EEPIS Journal, Vol. 2, No. 9, 2004.
ISSN: 0852-2863

Ali Ridho Barakbah, Kohei Arai, "Reversed pattern of moving variance for accelerating automatic clustering", EEPIS journal, pp.15-21, Vol. 9, No. 2, December 2004, Surabaya, Indonesia.

Reversed pattern of moving variance for accelerating automatic clustering

Ali Ridho Barakbah and Kohei Arai
Electronic Engineering Polytechnic of Surabaya and Saga University, ridho@eepis-its.edu

Abstract

This paper proposes a new approach to accelerate a construction of automatic clusters. It initiates an analyse of the moving variance of clusters for each stage of cluster construction, then observes the pattern to find the global optimum as well as avoid the local optima. Using two constraints, valley-tracing and hill-climbing, to find the global optimum, the proposed approach reverses the pattern in order to accelerate the automatic clustering. Experiment result performs the effectiveness of the proposed approach in this paper.

1. Introduction

Clustering is an exploratory data analysis tool that deals with the task of grouping objects that are similar to each other [2, 6, 12]. For many years, many clustering algorithms have been proposed and widely used. It can be divided into two categories, hierarchical and non-hierarchical methods. It is commonly used in many fields, such as data mining, pattern recognition, image classification, biological sciences, marketing, city-planning, document retrieval, etc. The clustering means process to define a mapping f:D→C from some data D={t_1, t_2, …t_n} to some clusters C={c_1, c_2,…, c_n} based on similarity between t_i.

The task of finding a good cluster is very critical issues in clustering. Cluster analysis constructs good clusters when the members of a cluster have a high degree of similarity to each other (internal homogeneity) and are not like members of other clusters (external homogeneity) [3, 8]. In fact, most authors find difficulty in describing clustering without some suggestions for grouping criteria. For example, "the objects are clustered or grouped based on the principles of maximizing the inter-class similarity and minimizing the intra-class similarity" [8]. One of the methods to define a good cluster is variance constraint [7] that calculates the cluster density with variance within cluster (V_w) and variance between clusters (V_b) [4, 12]. The ideal cluster has minimum V_w to express internal homogeneity and maximum V_b to express external homogeneity.

It is common that most of the clustering methods require the users to provide the number of clusters as input. But, in some clustering cases the users have not an idea to determine the number of clusters. Hence, they usually try it with different number of clusters. It makes very difficult, especially if the clustering case is not easy to observe. A genetic algorithm was proposed to search optimal clusters [1]. But, it still requires the user to provide the number of clusters in a priori. Tseng and Yang proposed a genetic clustering algorithm [5]. The clustering algorithm will automatically search for proper number of clusters and classify the objects into these clusters at the same time. However, before using the genetic clustering, this algorithm utilized the single linkage hierarchical method to reduce the size of data set if the size is large. In 2004, Barakbah and Arai proposed a new approach to make automatic clustering with purely utilizing the single linkage hierarchical method [9]. The algorithm identify the moving average of cluster construction for each stage. In this paper, we propose an improved approach with two constratints, valley-tracing and hill-climbing, to find the global optimum and make the automatic clustering with analyzing the moving average. Besides it improves the acceleration of automatic clustering with reversing the pattern of moving variance.

The remaining part of the paper is organized as follows. In Section 2, the basic concept of single linkage hierarchical algorithm is introduced. In Section 3, the cluster density is described. Section 4 describes the two constraints, valley-tracing and hill-climbing. Experimental results of the two constraints as well as the applicability to make automatic clustering is described in Section 5. Section 6 describes the improvement of the approach with reversing the pattern of moving variance to accelerate the automatic clustering. The paper is concluded in Section 7.

2. Single linkage hierarchical algorithm

One of the most famous methods in clustering is that classified method as hierarchical clustering. In hierarchical clustering the data are not partitioned into a particular cluster in a single step. It runs with making a single cluster that has similarity, and then continues iteratively. Hierarchical clustering algorithms can be either agglomerative or divisive [6, 10, 11]. Agglomerative method proceeds by series of fusions of the "n" similar objects into groups, and divisive method, which separate "n" objects

EEPIS Journal, Vol. 2, No. 9, 2004.
ISSN: 0852-2863

Ali Ridho Barakbah, Kohei Arai, "Reversed pattern of moving variance for accelerating automatic clustering", EEPIS journal, pp.15-21, Vol. 9, No. 2, December 2004, Surabaya, Indonesia.

successively into finer groupings. Agglomerative techniques are more commonly used.

One of similarity factors between objects in hierarchical methods is a single link that similarity closely related to the smallest distance between objects [2]. Therefore, it is called single linkage hierarchical algorithm. Euclidian distance is commonly used to calculate the distance in case of numerical data sets [11]. For two dimensional dataset, it performed as:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \quad (1)$$

The algorithm of single linkage clustering method is composed of the following steps:

1. Begin with an assumption that every point "n" is it's own cluster c_i, where i=1..n.
2. Find the nearest distance between $m(c_r)$ and $m(c_u)$, where $r \neq u$ and $m(c_j)$ is members of cluster c_j.
3. Merge c_r and c_u into new cluster c_a where $m(c_a)$ is members fusion of c_r and c_u.
4. Repeat until it reaches optimum

3. Cluster density

The density of cluster can be determined by the variance within cluster and variance between clusters. The ideal cluster has a low variance within cluster and a high variance between clusters [4, 12].

If there is some cluster c_i, where i=1…k, and each of them have members x_i, where i=1…n and n is total members of each clusters, and $\bar{x}_p$ is the centroid of cluster p. Then, the variance of cluster p (δ_p^2) can be calculated as:

$$\delta_p^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}_p\right)^2 \quad (2)$$

If N is total numbers of members in all clusters, variance within cluster (v_w^2) can be defined as:

$$v_w^2 = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1)\delta_i^2 \quad (3)$$

Then, variance between clusters (v_b^2) quantifies the variability of the group mean around the grand mean ($\bar{x}$), and hence the component of group differences. It is defined as:

$$v_b^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \quad (4)$$

Because an ideal cluster has minimum v_w^2 and maximum v_b^2, it means the ideal cluster has minimum v, where:

$$v = \frac{v_w^2}{v_b^2} \times 100\% \quad (5)$$

However, eventhough minimum v expresses the ideal cluster, we can not apply it directly to find the global optimum. There are some experiments proves that in some cases, minimum v reaches the local optima of cluster construction. For example, in case of Fig. 1 with n=50, minimum v=0.15 resides in stage 1 with 49 total cluster. Stage 2 performs v=0.18 with 44 total cluster. But, actually the ideal cluster resides in stage 15 with 6 total clusters where v=0.22.

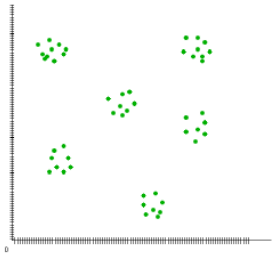

Figure 1. A case of clustering

Therefore, minimum v can not be used directly to find the global optimum. If we force to apply minimum v directly to identify the global optimum, in some cases, it may fall in local optima. To solve this problem, this paper proposes the new approach to find the global optimum and avoid the local optima.

4. Valley-tracing and hill-climbing approach

4.1. Identifying pattern of moving variance

Single linkage hierarchical algorithm is very thorough to make analysis every states of cluster construction stage by stage. Therefore, this paper used the single linkage as appropriate method in order to identify the moving variance from each stages of cluster construction.

Figure 2 shows the moving variance from each stages of cluster construction of case performed in Fig. 1. There we can also see that the global optimum resides in stage 15, with 6 total cluster.

For finding the global optimum of cluster construction and avoid the local optima, we propose a new approach to solve the case. First of all we try to describe all patterns of the moving variance, then analyze the possibility of the global optimum that resides in certain places.

EEPIS Journal, Vol. 2, No. 9, 2004.
ISSN: 0852-2863

Ali Ridho Barakbah, Kohei Arai, "Reversed pattern of moving variance for accelerating automatic clustering", EEPIS journal, pp.15-21, Vol. 9, No. 2, December 2004, Surabaya, Indonesia.

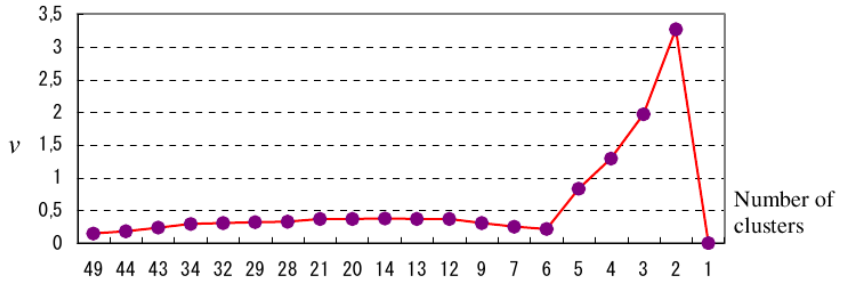

Figure 2. Moving variance of cluster construction for each stages

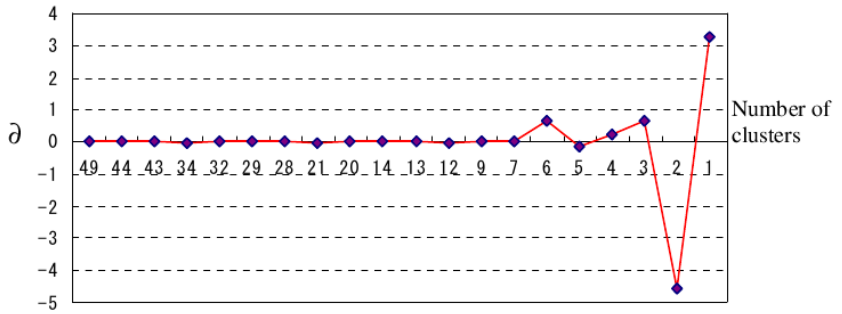

Figure 3. Transformation of moving variance into differential values

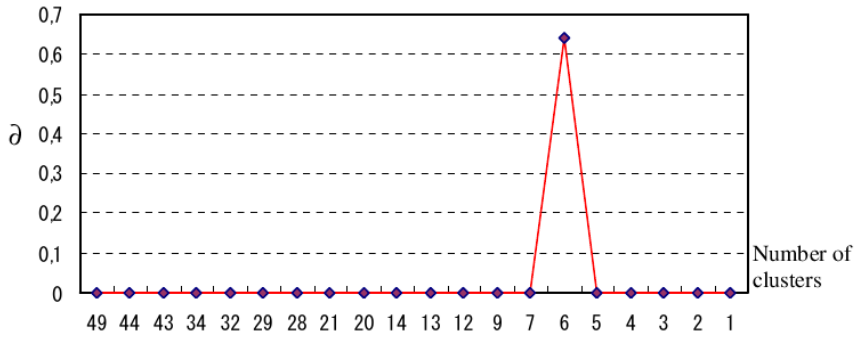

Figure 4. Differential values with hill-climbing (α=2)

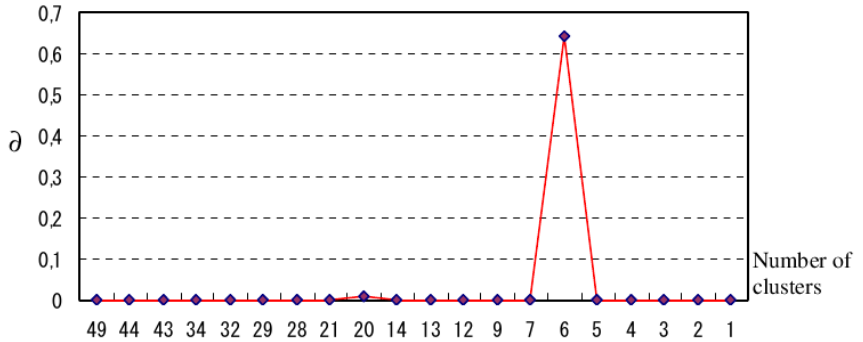

Figure 5. Differential values with valley-tracing

EEPIS Journal, Vol. 2, No. 9, 2004.
ISSN: 0852-2863

Ali Ridho Barakbah, Kohei Arai, "Reversed pattern of moving variance for accelerating automatic clustering", EEPIS journal, pp.15-21, Vol. 9, No. 2, December 2004, Surabaya, Indonesia.

Then, we look for position of the possible global optimum and classify into two contraints, hill-climbing and valley-tracing.

4.2. Hill-climbing

Based on many experiments, the global optimum usually resides in the stage that has far different value with its next stage. If it is visualized, it seems like a hill, as figured in Fig. 2. The effort to find the global optimum considers to climb the hill for each cluster construction. It has altitude value to determine how possible the climbing hill to be a global optimum. Then, we describe that the possibility to find the global optimum by hill-tracing resides in stage *i* fulfilled:

$$v_{i+1} > \alpha \cdot v_i \quad (6)$$

where α is altitude value. In this paper, we use 3 different values of altitude, α=2,3 and 4.

4.3. Valley-tracing

In this way, the possible optimum place can be traced in the valley of moving variance. From analyzing the pattern of moving varince, we describe that the possibility to find the global optimum by valley-tracing resides in stage fulfilled:

$$(v_{i-1} \geq v_i) \cap (v_{i+1} > v_i) \quad (7)$$

for i=1…n, and n is latest stages of cluster construction.

4.4. Considering differential value

As we described before that minimum v can not express value of global optimum. It needs transformation into certain value recognized absolutely as considerable value of global optimum.

The differential value between v for each stages can be considered [9]. Fig. 6 shows the differential value of v_i.

Then, we identify the differential value of altitude ∂ for each stages. It can be defined as:

$$\begin{aligned} \partial &= (v_{i+1} - v_i) + (v_{i-1} - v_i) \\ &= (v_{i+1} + v_{i-1}) - (2 \,.\, v_i) \end{aligned} \quad (8)$$

In order to avoid the local optima and find the global optimum, it can be derived from maximum of ∂ that fulfilled Eq. (8).

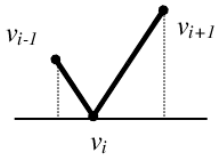

Figure 6. Differential value of v_i

Fig. 3 performs the trasformation of moving variance shown by Fig. 2 into differential values. From Fig. 3 we can see that the global optimum has been not determined yet because it may perform the closeness of values in some stages. Therefore we apply the two constraints, valley-tracing and hill-climbing. Fig. 4 and Fig. 5. performs the result of applied hill-climbing and valley-tracing in the moving average. We see that the global optimum of differential values after applying the two constraints can be determined as well as avoid the local optima because the uniqueness of highest values in the differential values.

4.4. Making automatic clustering

To construct cluster automatically, we put the additional variable λ as a threshold value to get a maximum ∂. The more complex clustering case needs smaller λ to set as more precise as possible. By setting the value of λ, the well-separated cluster will be constructed. In this paper, we use various values of λ from 0.05-0.5.

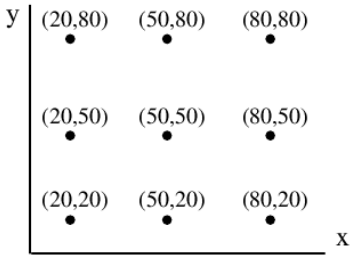

Figure 7. Ilustration of 9 positions to generate random normal data distribution.

5. Experimental result

We apply our proposed approach in the random normal data distribution. For experimental purpose, we use 2 dimentional data set (x and y). Then, we determine 9 nodes, as figured in Fig. 7, as positions to generate randomly each data clusters.

EEPIS Journal, Vol. 2, No. 9, 2004.
ISSN: 0852-2863

Ali Ridho Barakbah, Kohei Arai, "Reversed pattern of moving variance for accelerating automatic clustering", EEPIS journal, pp.15-21, Vol. 9, No. 2, December 2004, Surabaya, Indonesia.

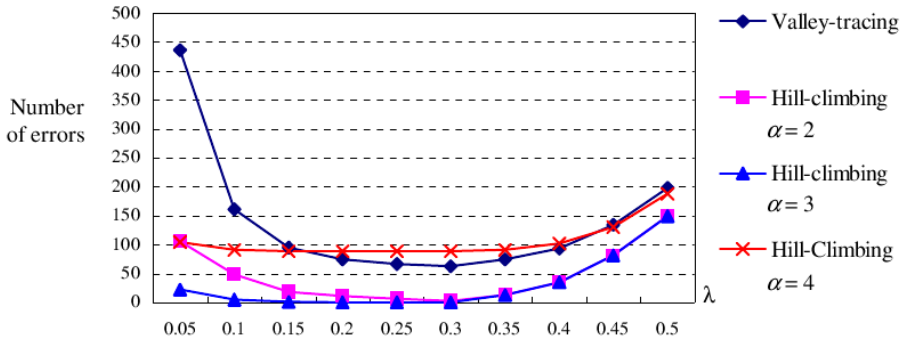

Figure 8. Comparing four constraints for automatic clustering

Those nodes are generated randomly with 9 maximum nodes, with minimum total data=6 and minimum total clusters=3. Next, we generate random cluster data distribution around each nodes position inside diameter=5 because we want to intent well-separated clusters. Numbers of data for each them are generated randomly and ≤ 10. With this model, it can generate thousands of different combination for normal data distribution. In this paper, we made 1000 experiments.

In the experiment, we involved four constraints, valley-tracing and three of hill-climbing with different altitude, α=2,3, and 4. We compute the error percentage of each constraints. We also compute the gap distance (φ) between global optimum and candidate global optimum. The high φ express the high possibility to apply threshold in order to make automatic clustering. Table 1 performs the error percentage and φ from 1000 experiments.

Table 1
Comparing four constraints

	Error (%)	φ
Valley-tracing	8.9	0.7861
Hill-climbing, α =2	29.6	10.946
Hill-climbing, α =3	13.6	11.353
Hill-climbing, α =4	12.6	10.512

We can see in Table 1 that Valley-tracing is superior than the others because the constraint considers not only the next stage from current stage of cluster construction, but also involves the previous stage. It is able more to filter the global optimum and eliminate the local optima.

We also establish practical applicability of four constraints for automatic clustering with applying various λ from 0.05-0.5. Figure 8 shows the numbers of error from four constraints with different λ.

From experiment results in Fig. 8 we can see that λ=0.3 is the ideal threshold to make automatic clustering. It happened in all constraints that λ=0.3 made the lowest error of automatic clustering. Table 2 performs the percentage of all error from 1000 experiments.

Table 2
The error percentage for automatic clustering

	Error (%)
Valley-tracing	43.18
Hill-climbing, α = 2	14.54
Hill-climbing, α = 3	9.48
Hill-climbing, α = 4	32.8

From Table 2 we can see that Hill-climbing is relatively more robust to make automatic clustering rather than Valley-tracing. The low φ for Valley-tracing, as shown in Table 1, made this constraint computed 43.18% of error to make automatic clustering. The highest φ reached by Hill-climbing with α=3 and made this constraint computed 9.48% of error for automatic clustering.

6. The improvement of the approach

In this paper we make the improvement of the proposed approach to accelerate the automatic clustering. It reverses the pattern of moving variance. First of all we change the agglomerative that we used for single linkage hierarchical algorithm into divisive in order to reverse the pattern of moving variance. Then, to apply the two constraints, valley-tracing and hill-climbing, in the

Ali Ridho Barakbah, Kohei Arai, "Reversed pattern of moving variance for accelerating automatic clustering", EEPIS journal, pp.15-21, Vol. 9, No. 2, December 2004, Surabaya, Indonesia.

reversed pattern we should redefine Eq. (6) and Eq. (7) in order to track continous points in the pattern.

Because of the reversion of the pattern, where the stage of *i*+1 will be *i*-1 in the pattern, we can modifies Eq. (6) as below:

$$v_{i-1} > \alpha \cdot v_i \tag{9}$$

It is also applied for Eq. (7) with modifying as:

$$(v_{i+1} \geq v_i) \cap (v_{i-1} > v_i) \tag{10}$$

We do not need to modify Eq. (8) because the result of reversion gives the equal equation.

To ensure the better performance of the reversion, we make the experimental applicability. We compute 1000 experiments using two constraints, valley-tracing and hill-climbing (α=2, α=3 and α=4), with λ=0.05-0.5. We record the performance of automatic clustering with normal moving average and with reversion.

Table 3 shows the comparison results of average numbers of stage to construct automatic clustering between normal and reversed moving average. We can see the better performance with shorter numbers of stage when we make the automatic clustering with reversion.

Table 3
Average numbers of stage between normal and reversed moving average

	Normal	Reversion
Valley-tracing	32.3555	5.1595
Hill-climbing, α=2	32.4388	5.0762
Hill-climbing, α=3	35.7502	4.7648
Hill-climbing, α=4	33.1537	4.3613

Table 4 performs the comparison results of average computation time of automatic clustering between normal and reversion.

Table 4
Average computation time between normal and reversed moving average

	Normal (ms)	Reversion (ms)
Valley-tracing	3943.178	682.564
Hill-climbing, α=2	3954.934	675.431
Hill-climbing, α=3	3982.173	644.258
Hill-climbing, α=4	4035.886	590.193

From Table 5 we can see the reduced time of computation with reversed pattern of moving average.

Table 5
The reduced time of computation with reversion

	Reduced time (%)
Valley-tracing	82.69
Hill-climbing, α=2	83.04
Hill-climbing, α=3	83.82
Hill-climbing, α=4	85.38

For all constraints those we used, the reversion can reduce 83.73% of computation time compared with normal pattern of moving average.

7. Conclusion

The proposed approach can solve the clustering problem and create well-separated clusters. From the experimental results with some various random normal data distribution clustering cases, Valley-tracing is better to find the global optimum as well as avoid the local optima, because it considers the next and the previous stage from current stage of cluster construction. However, it caused low gap distance (φ) between global optimum and candidate global optimum. It evokes the difficulty to determine the appropriate threshold in order to make automatic clustering. Among the four constraints, Hill-climbing with α=3 is relatively better used to make automatic clustering. From the experiments, the appropriate threshold is converged around λ=0.3 for all constraints. The new approach in this paper to accelerate the automatic clustering by reversing the pattern of moving variance performs the faster construction time of automatic clustering. From experimental results it can reduce 83.73% of computation time for constructing automatic clusters.

References

[1] C.A. Murthy, N. Chowdhury, *In search of optimal clusters using genetic algorithms*, Pattern Recognition Lett. 17 (1996), 825-832.

[2] G. Karypis, E.H. Han, V. Kumar, *Chameleon: a hierarchical clustering algorithm using dynamic modeling*, IEEE Computer: Special Issue on Data Analysis and Mining 32(8):68W5, 1999.

[3] G.A. Growe, *Comparing algorithms and clustering data: components of the data mining process,* thesis, Department of Computer Science and Information Systems, Grand Valley State University, 1999.

[4] S. Ray, R.H. Turi, *Determination of number of clusters in k-means clustering and application in colour image segmentation*, 4th ICAPRDT Proc., pp.137-143, 1999.

Ali Ridho Barakbah, Kohei Arai, "Reversed pattern of moving variance for accelerating automatic clustering", EEPIS journal, pp.15-21, Vol. 9, No. 2, December 2004, Surabaya, Indonesia.

[5] L.Y. Tseng, S.B. Yang, *A genetic approach to the automatic clustering problem*, Pattern Recognition Lett. 34 (2001), 415-424.

[6] M. Halkidi, Y. Batistakis, M. Vazirgiannis, *Clustering algorithms and validity measures*, In: Proc. the 13th International Conference on Scientific and Statistical Database Management, IEEE Computer Society, George Mason University, 2001.

[7] C.J. Veenman, M.J.T. Reinders, E. Backer, *A maximum variance cluster algorithm*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 9, pp. 1273-1280, 2002.

[8] E.V. Castro, *Why so many clustering algorithms-a position paper*, ACM SIGKDD Explorations Newsletter, Volume 4, Issue 1, pp. 65-75, 2002.

[9] A.R. Barakbah, K. Arai, *Identifying moving variance to make automatic clustering for normal data set*, In. Proc. IECI Japan Workshop 2004 (IJW 2004), Musashi Institute of Technology, Tokyo.

[10] D. Frossyniotis, A. Likas, A. Stafylopatis, *A clustering method based on boosting*, Pattern Recognition Lett. (2004) (accepted).

[11] P.A. Vijaya, M.N. Murty, D.K. Subramanian, *Leaders–subleaders: an efficient hierarchical clustering algorithm for large data sets*, Pattern Recognition Lett. 25 (2004), 505–513.

[12] W.H. Ming, C.J. Hou, *Cluster analysis and visualization*, Workshop on Statistics and Machine Learning, Institute of Statistical Science, Academia Sinica, 2004.

[13] A.R. Barakbah, K. Arai, *Determining constraints of moving variance to find global optimum and make automatic clustering*, In. Proc. Industrial Electronics Seminar (IES) 2004, Electronics Engineering Polytechnic Institute of Surabaya (EEPIS) – ITS, Surabaya.