Rep. Fac. Sci. Engrg.
Saga Univ.
36-1 (2007),9-15

Reports of the Faculty of Science and Engineering,
Saga University, Vol. 36, No.1, 2007

Cluster construction method based on global optimum cluster determination with the newly defined moving variance

By
Kohei ARAI* and Ali Ridho Barakbah**

Abstract: Generally, non-hierarchical clustering is easy an algorithm and is used abundantly from a by no means bad cluster result being obtained comparatively quickly. However, it is not rare to lapse into the class result (for it to be equivalent to the partial solution in an optimization problem) which depended for the cluster result in this clustering method on the set-up initial cluster center of gravity, and was mistaken depending on this setup. Even if it mistakes a setup of the initial cluster center of gravity, the proposed method of finding out the global optimal solution (the optimal cluster of which the input data should be belonged to) is proposed. That is, moving variance is defined as the within cluster variance at the time of determining an imputed cluster, and it judges having reached the global optimal solution based on this tendency. By the experiment using the widely available test data, by comparing the cluster results of this proposal technique and the existing non-hierarchical clustering technique showed the predominancy of the proposal technique
Key words: Single Linkage Hierarchical clustering, cluster density, global optimum, automatic clustering

1. Introduction

Clustering is an exploratory data analysis tool that deals with the task of grouping objects that are similar to each other [1, 4, 10]. 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) [2, 6]. 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" [6]. One of the methods to define a good cluster is variance constraint [5] that calculates the cluster density with variance within cluster (V_w) and variance between clusters (V_b) [3, 10]. The ideal cluster has minimum V_w to express internal homogeneity and maximum V_b to express external homogeneity.

The following section describes the proposed clustering method together with the existing typical clustering methods which are referred to the proposing the method. The proposed method is compared to the conventional methods with simulation and widely available data for clustering performance evaluation.

2. Proposed method

2.1 Single linkage clustering method

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 [4, 7, 9]. Agglomerative method proceeds by series of fusions of the "n" similar objects into groups, and divisive method, which separate "n" objects 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

Received on Apr.28 2007
*Department of Information Science
**EEPIS:Electric Engineering Politechnique in Surabaya

[1]. Therefore, it is called Single Linkage Clustering Method (SLHM). Euclidian distance is commonly used to calculate the distance in case of numerical data sets [9]. 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 reached an optimum

2.2 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 [3, 10].

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 δ_p is the center of gravity of cluster p, than variance of cluster p (δ_p^2) can be calculated as:

$$\delta_p^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y}_p)^2 \quad (2)$$

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

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

Then, variance between clusters (δ_b^2) quantifies the variability of the group mean around the grand mean (y), and hence the component of group differences. This is defined as:

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

Because an ideal cluster has minimum δ_w^2 and maximum δ_b^2, so based on this statement, it means the ideal cluster has minimum P where:

$$P = \frac{\delta_w^2}{\delta_b^2} \times 100 \quad (5)$$

2.3 The difficulty to find the global optimum

However minimum P expresses the ideal cluster, we can not apply directly to find the global optimum. There is some experiments prove that in some cases, minimum P reaches the local optima of cluster construction. For example, in case of Fig. 1, minimum P=0.15 resides in stage 1 with 49 total cluster. Stage 2 performs P=0.18 with 44 total cluster. But, actually the ideal cluster resides in stage 15 with 6 total cluster where P=0.22. Therefore, minimum P can not be used directly to find the global optimum. If we force to apply minimum P directly to identify the global optimum, in some cases, it may fall in local optima. To solve this problem, this paper proposed the new formulation to find the global optimum and avoid the local optima.

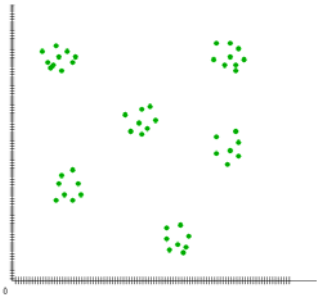

Fig. 1. A two dimensional case of clustering problem with n=50

2.4 Cluster construction

SLHM is very thorough to make analysis every states of cluster construction stage by stage. Therefore, this paper used SLHM as appropriate method in order to identify the moving variance from each stage of cluster construction. Fig. 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 3 total clusters.

2.5 Identifying pattern of moving variance

For finding the global optimum of cluster construction and avoid the local optima, we propose new formulation 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 the valley of patterns. Table 1 performs the possible patterns to get the global optimum. From analyzing the pattern in the Table 1, we can describe that the possibility to find the global optimum resides in stage fulfilled:

$$P_{i-1} >= P_i \quad \text{and} \quad P_{i+1} > P_i \quad (6)$$

for i=1..n, and n is latest stages of cluster construction. Then, we identify the different value of altitude ∂ for each stage, as figured at Fig. 3, it can be defined:

$$\partial = (P_{i+1} - P_i) + (P_{i-1} - P_i) = (P_{i+1} + P_{i-1}) - (2 \times P_i) \quad (7)$$

In order to avoid the local optima and find the global optimum, it can be derived from maximum of ∂ that fulfilled Eq.(7).

To construct cluster automatically, we put the additional variable λ as a threshold value to get a maximum of ∂. 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.

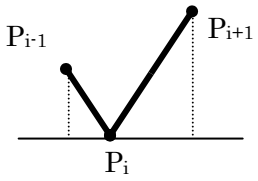

Fig. 3. Different value of altitude

3. Accuracy of the proposed method

3.1 Point datasets

We examined our proposed method to some of different cases for the shape independent clustering. It covered determining the cluster density as well as the global optimum. Various cases those examined can determine the accuracy of the proposed method. For every case, we record the valuable data that has values moving pattern at each stage. In our experimental cases, we use λ=0.1 to reach the global optimum.

We also use an additional variable φ to express the different values between max(∂) and ∂_i that has closer value to max(∂).

$$\varphi = \frac{\max(\partial)}{\text{closer value to } \max(\partial)} \quad (8)$$

The value of φ can show a distant value to get global optimum. The large φ, at least $\varphi \geq 2$, expresses possibility to construct well-separated cluster. It avoids cluster construction reaching any local optima. If the closer value is non-positive, the value of φ will be ϕ, means the global optimum is absolutely right.

Fig.5 shows how the proposed method works for avoidance of the local minima in comparison to the existing case as is shown in Fig4.

It is found that our proposed method is superior to solve the clustering case. The result shows the accuracy of ∂ to express the global optimum, as viewed in Fig. 6. We use λ=0.1 to reach the global optimum. The experimental result showed that the maximum of ∂ = 1.39, in the stage which numbers of well-separated cluster is 3. It is proved that the global optimum will be reached with 3 numbers of clusters. The value φ = 19.8571. We applied the proposed method to solve some various clustering cases (Fig.7-13). The result of clustering construction is indicated with a different color.

3.2 Real world datasets

The real world datasets used are Iris data, Wine data, Fossil data, Ruspini data, Letter Recognition data and New Thyroid data which are widely used and well known datasets for evaluation of clustering algorithms.

The raw data of the real world datasets are used because comparison of clustering performance between the proposed method and the other existing method (random designation of initial cluster center) is concerned. Clustering performance of Single Linkage, K-means clustering with random designated initial cluster center is compared to the proposed method.

The following error percentage which is calculated from the number of misclassified patterns and the total number of patterns in the datasets is evaluated.

$$Error = \frac{Numberofmisclassified}{Numberofpatterns} x100\%$$

(1) Iris dataset

This dataset is from the UCI Repository[11]. This dataset contains information about Iris flowers. There are three classes of Iris flowers, namely Iris Setosa, Iris Versicolor and Iris Virginica. The dataset consists of 150 examples with 4 attributes. One class is well separable against the other two. The others have a large overlap.

(2)Wine dataset

We also obtained this dataset from UCI Repository. The data is the result of a chemical analysis of wines grown in a region in Italy but derived from three different cultivars. There are three classes. The dataset consists of 178 examples each with 13 continuous attributes. The dataset contains distribution 59 examples of class 1, 71 examples for class 2 and 48 examples for class 3.

(3) Fossil dataset

The Fossil data is obtained from Chernoff[12]. It consists of 87 nummulitidae specimens from Eocene yellow limestone formation of northwestern Jamaica. There are three 6 attributes with 3 classes which the distribution is 40 examples of class 1, 34 examples of class 2 and 13 examples of class 3.

(4) Ruspini dataset

The Ruspini dataset represents a simple, well-known example that is commonly used as a benchmark problem in evaluating clustering methods and is widely available, incorporated as a built-in data object in both R and

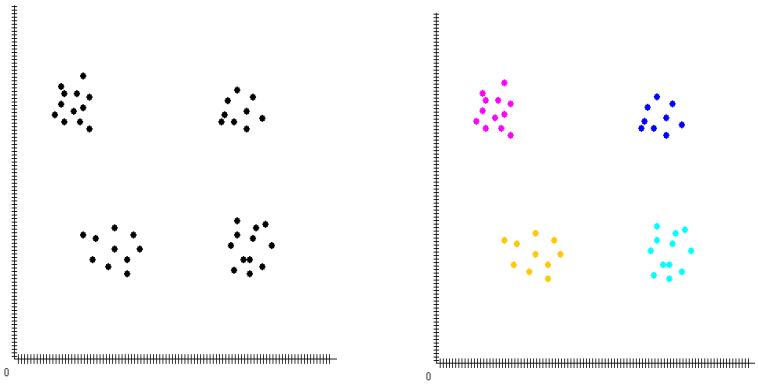

Fig. 7. 4 data set, n=43, λ=0.1, max(∂) = 1.33, φ = 19.

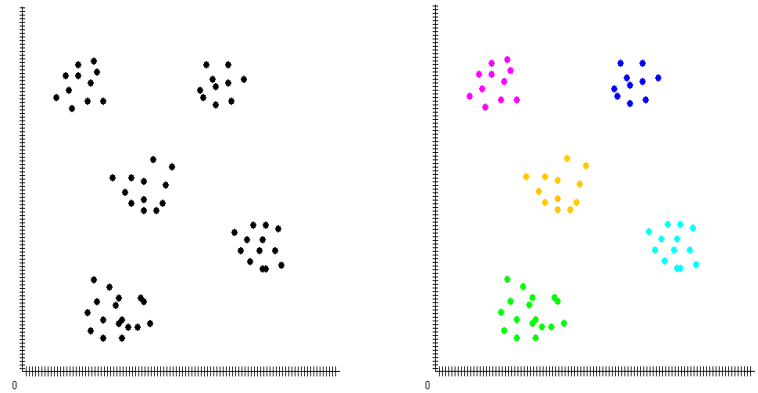

Fig. 8. 5 data set, n=63, λ=0.1, max(∂) = 0.6, φ = 20.

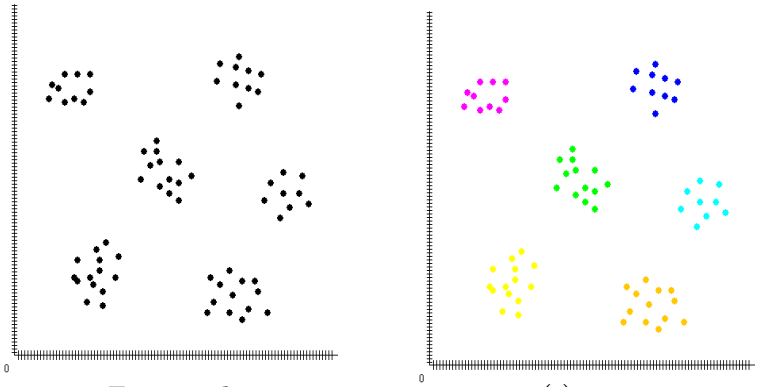

Fig. 9. 6 data set, n=69, λ=0.1, max(∂) = 0.52, φ = 13.

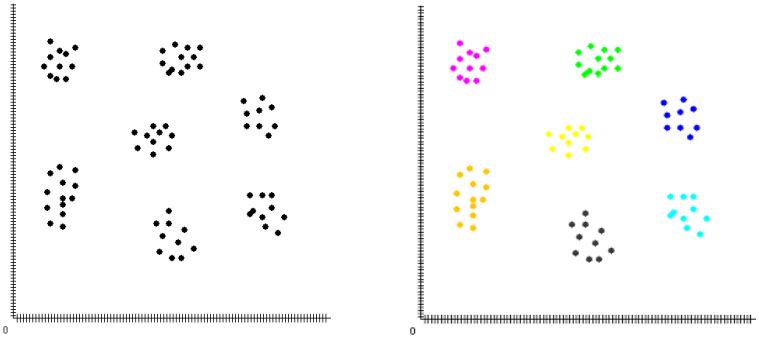

Fig. 10. 7 data set, n=75, λ=0.1, max(∂) = 0.43, φ = 5.375.

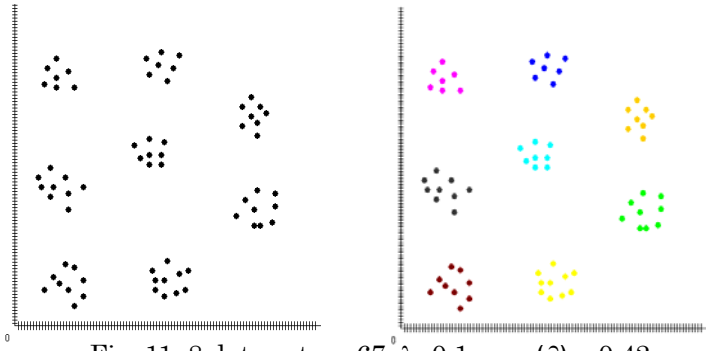

Fig. 11. 8 data set, n=67, λ=0.1, max(∂) = 0.42, φ = 8.4.

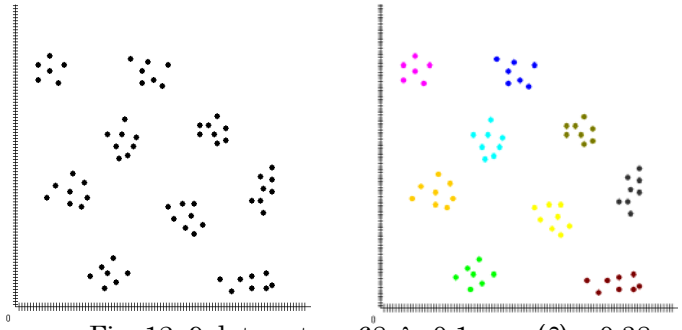

Fig. 12. 9 data set, n=68, λ=0.1, max(∂) = 0.38, φ = 9.5.

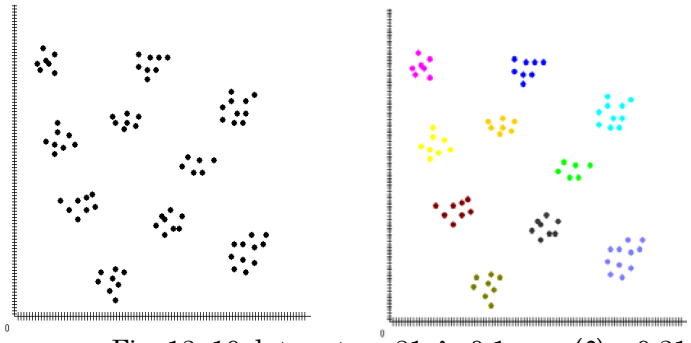

Fig. 13. 10 data set, n=81, λ=0.1, max(∂) = 0.21, φ = 5.25.

S-plus statistics packages [13]. The dataset consists of 75 bi-variate attribute vectors. There are five classes. The dataset contains 23, 20, 17 and 15 in classes 1, 2, 3 and 4, respectively.

(5) Letter recognition dataset

This dataset obtained from UCI Repository. The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts), which were then scaled to fit into a range of integer values from 0 through 15. The training data consists of first 16000 items and then used the resulting model to predict the letter category for the remaining 4000. For experimental purpose we have taken 595 patterns of letter A and 597 patterns of letter D from the training dataset.

(6) New thyroid dataset

The new thyroid dataset is also obtained from UCI Repository. The dataset contains information about classification whether a patient's thyroid to the class euthyroidism, hypothyroidism or hyperthyroidism. The diagnosis (the class label) was based on a complete medical record, including anamnesis, scan etc. The dataset consists 5 attributes, with 215 examples. The distribution is 150 of class euthyroidism, 35 of class hypothyroidism and 30 of class hyperthyroidism.

As is shown in Table 2-7, the clustering error of the proposed method is smaller than those of the single linkage clustering and k-means clustering with randomly designated initial cluster center. In terms of the simplicity of the dataset, Ruspini is the simplest followed by Fossil, Letter, Iris, New thyroid and Wine datasets.

This order is almost identical to the clustering error of the proposed method and Single linkage clustering. Turns out, the error if the Letter dataset is smallest followed by Fossil, Ruspini, Iris, New thyroid and Wine datasets for k-means clustering with randomly designated initial cluster center.

The averaged clustering error of the proposed method is 11.22% and is 34.28% smaller than that of k-means clustering with randomly designated initial cluster center (17.13%) and also is around 60% smaller than that of single linkage clustering (30.5%).

Table 2 Iris dataset

	Error (%)
Single Linkage	32.000
K-means using random initialization	17.7507
Proposed clustering method	10.6667

Table 3 Wine dataset

	Error (%)
Single Linkage	57.3034
K-means using random initialization	32.6197
Proposed clustering method	29.7753

Table 4 Fossil dataset

	Error (%)
Single Linkage	13.7931
K-means using random initialization	8.5931
Proposed clustering method	4.5977

Table 5 Ruspini dataset

	Error (%)
Single Linkage	0
K-means using random initialization	13.7787
Proposed clustering method	0

Table 6 Letter recognition dataset

	Error (%)
Single Linkage	49.8322
K-means using random initialization	8.2326
Proposed clustering method	8.2215

Table 7 New thyroid dataset

	Error (%)
Single Linkage	29.7674
K-means using random initialization	20.9842
Proposed clustering method	13.9535

4. Conclusions

From the experimental results with some various clustering cases, the proposed method can solve the clustering problem and create well-separated clusters. The variable φ showed in those cases that the possibility of constructing well-separated clusters is high, implies that the proposed method can also avoid any local optima and find the global optimum. The threshold of λ is easy to set ensuring reach the global optimum. For more the amorphous shape independent cases need smaller λ to set as more precise as possible. By setting the value, λ, the well-separated cluster will be constructed. The very high value of φ for normal data sets proves that the proposed method is able to solve the clustering problems.

It is found that the proposed method achieved 34.28% and 60% improvement in terms of clustering error in comparison to the k-means clustering with randomly designated initial cluster center and the single linkage clustering method, respectively.

References

[1]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.

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

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

[4]M. Halkidi, Y. Batistakis, M. Vazirgiannis, *Clustering algorithms and validity measures*, proceedings of the 13th International Conference on Scientific and Statistical Database Management, July 18–20. IEEE Computer Society, George Mason University, Fairfax, Virginia, USA, 2001.

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

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

2002.

[7]D. Frossyniotis, A. Likas and A. Stafylopatis, *A clustering method based on boosting*, Pattern Recognition Letters, 25, 6, 641-654, 2004..

[8]S. Bandyopadhyay, *An automatic shape independent clustering technique*, Machine Intelligence Unit, Journal of Pattern Recognition Society, volume 37, number 1, 2004.

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

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

[11] UCI Repository (http://www.sgi.com/tech/mlc/db/)

[12]Yi-tsuu, C., Interactive Pattern Recognition. Marcel Dekker Inc., New York and Basel.、1978.

[13] Pearson, R. K., Zylkin, T., Schwaber, J.S., Gonye, G.E., Quantitative evaluation of clustering results using computational negative controls. Proc. 2004 SIAM International Conference on Data Mining, Lake Buena Vista, Florida, 2004..