Ali Ridho Barakbah, Kohei Arai, "Identifying moving variance to make automatic clustering for normal data set", IECI Japan Workshop 2004 (IJW 2004), May 22, 2004, Musashi Institute of Technology, Tokyo.

Proceedings of the IECI Japan Workshop 2004
Musashi Institute of Technology, Japan

ISSN 1344-7491
pp. 26 - 30

Identifying moving variance to make automatic clustering for normal data set

Ali Ridho Barakbah, Kohei Arai *
Department of Information Science, Saga University, Japan

Abstract

This paper proposed new approach to make cluster construction automatically for normal data set. The proposed method for automatic cluster construction is based on identifying moving variance of cluster for each stage of cluster construction, then analyzing the pattern to find the global optimum. After that, this paper proposed a new formulation to stop the construction of the cluster where it is in the global optimum, as well as to avoid the local optima. Experiment results will perform the effectiveness of the proposed method in this paper.

Keywords: Single Linkage Hierarchical Method, 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.

2. SINGLE LINKAGE HIERARCHICAL 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 [1]. Therefore, it is called Single Linkage Hierarchical Method (SLHM). Euclidian distance is commonly used to calculate the distance in case of numerical data sets [9]. For two dimensional data set, it performed as:

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

The algorithm of Single Linkage clustering method is composed of the following steps:

* Dept. of Information Science, Saga University, 1 Honjo, Saga 840-8502 Japan; tel.: +81-952-28-8567; fax: +81-952-28-8650; e-mail: ridho@ip.is.saga-u.ac.jp

Ali Ridho Barakbah, Kohei Arai, "Identifying moving variance to make automatic clustering for normal data set", IECI Japan Workshop 2004 (IJW 2004), May 22, 2004, Musashi Institute of Technology, Tokyo.

Proceedings of the IECI Japan Workshop 2004
Musashi Institute of Technology, Japan

ISSN 1344-7491
pp. 26 - 30

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≠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

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 [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 $\overline{y}_p$ is the center of gravity of cluster p, than variance of cluster p (δ_p^2) can be calculated as:

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

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 ($\overline{y}$) and hence the component of group differences. This is defined as:

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

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:

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

4. 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 are some experiments proved 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 clusters. Stage 2 performs P=0.18 with 44 total clusters. But, actually the ideal cluster resides in stage 15 with 6 total clusters where P=0.22.

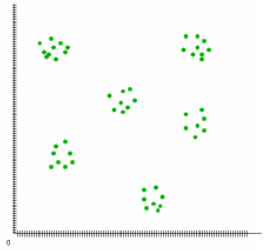


Fig. 1. A case of clustering with n=50

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.

5. 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.

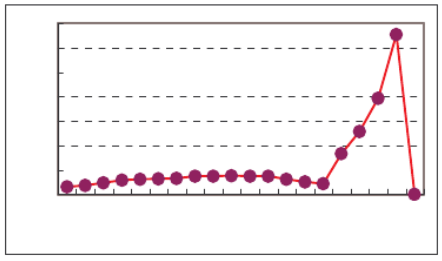


Fig. 2. Moving variance of cluster construction for each stage

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

Ali Ridho Barakbah, Kohei Arai, "Identifying moving variance to make automatic clustering for normal data set", IECI Japan Workshop 2004 (IJW 2004), May 22, 2004, Musashi Institute of Technology, Tokyo.

Proceedings of the IECI Japan Workshop 2004
Musashi Institute of Technology, Japan

ISSN 1344-7491
pp. 26 - 30

6. IDENTIFYING PATTERN OF MOVING VARIANCE

For finding the global optimum of cluster construction and avoid the local optima, we propose a new formulation the 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 possibility of the patterns to get the global optimum.

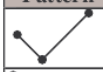
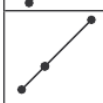
Pattern	Possible?	Pattern	Possible?
	√		X
	√		X
	√		X
	X		√
	X		X
	X		X
	X		X
	X		X

Table 1. Possibility of patterns to be a global optimum

From analyzing the pattern in the Table 1, we can describe that the possibility to find the global optimum resides in stage which fulfilled:

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

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

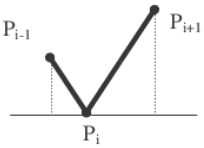


Fig. 3. Different value of altitude

Then, we identify the different value of altitude ∂ for each stage, as figured at Fig. 3, which can be defined:

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

In order to avoid the local optima and find the global optimum, it can be derived from maximum of ∂ that fulfilled Eq. 6.

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.

7. ACCURACY OF THE PROPOSED METHOD

We examined our proposed method to some of different cases for the normal data set 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(∂).

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

The value of φ can show a distant value to get global optimum. The large φ, at least φ≥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.

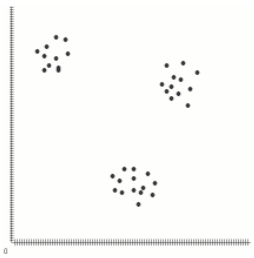


Fig. 3. The case of clustering with n=37

Ali Ridho Barakbah, Kohei Arai, "Identifying moving variance to make automatic clustering for normal data set",
IECI Japan Workshop 2004 (IJW 2004), May 22, 2004, Musashi Institute of Technology, Tokyo.

Proceedings of the IECI Japan Workshop 2004
Musashi Institute of Technology, Japan

ISSN 1344-7491
pp. 26 - 30

Fig. 4 shows how the proposed method works for avoidance of the local minima in comparison to the existing case as is shown in Fig. 3.

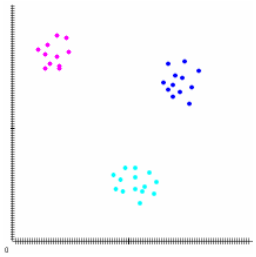


Fig. 4. The result of the case

It is found that our proposed method is able to solve the clustering case. The result shows the accuracy of ∂ to express the global optimum, as viewed in Fig. 5. 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 total numbers of cluster, with the value φ = 19.8571.

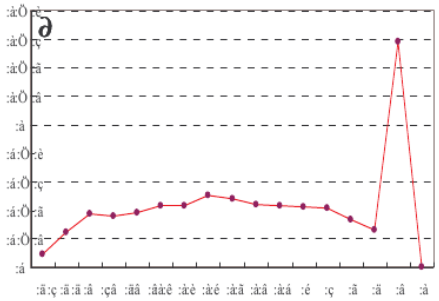


Fig. 5. The moving values of ∂ at each stage

We applied the proposed method to solve some various clustering cases (Fig.6 - Fig.12). The result of clustering construction is indicated with a different color.

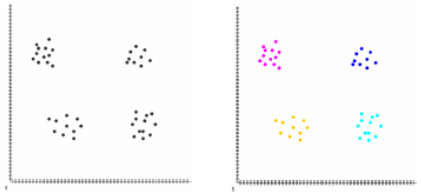


Fig. 6. 4 data set, n=43, λ=0.1, max(∂) = 1.33, φ = 19.

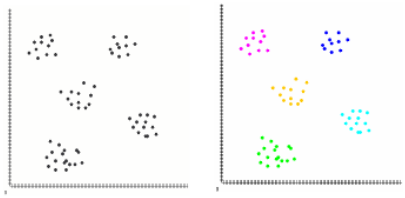


Fig. 7. 5 data set, n=63, λ=0.1, max(∂) = 0.6, φ = 20.

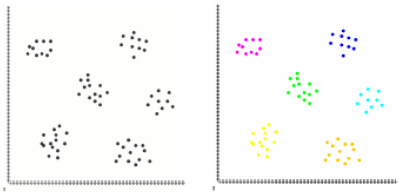


Fig. 8. 6 data set, n=o›, λ=0.1, max(∂) = 0.52, φ = 13.

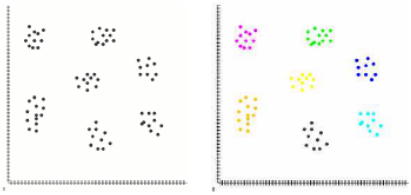


Fig. 9. 7 data set, n=75, λ=0.1, max(∂) = 0.43, φ = 5.375.

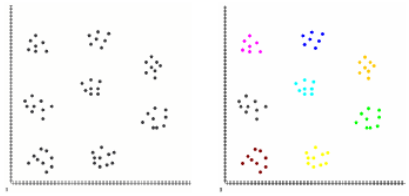


Fig. 10. 8 data set, n=67, λ=0.1, max(∂) = 0.42, φ = 8.4.

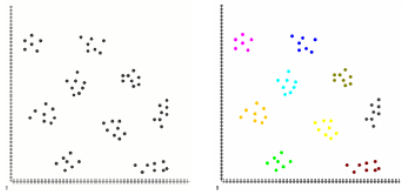


Fig. 11. 9 data set, n=68, λ=0.1, max(∂) = 0.38, φ = 9.5.

Ali Ridho Barakbah, Kohei Arai, "Identifying moving variance to make automatic clustering for normal data set",
IECI Japan Workshop 2004 (IJW 2004), May 22, 2004, Musashi Institute of Technology, Tokyo.

Proceedings of the IECI Japan Workshop 2004
Musashi Institute of Technology, Japan

ISSN 1344-7491
pp. 26 - 30

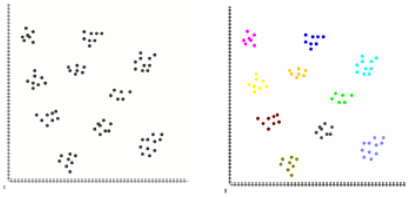


Fig. 12. 10 data set, n=81, λ=0.1, max(∂) = 0.21, φ = 5.25.

8. CONCLUSION

From the experimental results with some various clustering cases for normal data set, 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 complex clustering 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 of value φ proved that the proposed method is able to solve the clustering cases for normal data set.

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, *Comparing 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, *Determination 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 (2004) (Accepted).

[8] S. Bandyopadhyay, *An automatic shape independent clustering technique*, Machine Intelligence Unit, Journal of Pattern Recognition Society, volume 37, number 1, January 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 (2004) 505–513.

[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.