd_j). The closeness level of cumulative compared to the context of distribution of cumulative closeness level presented in each cluster, is modeled with the procedure of css_k, css_i and css_j, respectively. Therefore, basically, if cd_k involving an increase in context C_k is higher than both in steps cd_i and cd_j involved in the context of C_i and C_j (namely if css_k is higher than both css_i and css_j), C_i and C_j will not be suited for global combination.

Firstly, let C_x be any given cluster in the dendrogram Δ resulting from the agglomeration process of the objects in X and let $\overline{cd_x}$ be the sample consisting of its own cumulative proximity level (cd_x) and the cumulative proximity levels of its nested clusters in dendrogram Δ. The cumulative standard score statistic of cluster $C_x(css_x)$ is defined as the standard score of cd_x with respect to $\overline{d_x}$ [16]:

$$css_x = \frac{cd_x - c\mu_x}{\sqrt{c\sigma_x - c\mu_x^2}} \quad (5)$$

where $c\mu_x$ and $c\sigma_x$ are the first and second moments of $\overline{cd_x}$ [16]:

$$\mu_x = \frac{1}{n_{dx}} \sum_{l=1}^{n_{dx}} \overline{cd_{xl}}, \quad \sigma_x = \frac{1}{n_{dx}} \sum_{l=1}^{n_{dx}} \overline{cd_{xl}^2} \quad (6)$$

$\overline{cd_{xl}}$ the lth observation in $\overline{cd_x}$ and n_{dx} the length of $\overline{cd_x}$ (i.e., the number of non-singleton clusters nested within C_x).

The GCSS criterion determines that the union between C_i and C_j into a new cluster C_k is a suitable merging if their cumulative standard score statistics (css_i and css_j) are greater than or equal to the following dynamic merging threshold [16]:

$$\begin{aligned} gcss_{th}(C_k, C_i, C_j, Y_{MIN}) &= gcss_{th}(css_k, N_i, \gamma_i, \mu_i, \sigma_i, N_j, \gamma_j, \mu_j, \sigma_j, Y_{MIN}) \\ &= css_k \Upsilon(N_i, N_j) \Psi_G(N_i, \gamma_i, \mu_i, \sigma_i, N_j, \gamma_j, \mu_j, \sigma_j, Y_{MIN}) \end{aligned} \tag{7}$$

where css_k is the cumulative standard score of C_k, $Y_{MIN} = 0.01N$, $\gamma_i = d_{ij} - d_i$, $\gamma_j = d_{ij} - d_j$ and μ_i, σ_i, μ_j and σ_j. The value of Y_{MIN} is defined as 1% of the number of clusters in C ($Y_{MIN} = 0.01Y$).

Therefore, the merging rule derived from the GCSS criterion is defined as follows [16].

- If the GCSS criterion is simultaneously met from both C_i ($css_i \geq gcss_{th}(C_k, C_i, C_j, Y_{MIN})$) and C_j ($css_j \geq gcss_{th}(C_k, C_i, C_j, Y_{MIN})$), C_i and C_j merge into a new cluster.
- Otherwise, the merging between C_i and C_j is rejected in global terms, so that they remain separated.

3. **Evaluation of Clustering Result.** This evaluation is intended to determine the appropriate grouping solution, here using the index validity of silhouete coefficient and cophenetic correlation coefficient.

3.1. **Silhouete coefficient.** The silhouette value for each point is a measure of how similar that point is to points in its own cluster, when compared to points in other clusters. The silhouette value for the ith point, $s(i)$, is defined as [14]

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{8}$$

The range of silhouette is $[-1, 1]$ [14].

3.2. **Cophenetic correlation coefficient (CPCC).** Cophenetic correlation coefficient measures the degree of similarity between P_c and the proximity matrix P. The cophenetic matrix P_c is defined in such a way that the element $P_c(i, j)$ represents the proximity level at which the two data points x_i and x_j are found in the same cluster for the first time. The CPCC index is defined as [5]

$$CPCC = \frac{\frac{1}{M}\sum_{i=1}^{n=1}\sum_{j=i+1}^{n} d_{ij}c_{ij} - \mu_p\mu_c}{\sqrt{\left(\frac{1}{M}\sum_{i=1}^{n=1}\sum_{j=i+1}^{n} d_{ij}^2 - \mu_p^2\right)\left(\frac{1}{M}\sum_{i=1}^{n=1}\sum_{j=i+1}^{n} c_{ij}^2 - \mu_c^2\right)}} \tag{9}$$

where $M = \frac{n(n-1)}{2}$ and μ_p, μ_c:

$$\mu_p = \frac{1}{M}\sum_{i=1}^{n=1}\sum_{j=i+1}^{n} d_{ij}, \quad \mu_c = \frac{1}{M}\sum_{i=1}^{n=1}\sum_{j=i+1}^{n} c_{ij} \tag{10}$$

where d_{ij} and c_{ij} are the (i, j) elements of matrices P and P_c, respectively. The $CPCC$ ranges from -1 to $+1$. The high value indicates great similarity between P and P_c [5].

4. **Result and Analysis.** This study uses the dataset of the UCI machine learning repository [17], i.e.:

- Iris (N = 150)
- Wine (N = 178)
- Wisconsin diagnostic breast cancer (WDBC) (N = 96)

Cluster analysis

From the clustering algorithm analysis of ALG algorithm there are 2 clusters in this dataset, dendrogram in Figure 1 shows this algorithm can well be mapped every existing object and compared with algorithm from previous research that LSS-GCSS showed a significant increase, from validation value of cophenetic correlation coefficient experiences an increase of ≥ 0.3 from the previous study and from the validation value of silhouette coefficient increased ≥ 0.19 from the previous study.

In the wine dataset Figure 2, there is a greater amount of data than the iris dataset and in the clustering context of objects within the denser wine density is more dense and varied, from the results of ALG clustering algorithm analysis there are 2 clusters

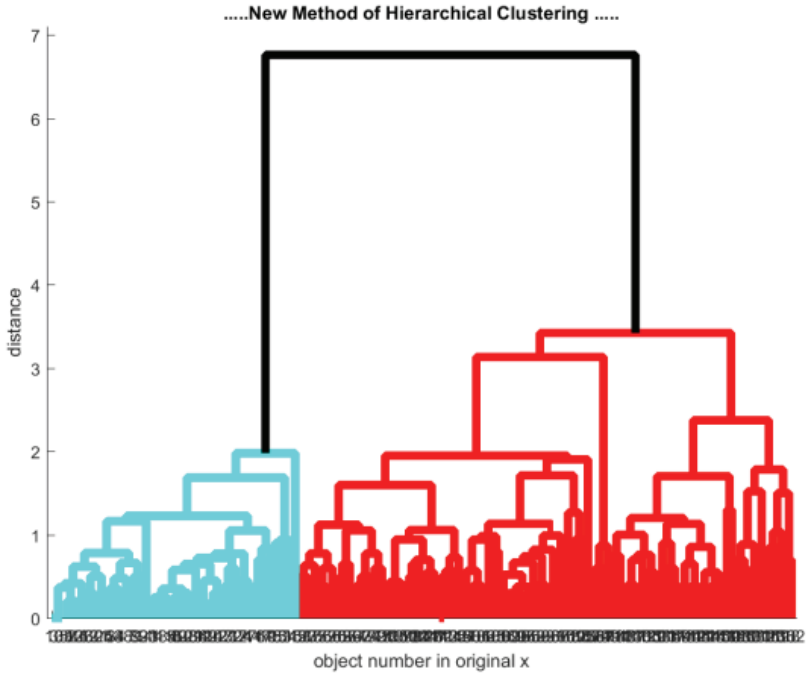

FIGURE 1. Dendrogram dataset iris ALG algorithm

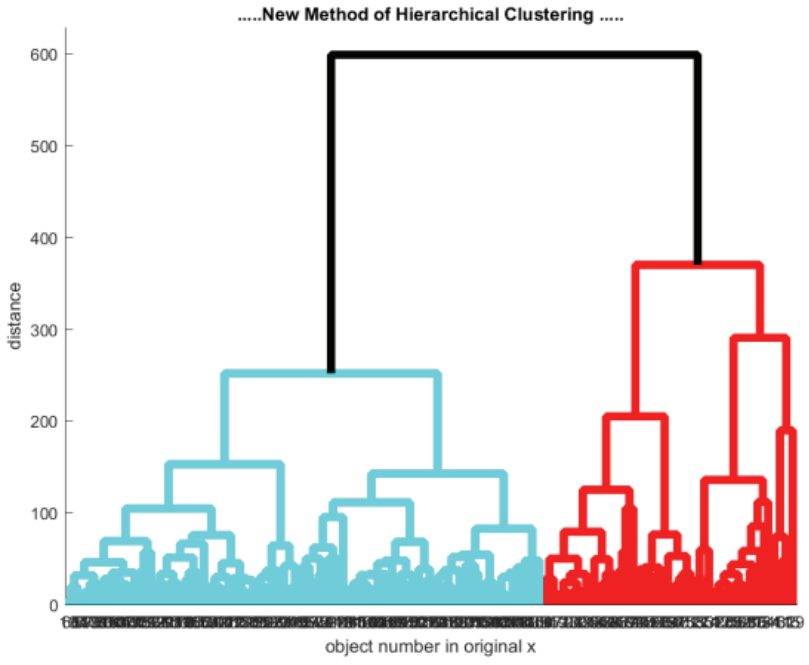

FIGURE 2. Dendrogram dataset wine ALG algorithm

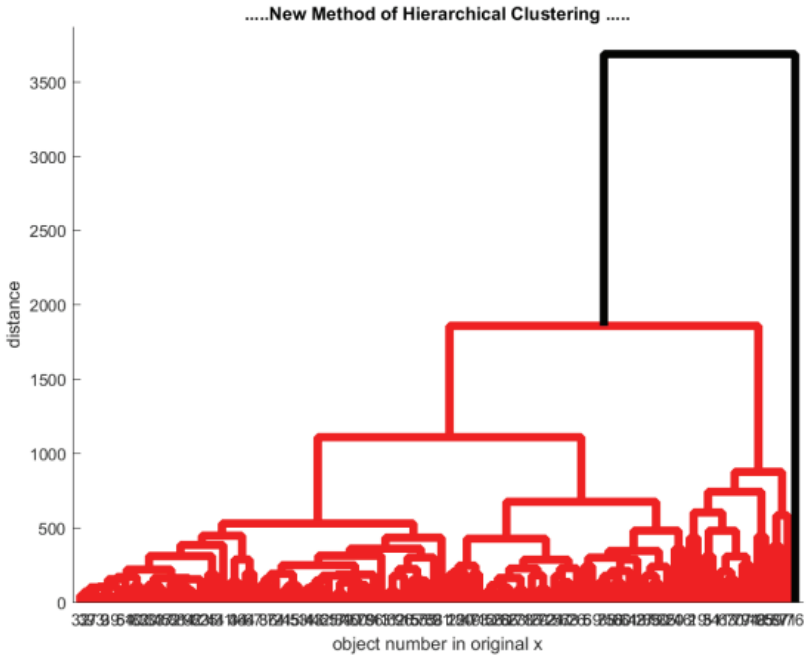

FIGURE 3. Dendrogram dataset WDBC ALG algorithm

TABLE 1. The results of the comparison of the validity of the silhouette coefficient and the cophenetic correlation coefficient, the LSS-GCSS algorithm and the ALG algorithm

	LSS-GCSS		*ALG*	
	CPCC	*S*	*CPCC*	*S*
Iris	0.568	0.484	0.8693	0.6785
Wine	0.726	0.392	0.7708	0.6278
WDBC	0.4395	0.456	0.8337	0.6787

in this dataset and compared with the algorithm from the previous research LSS-GCSS showed a significant increase, from the validation value of cophenetic correlation coefficient increased ≥ 0.04 from the previous study and from the validation value of silhouette coefficient increased ≥ 0.23 from previous research.

In the WDBC dataset Figure 3, from the ALG clustering algorithm analysis results, there are 2 clusters in this dataset and compared with the algorithm from the previous research, LSS-GCSS showed a significant increase, from the validation value of cophenetic correlation coefficient increased ≥ 0.39 from the previous research and from the validation value of silhouette coefficient, an increase of ≥ 0.2 from the previous study.

Comparison of LSS-GCSS algorithm and ALG algorithm can be seen in Table 1. From calculation of validation value using cophenetic correlation coefficient by testing in three dataset variables results in increased precision by ALG algorithm of +0.3013 on iris dataset, +0.0448 in wine dataset, +0.3942 on the WDBC dataset. An increase also occurring in the validation calculation using silhouette coefficient on three dataset variables results in improved precision by ALG algorithm of +0.1945 on the iris dataset, +0.2358 on the wine dataset, +0.2227 on the WDBC dataset.

5. **Conclusions.** Test results obtained that ALG algorithm successfully improved the flexibility of cluster isolation. This is evidenced by the validation test results using cophenetic correlation coefficient and silhouette coefficient. The result of ALG algorithm comparison with algorithm from previous research is LSS-GCSS. Three datasets tested from

UCI machine learning repository. The ALG algorithm outperforms the LSS-GCSS algorithm across all datasets. From ALG algorithm test result with various dataset categories, this algorithm proved able to overcome various scenarios with stable grouping getting the highest value on measuring value of cophenetic correlation and silhouette.

Acknowledgement. This research was supported by Universitas Nasional as donors. This paper is part of the Ph.D. research in school of graduate studies, Asia e University.

REFERENCES

[1] P. Cichosz, *Data Mining Algorithms: Explained Using R*, John Wiley & Sons, 2014.
[2] A. K. Mann and N. Kaur, Review paper on clustering techniques, *Global Journal of Computer Science and Technology*, 2013.
[3] Z. Nazari et al., A new hierarchical clustering algorithm, *International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, 2015.
[4] Y. Zhao, G. Karypis and U. Fayyad, Hierarchical clustering algorithms for document datasets, *Data Mining and Knowledge Discovery*, vol.10, no.2, pp.141-168, 2005.
[5] G. Gan, C. Ma and J. Wu, *Data Clustering: Theory, Algorithms and Applications*, American Statistical Association and the Society for Industrial and Applied Mathematics, PA, Pennsylvania, USA, 2007.
[6] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy. The Principles and Practice of Numerical Classification*, 1973.
[7] G. Karypis, E.-H. Han and V. Kumar, Chameleon: Hierarchical clustering using dynamic modeling, *Computer*, vol.32, no.8, pp.68-75, 1999.
[8] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.
[9] C. T. Zahn, Graph-theoretical methods for detecting and describing gestalt clusters, *IEEE Trans. Computers*, vol.100, no.1, pp.68-86, 1971.
[10] C. C. Aggarwal et al., Fast algorithms for projected clustering, *Proc. of the 1999 ACM SIGMOD International Conference on Management of Data*, vol.28, no.2, 1999.
[11] M. Steinbach, G. Karypis and V. Kumar, A comparison of document clustering techniques, *KDD Workshop on Text Mining*, vol.400, no.1, 2000.
[12] A. L. N. Fred and J. M. N. Leitao, Clustering under a hypothesis of smooth dissimilarity increments, *Proc. of the 15th International Conference on Pattern Recognition*, vol.2, 2000.
[13] A. L. N. Fred and J. M. N. Leitao, A new cluster isolation criterion based on dissimilarity increments, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.25, no.8, pp.944-958, 2003.
[14] P. J. Rouseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, vol.20, no.1, 1987.
[15] H. Aidos and A. Fred, A family of hierarchical clustering algorithms based on high-order dissimilarities, *Proc. of the 22nd European Signal Processing Conference (EUSIPCO)*, 2014.
[16] G. C. Rodríguez, *Parameter-Free Agglomerative Hierarchical Clustering to Model Learners' Activity in Online Discussion Forums*, Ph.D. Thesis, Internet Interdisciplinary Institute, 2014.
[17] M. Lichman, *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]*, University of California, School of Information and Computer Science, Irvine, CA, 2013.
[18] R. Raju, Bhvs and V. Kumari, Comparison of parameter free MST clustering algorithm with hierarchical agglomerative clustering algorithms, *International Journal of Computer Applications*, vol.34, no.4, pp.26-31, 2011.