A simulated study of measures for clustering

Akinobu Takeuchi
College of Social Relations,
Rikkyo University,
Tokyo, 171–0021, JAPAN
akitake@rikkyo.ac.jp

Hiroshi Yadohisa
Dept. of Math. & Comp. Sci.,
Kagoshima University,
Kagoshima, 890–0065, JAPAN
yado@sci.kagoshima-u.ac.jp

Koichi Inada
Dept. of Math. & Comp. Sci.,
Kagoshima University,
Kagoshima, 890–0065, JAPAN
inada@sci.kagoshima-u.ac.jp

1. Introduction

Several criteria for measuring the result of a clustering have been proposed. Example of this are the Cophenetic correlation coefficient (Sokal and Rohlf, 1962), the sum of squares (Hartigan, 1967), the Minkowski metrics (Jardine and Sibson, 1971) and so on. Yadohisa et al. (1999) proposed the distortion ratio based on the concept of the space distortion introduced by Lance and Williams (1967). Another criterion based on the dispersion of the clusters was proposed in Rubin (1967) and called well-structured. He defined the data as well-structured (l-group) if there exit a clustering $C_1, C_2, \ldots, C_l$ such that all within-clusters distances are smaller than all between-cluster distances. Using this concept, Fisher and Van Ness (1971) proposed a new admissibility of a clustering algorithm called well-structured admissible. Takeuchi, et al. (2000a) proposed a criterion for measuring the clustering result called "structured ratio". It includes the concept of well-structured as a special case, and represents some kind of goodness of the result of the clustering. However, the characterizations of these measures (e.g. the relationship between these measures and dispersion of objects) are still remained to be completed.

In this paper, we investigate the relationship between the structured ratio and the data analyzed . Before turning to a simulated study, the structured ratio will be introduced. In the simulation, we consider several factors which may have an effect on the ratio. The purpose of this paper is to characterize the ratio through the simulation.

2. Structured ratio

Here we introduce "Structured Ratio" (Takeuchi, et al., 2000a) based on the concept of Rubin's well-structured. For this purpose, we define W_h as dispersion within a cluster and B_h as dispersion between clusters.

We denote the dissimilarity between objects p and q as d_{pq} and the dissimilarity between clusters I and J as d_{IJ}. The number of objects to be classified is N and the cluster I at stage m $(1 \leq m < N)$ is denoted as $C_I(m)$. We write the fact that object p belongs to cluster $C_I(m)$ as $p \in C_I(m)$; the number of objects belonging to cluster $C_I(m)$ is written as n_I. To simplify the notation, we define ${}_{n_I}C_2 = 0$ if $n_I = 1$.

We assume that the clusters $C_I(m)$ are obtained by using some agglomerative hierarchical clustering algorithms. From this assumption, the number of the clusters at stage m is $N - m$.

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 and between cluster dispersion at stage m, respectively. We can define several dispersion measures of within a cluster and between clusters. For example,

$$\begin{aligned}
W_1(m) &= \max_I \max_{p,q \in C_I(m)} d_{pq}, \\
B_1(m) &= \min_{I,J} \min_{p \in C_I(m), q \in C_J(m)} d_{pq}, \\
W_2(m) &= \sum_I \left(\max_{p,q \in C_I(m)} d_{pq} \right) \Big/ (N - m), \\
B_2(m) &= \sum_{I,J} \left(\min_{p \in C_I(m), q \in C_J(m)} d_{pq} \right) \Big/ {}_{N-m}C_2, \\
W_3(m) &= \max_I \left(\sum_{p,q \in C_I(m)} d_{pq} / {}_{n_I}C_2 \right), \\
B_3(m) &= \min_{I,J} \left(\sum_{p \in C_I(m), q \in C_J(m)} d_{pq} / n_I n_J \right), \\
W_4(m) &= \sum_I \left(\sum_{p,q \in C_I(m)} d_{pq} / {}_{n_I}C_2 \right) \Big/ (N - m),
\end{aligned}$$

$$
\begin{aligned}
B_4(m) &= \sum_{I,J}\left(\sum_{p\in C_I(m),q\in C_J(m)} d_{pq}/n_I n_J\right)\Big/ {}_{N-m}C_2,\\
W_5(m) &= \sum_{I}\left(\sum_{p,q\in C_I(m)} d_{pq}\right)\Big/ \sum_{I} {}_{n_I}C_2,\\
B_5(m) &= \sum_{I,J}\left(\sum_{p\in C_I(m),q\in C_J(m)} d_{pq}\right)\Big/ \sum_{I,J} n_I n_J,
\end{aligned}
$$

where and hereafter we assume $I \neq J$.

Since the structured ratio is the ratio of dispersion within a cluster to between clusters, a smaller value is preferable in structured sense. The value of the structured ratio strongly depends on the dispersion measures.

3. Simulation

In this section, to make characterize the measures described in the above, we perform a simulation under paying attention to i) number of the clusters, ii) ratio of the number of objects into one cluster, and iii) dispersion index. The simulation is performed by the following processes;

1. Generate the center $\boldsymbol{\mu}_I = (\mu_I^1, \mu_I^2)$ of provisional cluster C_I, where $\mu_I^d \sim U(10, 40)$ $(d = 1, 2)$.

2. Generate the coordinates of 50 artificial objects. The coordinates of the k-th object belonging to C_I is represented by $\boldsymbol{X}_{Ik} = (X_{Ik}^1, X_{Ik}^2)$, where $X_{Ik}^d \sim N(\mu_I^d, \sigma_r)$ $(d = 1, 2), \sigma_r = r\sum_I(\mu_I^d - \mu_I^d)'(\mu_I^d - \mu_I^d)/(10M)$, M is a number of clusters, and r is a dispersion index.

3. Analyze this data by 8 popular hierarchical clustering algorithms (single linkage, complete linkage, weighted average, median, group average, centroid, Ward's, flexible $(\beta = -0.25)$).

4. Repeat 200 times from process 1 to 3 for each case, and calculate the average value of the SR_h of 200 trials.

In this paper, we generate $\boldsymbol{X}_{Ik}$ for the following factors that may influence the measures;

Number of clusters: $M = 2, 3, 4, 5$.
Number of objects: equal, 10%, 60% into one cluster.
Density index: $r = 1, 2, 3, 4, 5$.
Index of $SR_h(m)$: $h = 1, 2, 3, 4, 5$.

Then, the calculations will be performed on 2400 $(8 \times 4 \times 3 \times 5 \times 5)$ cases.

From the result, we obtain some information about the relationship between structured ratio and each of cases. For example, these measures attains its minimum close to $(50 - M)$ stage in almost cases; these measures must be independent on the number of the clusters, and so on. Specially, there are obvious differences between the inclinations of dispersion index and that of structured ratio.

REFERENCES

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

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

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

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

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

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

Takeuchi, A. Yadohisa, H., and Inada, K. (2000a) Measuring the structure of the agglomerative hierarchical clustering, *Proceedings of The International Conference on Measurement and Multivariate Analysis*, **1**, 18–20.

Takeuchi, A., Yadohisa, H, and Inada, I. (2000b), Evaluation and representation of the clustering results in agglomerative hierarchical clustering, *Nihonkoudoukeiryougakkai dai 28 kai taikaihappyouronbunsyarokusyuu*, 225–228 (in Japanese).

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.

RESUME

Le problèe de choisir un algorithme groupant partir de la myriade d'algorithmes est discuté ces dernières annèes. Beaucoup de chercheurs ont attaqué ce problème en utilisant le concept de l'admissibilité (par exemple Fisher et Van Ness (1971), Yadohisa, et autres (1999)). Takeuchi, et al. (2000) a proposé la mesure de dispersion la proportion structurée appelée. Dans ce papier, nous examinons le rapport entre la proportion structurée et les données analysées. Le but de ce papier devrait caractériser la proportion par la simulation.