Measures for the structure of clustering and admissibilities of its algorithm

Akinobu Takeuchi*, Hiroshi Yadohisa**, and Koichi Inada**
* College of Social Relations, Rikkyo (St. Paul's) University,
Nishi-Ikebukuro 3-34-1, Tokyo 171-8501, JAPAN,
E-mail: akitake@rikkyo.ac.jp
* Department of Mathematics and Computer Science, Kagoshima University,
Korimoto 1-21-35, Kagoshima 890-0065, JAPAN,
E-mail: (yado, inada)@sci.kagoshima-u.ac.jp

Abstract

The problem of selecting a clustering algorithm from the myriad of algorithms has been discussed in recent years. Many researchers have attacked this problem by using the concept of admissibility (e.g. Fisher and Van Ness, 1971, Yadohisa, et al., 1999). We propose a new criterion called the "structured ratio" for measuring the clustering results. It includes the concept of the well-structured admissibility as a special case, and represents some kind of "goodness-of-fit" of the clustering result. New admissibilities of the clustering algorithm and a new agglomerative hierarchical clustering algorithm are also provided by using the structured ratio. Details of the admissibilities of the eight popular algorithms are discussed.

Keywords and phrases: admissibility, AHCA (agglomerative hierarchical clustering algorithm), structure

1 Introduction

Several criteria for measuring the results of clustering algorithms have been proposed. Examples are the cophenetic correlation coefficient (Sokal and Rohlf, 1962), sum of squares (Hartigan, 1967), and Minkowski metrics (Jardine and Sibson, 1971). Takeuchi, et al. (1999) proposed the distortion ratio based on the concept of space distortion introduced by Lance and Williams (1967).

The well-structured criterion proposed by Rubin (1967) is another measure and is based on the dispersion of clusters. He defined data as well-structured (l-group) if there exist clusters $C_1, C_2, \ldots, C_l$ such that all within-cluster distances are smaller than the smallest between-cluster distance. Using this concept, Fisher and Van Ness (1971) proposed a new clustering algorithm admissibility called the well-structured admissible.

In this paper, we propose a new criterion for measuring clustering results called the "structured ratio". It includes the well-structured concept as a special case, and represents some kind of goodness-of-fit of a clustering result. New admissibilities and a new agglomerative hierarchical clustering algorithm (AHCA) are also provided by using the structured ratio, and details of the admissibilities of the eight popular algorithms are discussed.

Cluster I at stage m ($1 \leq m < N$) is denoted as $C_I(m)$. We denote the dissimilarity between objects p and q by d_{pq}, the dissimilarity between $C_I(m)$ and $C_J(m)$ by d_{IJ}, and the number of objects to be clustered by N. We use the standard set theoretic notation $p \in C_I(m)$ to indicate that object p belongs to $C_I(m)$; the number of objects belonging to $C_I(m)$ is denoted by n_I. To simplify notation, we define ${}_{n_I}C_2 = 0$ if $n_I = 1$.

We assume that clusters $C_I(m)$ are obtained using some AHCAs. From this assumption, the number of the clusters at stage m is $N - m$. When $C_T(m)$ and $C_K(m)$ are combined at stage m and $C_T(m)$ is not a singleton, it is assumed that $C_T(m)$ was formed from $C_I(t)$ and $C_J(t)$, which were combined at stage t ($1 < t < m$), and that $C_K(m)$ is a singleton or was formed from $C_{I'}(t')$ and $C_{J'}(t')$, which were combined at stage t' ($1 \leq t' < t$). Hereafter, we assume this relationship between the two combined clusters, without loss of generality, and we assume $d_{IJ} < d_{IK} \leq d_{JK}$.

We abbreviate the single linkage algorithm as SL, the complete linkage algorithm as CL, the weighted average algorithm (WPGMA) as WA, the median algorithm (WPGMC) as MD, the group average algorithm (UPGMA) as GA, the centroid algorithm (UPGMC) as CE, the minimum variance algorithm (Ward's method) as WD, and the flexible algorithm with $\beta = -0.25$ (see Gordon, 1996) as FX.

2 Structured measures

Here we define the "structured ratio" as an extension of the well-structured concept that was first proposed by Rubin (1967). We define W_h as the dispersion within a cluster and B_h as the dispersion between clusters.

Definition 1: The structured ratio at stage m ($< N-1$) is defined as:

$$SR_h(m) = W_h(m)/B_h(m), \tag{1}$$

where $W_h(m)$ and $B_h(m)$ are within cluster and between cluster dispersions at stage m, respectively. We define several measures of within cluster and between cluster dispersion. For example, for $I \neq J$, which we assume hereafter,

$$W_1(m) = \max_I \max_{p,q \in C_I(m)} d_{pq}, B_1(m) = \min_{I,J} \min_{p \in C_I(m), q \in C_J(m)} d_{pq},$$

$$W_2(m) = \sum_I \left(\max_{p,q \in C_I(m)} d_{pq} \right) \Big/ (N-m),$$

$$B_2(m) = \sum_{I,J} \left(\min_{p \in C_I(m), q \in C_J(m)} d_{pq} \right) \Big/ {}_{N-m}C_2,$$

$$W_3(m) = \max_I \left(\sum_{p,q \in C_I(m)} d_{pq}/{}_{n_I}C_2 \right), B_3(m) = \min_{I,J} \left(\sum_{p \in C_I(m), q \in C_J(m)} d_{pq}/n_I n_J \right),$$

$$W_4(m) = \sum_I \left(\sum_{p,q \in C_I(m)} d_{pq}/{}_{n_I}C_2 \right) \Big/ (N-m),$$

$$B_4(m) = \sum_{I,J} \left(\sum_{p \in C_I(m), q \in C_J(m)} d_{pq}/n_I n_J \right) \Big/ {}_{N-m}C_2,$$

$$W_5(m) = \sum_I \left(\sum_{p,q \in C_I(m)} d_{pq} \right) \Big/ \sum_I {}_{n_I}C_2,$$

$$B_5(m) = \sum_{I,J} \left(\sum_{p \in C_I(m), q \in C_J(m)} d_{pq} \right) \Big/ \sum_{I,J} n_I n_J.$$

Using the same dispersion measures, we define another ratio for representing the structure of clustering results, while the structured ratio is defined for each combination.

Definition 2: The total structured ratio is defined as:

$$TSR_h(N-L) = \sum_{m=1}^{N-L} SR_h(m)/(N-L), \tag{2}$$

where L ($1 < L < N$) is the number of clusters selected.

The total structured ratio can be used to measure the structure of clustering algorithms. If we would like to measure the final results of a clustering algorithm, $SR_h(N-L)$ may be more appropriate than the total structured ratio.

Since the structured ratio is the ratio of dispersion within a cluster to dispersion between clusters, a smaller value is preferable in terms of the concept of structure. However, the value of the structured ratio depends heavily on the dispersion measure. The characterization of dispersion measures still remains to be completed. However, we can obtain some useful information from the structured ratio using the following properties.

Property 1: $W_1(m)$ and $B_1(m)$ are monotone increasing functions.

Property 2: If $B_1(m+1)/B_1(m) \le W_1(m+1)/W_1(m)$ for all m, then SR_1 is a monotone increasing function.

Property 3: For any m $(< N-1)$, the following inequalities hold;

$$SR_4(m) \le SR_2(m) \le SR_1(m), \quad SR_4(m) \le SR_3(m) \le SR_1(m).$$

Property 4: If

$$W_3 \le \min_I \max_{p,q \in C_I(m)} d_{pq} \quad \text{and} \quad \max_{I,J} \left(\min_{\substack{p \in C_I(m) \\ q \in C_J(m)}} d_{pq} \right) \le B_3$$

hold, then the following inequalities hold.

$$SR_4(m) \le SR_3(m) \le SR_2(m) \le SR_1(m).$$

Property 5: The following equation hold for all m $(< N-1)$;

$$\sum_I {}_{n_I}C_2 W_5(m) + \sum_{I,J} n_I n_J B_5(m) = \sum_{p,q} d_{pq}.$$

3 ζ-structured admissibility

In this section, we propose some admissibilities of the clustering by using the structured ratio and the total structured ratio defined in previous section. Using the structured ratio, we redefine the condition of the well-structured (L-group) admissible first proposed by Fisher and Van Ness (1971), as follows.

An algorithm is well-structured (L-group) admissible if and only if the following equation is satisfied for any well-structured (L-group) data;

$$SR_1(N-m) < 1.$$

We defined an admissibility including this as a special case.

Definition 3: Suppose an algorithm classifies objects to L $(1 < L < N)$ clusters at stage m $(= N - L)$. If the following inequality is satisfied, the algorithm is ζ-structured (L-group) admissible;

$$SR_h(m) < \zeta. \tag{3}$$

The ζ-structured (L-group) admissible is defined at one combined stage. Next we define admissibilities for the entire set of combined stages.

Definition 4: Suppose an algorithm classifies objects to L $(1 < L < N)$ clusters at stage m $(= N - L)$. If the following inequality is satisfied for all n $(\le m)$, the algorithm is ζ-structured (perfect) admissible;

$$SR_h(n) < \zeta. \tag{4}$$

Similarly, we can define admissibility by using the total structured ratio.

Definition 5: Suppose an algorithm classifies objects to L $(1 < L < N)$ clusters. If the following inequality is satisfied, the algorithm is ζ-total structured admissible;

$$TSR_h(N-L) < \zeta. \tag{5}$$

As is obvious from these definitions, the concept of ζ-structured admissible is determined only at the stage when the data is separated into L clusters. This admissibility is a looser condition than the ζ-structured (perfect) admissible. In fact, for a small value of L, it is necessary to select quite a large value of ζ when the algorithm is ζ-structured (perfect) admissible.

These admissibilities satisfy the following properties.

Property 6: If an AHCA is ζ-structured (perfect) admissible, then the algorithm is ζ-structured (L-group) admissible and ζ-total structured admissible.

Property 7: If an AHCA is 1-structured (L-group) admissible, then the algorithm is well-structured (L-group) admissible, as proposed by Fisher and Van Ness (1971).

4 New algorithm

In this section, we defined a new AHCA which has the minimum $SR_h(m)$ at the stage m.

Definition 6: An AHCA, that combines $C_I(m-1)$ and $C_J(m-1)$ to make a new cluster at stage m by minimizing $SR_h(m)$ is called MSR_h algorithm.

5 A numerical example

Here, we analyze an artificial dataset in two-dimensional space (see, Figure 1). We anticipate that this data can be separated into three clusters.

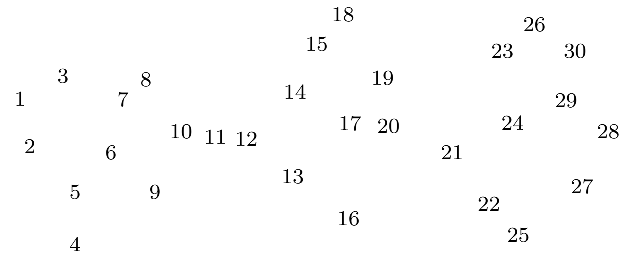

Figure 1: Scatter plots of 30 objects

Here, we analyze the data using eight popular AHCAs and the MSR_1 algorithm.

The SR_1 of the results of these algorithms are represented on the ordinate in Figures 2 and 3. The values of $SR_1(27)$ and $TSR_1(27)$ are shown in these figures and the abscissa shows the combined stage. We select the 27th stage because there are three clusters that combine at this stage.

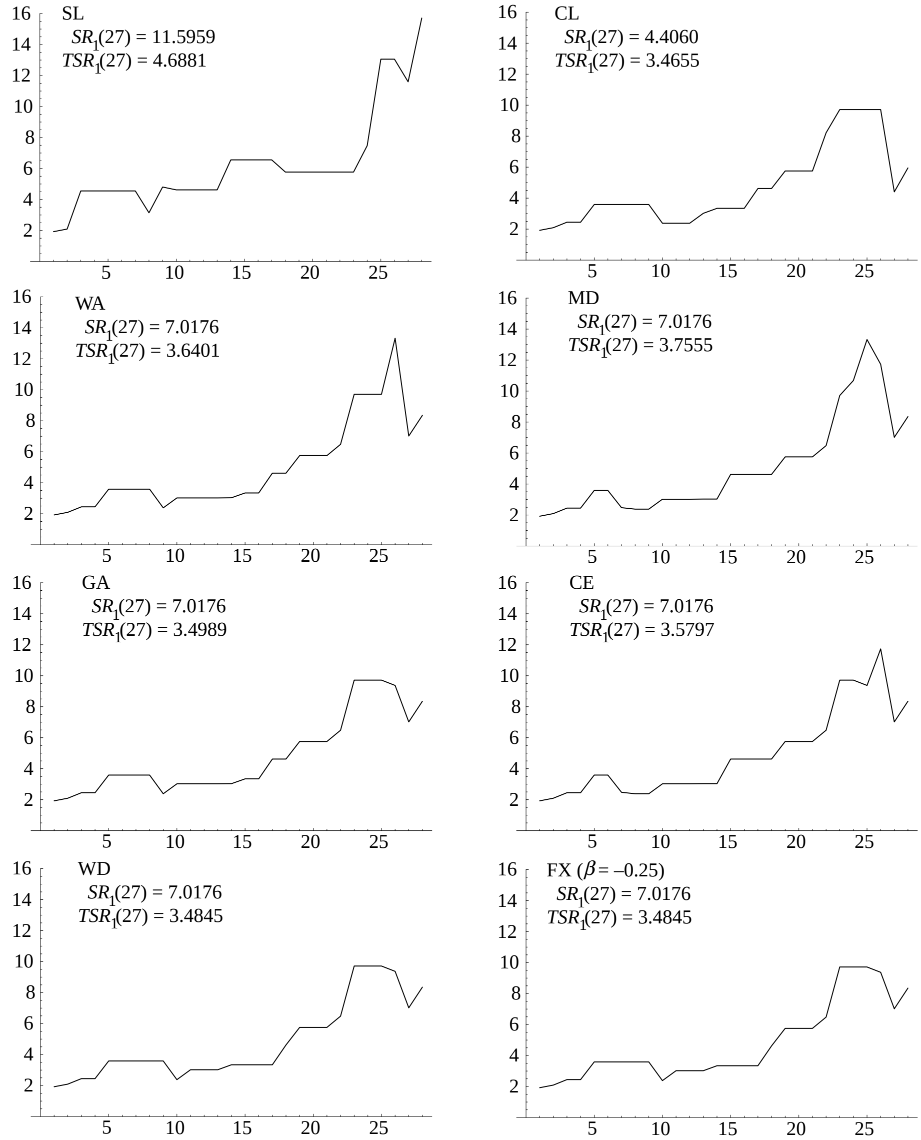

Figure 2: Structured ratios of 8 popular AHCAs

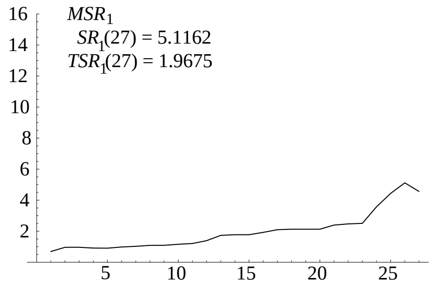

Figure 3: The structured ratio of MSR_1

The structured admissibilities for $\zeta = 5, 7.5, 10$ or $\zeta = 3.5, 4, 5$ are indicated in Table 1. Generally, from the definitions, the structured admissibilities are sensitive concepts in contrast to the total structured admissibilities. For example, most algorithms are not 5-structured (3-group or perfect) admissible, but they are 5-total structured admissible. By changing the value of ζ, we can control the condition of the structured admissibilities. For example, the structured (perfect) admissibilities are changed from 'No' to 'Yes' by decreasing the value of ζ at the assessment of CL, GA, WD, FX, and MSR_1 algorithms, respectively.

Table 1: ζ-structured admissibilities of the AHCAs

Admissible	SL	CL	WA	MD	GA	CE	WD	FX	MSR_1
5-structured (3-group)	No	Yes	No	No	No	No	No	No	No
7.5-structured (3-group)	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
10-structured (3-group)	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
5-structured (perfect)	No	No	No	No	No	No	No	No	No
7.5-structured (perfect)	No	No	No	No	No	No	No	No	Yes
10-structured (perfect)	No	Yes	No	No	Yes	No	Yes	Yes	Yes
3.5-total structured	No	No	No	No	No	No	No	No	Yes
4-total structured	No	Yes	No	No	Yes	No	Yes	Yes	Yes
5-total structured	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

6 Discussion

We consider the concept of well-structured, in which the desired classification condition has similar objects classified to the same cluster with small within-cluster dispersion, and dissimilar objects classified to different clusters with large between-cluster dispersion. However, the concept of well-structured is very strict, and is determined for specific data. The equation of the condition for a well-structured is only satisfied by a large L in data, and popular AHCAs are not satisfied with small values of L.

The structured concept using the structured ratio that we proposed can be used for any data and can control the condition of judgement. In addition, our concept can select from many criteria that is most suited for the user's purpose, not only the particular criterion. Thus, the concept of the structured includes the existing well-structured concept and can be used for more general and extensive cases. Additionally, this concept can numerically measure the degree of structure, so it can be used in a manner similar to admissibilities of an algorithm. By using this concept, we believe analysts can better select algorithms to obtain a desired result.

References

[1] Fisher, L. and Van Ness, J. (1971), Admissible clustering procedures, *Biometrika*, **58**, 91–104.

[2] Gordon, A. D. (1996), Hierarchical classification, In: *clustering and classification*, P. Arabie, L. Hubert and G. Soete (Eds.), World Scientific, New Jersey, 65–121.

[3] Hartigan, J. A. (1967), Representation of similarity matrices by trees, *Journal of the American Statistical Association*, **62**, 1140–1158.

[4] Jardine, N. and Sibson, R. (1971), *Mathematical taxonomy*, London, Wiley.

[5] Lance, G. N. and Williams, W. T. (1967), A general theory of classificatory sorting strategies: 1. hierarchical systems, *The Computer Journal*, **9**, 373–380.

[6] Rubin, J. (1967), Optimal classification into groups: an approach for solving the taxonomy problem, *Journal of Theoretical Biology*, **15**, 103–144.

[7] Sokal, R. R. and Rohlf, F. J. (1962), The comparison of dendrograms by objective methods, *Taxon*, **11**, 33–40.

[8] Yadohisa, H., Takeuchi, A. and Inada, I. (1999), Developing criteria for measuring space distortion in combinatorial cluster analysis and methods for controlling the distortion, *Journal of Classification*, **16**, 45–62.