Communications for Statistical Applications and Methods
2014, Vol. 21, No. 3, 225–234

DOI: http://dx.doi.org/10.5351/CSAM.2014.21.3.225
ISSN 2287-7843

Symbolic Cluster Analysis for Distribution Valued Dissimilarity

Yusuke Matsui[1,a], Hiroyuki Minami[b], Masahiro Misuta[b]

[a]Graduate School of Information Science and Technology, Hokkaido University, Japan
[b]Information Initiative Center, Hokkaido University, Japan

Abstract

We propose a novel hierarchical clustering for distribution valued dissimilarities. Analysis of large and complex data has attracted significant interest. Symbolic Data Analysis (SDA) was proposed by Diday in 1980's, which provides a new framework for statistical analysis. In SDA, we analyze an object with internal variation, including an interval, a histogram and a distribution, called a symbolic object. In the study, we focus on a cluster analysis for distribution valued dissimilarities, one of the symbolic objects. A hierarchical clustering has two steps in general: *find out* step and *update* step. In the *find out* step, we find the nearest pair of clusters. We extend it for distribution valued dissimilarities, introducing a measure on their order relations. In the *update* step, dissimilarities between clusters are redefined by mixture of distributions with a mixing ratio. We show an actual example of the proposed method and a simulation study.

Keywords: Symbolic Data Analysis, hierarchical clustering, big data, distribution valued data, network latency.

1. Introduction

These days, analysis of large complex data is a salient topic. We need to utilize significant computer resources to handle huge amount of data. Big Data is a recent keyword among researchers in data analysis and often characterized by "3V", *i.e., Volume, Velocity*, and *Variety*. When we consider their analysis, we should pay attention to data descriptions as well as its methodological efficiency. Diday and Brito (1989) proposed Symbolic Data Analysis (SDA), the new statistical framework, which extends conventional data analysis to various data descriptions, such as an interval, a histogram, and a distribution.

SDA is not only a method for knowledge extraction from large databases as *Data Mining*, but also an application tool to retrieve underlying concepts from extracted knowledge. The units of objects might be categories, classes or concepts. The objects with structured *internal variations* are called *Symbolic Data* (Bock and Diday, 2000). Typically, the data are multi valued data, interval valued data, modal data, and distribution valued data: Schweitzer (1968) remarked, "Distributions are the numbers of the future" and Diday and Vrac (2005) quoted the phrase and studied distribution valued data.

Many methods on SDA have been proposed (Bock and Diday, 2000; Billard and Diday, 2006; Diday and Noirhomme-Fraiture, 2008). Clustering remains an efficient method to identify unknown subgroups (Huh, 2002). Symbolic clustering is developed to find homogeneous subgroups of symbolic data. There are many studies for multi valued data, interval valued data, and modal data.

[1] Corresponding author: Information Initiative Center, Hokkaido University. N11-W5, Kita-ku, Sapporo, Hokkaido, 060-0811, Japan. E-mail: matsui@iic.hokudai.ac.jp

Published online 31 May 2014 / journal homepage: http://csam.or.kr

Symbolic clustering for distribution valued data is a popular topic. Diday and Vrac (2005) explored mixture decomposition of distribution valued data. Katayama *et al.* (2009) proposed hierarchical symbolic clustering, and Terada and Yadohisa (2010) studied k-means-like symbolic clustering for distribution valued data. These methods assumed that each input *object* is described as distribution valued data.

When we analyze dissimilarities, we usually use Multidimensional Scaling(MDS) or hierarchical clustering. Mizuta and Minami (2012) proposed MDS for distribution valued dissimilarities. We propose hierarchical symbolic clustering for distribution valued *dissimilarities*, *i.e.,* we assume that dissimilarities are represented by distributions. We show an actual example of the proposed method and a simulation study.

2. Hierarchical Clustering and Hierarchical Symbolic Clustering

In hierarchical clustering, two types of input data are considered. One is that data are p variate n vectors. In this case, we define dissimilarity measures between vectors. The other case is that, dissimilarities $\{s_{ij}\}$ for a pair of objects $i, j\,(i, j = 1, 2, \ldots, n)$ are measured in direct ways.

For hierarchical symbolic clustering, the type of n objects with p variables are much studied and a variety of dissimilarities are proposed (Billard and Diday, 2006). The dissimilarities could be incorporated with standard hierarchical clustering algorithms.

Katayama *et al.* (2009) studied hierarchical symbolic clustering where each n object is described as multivariate distribution valued data. Symmetric Kullback-Leibler divergence is applied to define the dissimilarities between the objects; consequently, the remains of the procedures are followed by a standard clustering scheme.

It is rarely considered that dissimilarities are described as symbolic data when they are directly measured. In the case, standard hierarchical clustering algorithms are not applicable since they are not designed for dissimilarities described by symbolic data; therefore, we need to develop a new clustering procedure for dissimilarities represented as symbolic data.

3. Proposed Method

Suppose there are N objects and we put S_{ij} as a dissimilarity between an object i and an object j. We assume the dissimilarities S_{ij} are given as distributions. We propose a clustering method for the dissimilarities. The clusters are denoted as $\{C_k;\ k = 1, 2, \ldots, K\}$.

Initialization

We initialize $K = N$ and $C_i = \{i\}$ $(i = 1, 2, \ldots, N)$.

Find out step

We define the nearest pair (i, j) as follows

$$(i, j) := \underset{q,r}{\operatorname{argmax}} \left\{ \Pr\left\{ S\left(C_i, C_j\right) \leq S\left(C_q, C_r\right);\ (i, j) \neq (q, r)\,,\ q < r,\ \text{for all}\, q, r \right\} \right\}. \tag{3.1}$$

We evaluate $\Pr\{S(C_i, C_j) \leq S(C_q, C_r)\}$ by relative frequencies.

Suppose that we have M realizations of $\{S(C_i, C_j);\ i, j = 1, 2, \ldots, K,\ i < j,\ \text{for all}\, i, j\}$, and we denote b^{th} $(b = 1, 2, \ldots, m)$ realization by $\{s_{ij,b}\}$.

Relative frequencies of $\{S(C_i, C_j) \leq S(C_q, C_r); (i, j) \neq (q, r),\ q < r, \text{ for all } q, r\}$ are

$$\frac{\sum_{b=1}^{m} \mathbf{1}_{ij,b}}{m}, \tag{3.2}$$

where $\mathbf{1}_{ij,b}$ is defined by

$$\mathbf{1}_{ij,b} := \begin{cases} 1, & \text{if } (i, j) = \underset{q,r}{\operatorname{argmin}} \left\{ s_{qr,b} \right\}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

Thus we obtain the nearest pair by

$$(i, j) = \underset{i,j}{\operatorname{argmax}} \left\{ \frac{\sum_{b=1}^{m} \mathbf{1}_{ij,b}}{m} \right\}. \tag{3.4}$$

Update step

Now the clusters C_i and C_j are merged into a new cluster; $C_{new} = C_i \cup C_j$. Then we have to update dissimilarities between them. We mix the distributions of $S(C_i, C_q)$ and $S(C_j, C_q)$ with a mixing proportion p, taking account of the number of clusters in C_i and C_j;

$$p = \frac{N_{C_i}}{N_{C_i} + N_{C_j}}, \tag{3.5}$$

where N_{C_i} is the number of the objects in C_i. We put a cumulative distribution function $F_{C_iC_j}(s)$ of $S(C_i, C_j)$, then a new distribution valued dissimilarities $S(C_{new}, C_q)$ are defined by

$$F_{C_{new}C_q}(s) = pF_{C_iC_q}(s) + (1 - p) F_{C_jC_q}(s). \tag{3.6}$$

We decrement K by 1 and repeat the steps until $K = 1$.

We implement the proposed method with R language. In conventional hierarchical clustering algorithms, the dissimilarity matrix is used in the procedures. However, since we deal with distribution valued dissimilarities, we consider dissimilarity arrays with 3 dimensions `s[i,j,m]` $(m = 1, 2, \ldots, M)$ in which 3rd dimension is a m^{th} value from dissimilarity S_{ij}. In the *find out* step, we randomly extract M' $(M' \leq M)$ values from the array `s[i,j,]`. We adopt (3.3) and (3.4) to those values then we get the index of the nearest pair of the clusters; (i, j). To recompute the dissimilarity $S(C_i \cup C_j, C_q)$ in the *update* step, we resample $M' \times p$ and $M' \times (1 - p)$ values from `s[i,q,]` and `s[j,q,]` respectively where p is defined in (3.5), then they are randomly mixed so that we get a new element of the dissimilarity array. The treatment of ties of dissimilarities is important. We adopt a random selection in those cases.

4. Actual Example

We perform the proposed method with actual data set on network delays. The virtual distance between two sites in the Internet can be assumed by Round Trip Time (RTT). We obtained the dataset which had been performed to 35 sites for over 2 years. Each measurement was carried out every 5 minutes. The classification of the sites provides us an insight on network performance and future reinforcement.

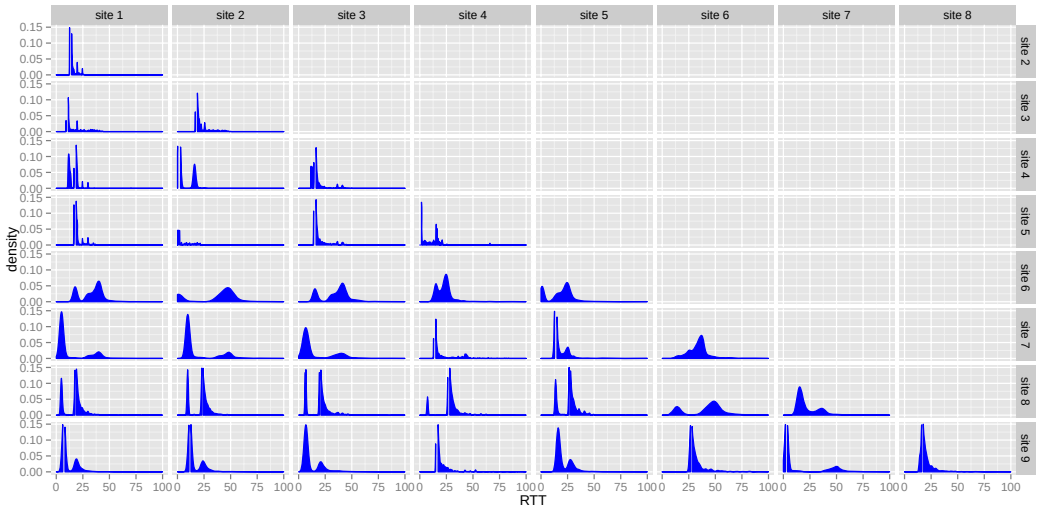


Figure 1: *Plot of distribution valued dissimilarity array (RTT data)*

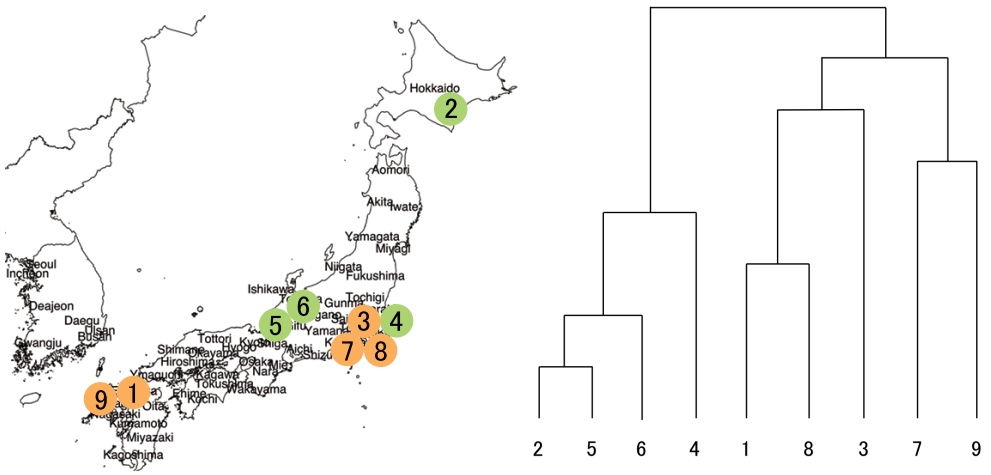
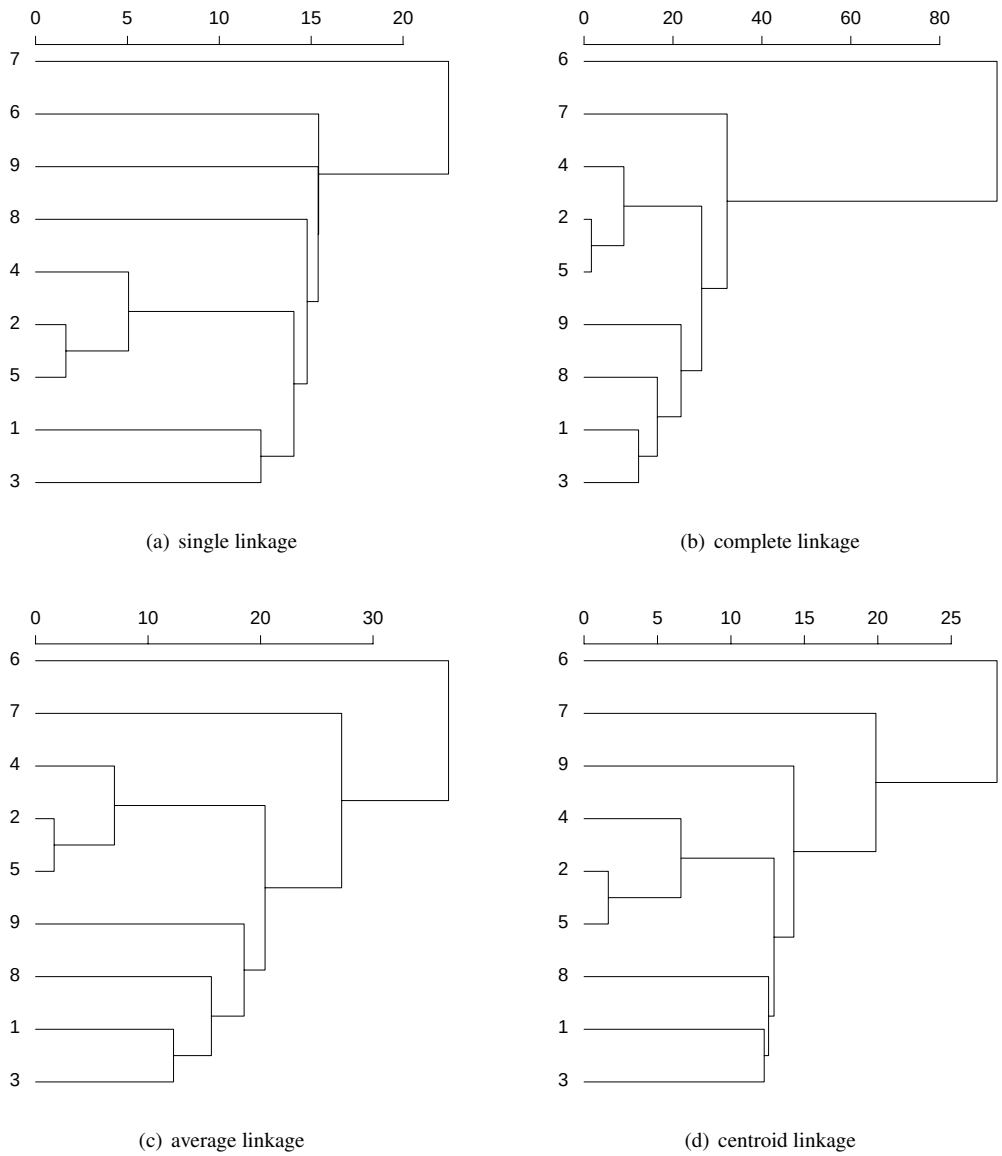


Figure 2: *Results of the proposed method (RTT data)*

Unfortunately, the dataset has many NAs. In the end, we obtain the complete (measured by both directions) data among 9 sites over 4 days. RTTs were measured from both directions, then we averaged them. We label the sites from *site* 1 to *site* 9.

Figure 1 shows the array of the distribution valued dissimilarities between each pair of the 9 sites. Most have two peaks. The peaks reminds us of an economical scandal and the negative effects in the Stock Market in Japan. We are wondering that it accelerated the amount of the network traffic.

We apply the proposed method (Figure 2). To compare the results to the conventional ones, we take the means over the distributions and apply the conventional hierarchical clustering, single linkage, complete linkage, average linkage, centroid linkage to them (Figure 3).



(a) single linkage (b) complete linkage

(c) average linkage (d) centroid linkage

Figure 3: *Conventional hierarchical clustering with means of RTT*

We interpret Figure 2 and Figure 3. In Figure 2, we classify the data into 2 clusters- *site* 2, 4, 5, 6 and *site* 1, 3, 7, 8, 9. The difference may be from network bandwidths. *Site* 2, 5, 6 belong to eastern area of Japan, and *site* 1, 3, 7, 8, 9 belong to western area (*site* 4 is considered as an exception since it is the Internet exchange point in Japan). From the interpretations, we reconfirm that the bandwidth of eastern area in Japan is narrower than that of western area in the period. In Figure 3, there is no appropriate interpretation even if we change the number of the classes.

Table 1: Coordinates of initial points

	x_1	x_2
1	−1.80	2.18
2	−2.07	3.02
3	1.64	1.84
4	1.16	5.39
5	0.16	−2.36
6	−0.15	1.04
7	−3.60	−0.90
8	0.45	1.01
9	2.82	−5.33

Table 2: Distances of initial points

	1	2	3	4	5	6	7	8
2	0.882							
3	3.457	3.893						
4	4.366	4.006	3.582					
5	4.945	5.824	4.453	7.814				
6	2.006	2.758	1.961	4.543	3.414			
7	3.567	4.208	5.913	7.888	4.034	3.958		
8	2.536	3.223	1.451	4.437	3.382	0.601	4.478	
9	8.817	9.676	7.266	10.848	3.987	7.028	7.800	6.768

5. Simulation Study

We show a simulation study. We assume that there are 9 objects and their dissimilarities are assumed by gamma distributions. The parameters are assigned under the condition that the means of distributions are fixed. That is, if we apply conventional hierarchical clustering with means of dissimilarities, the results must be the same.

We generate 9 points in two dimensional space with multivariate normal distribution which are from $N\left(\binom{1}{1}, 5\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)$ (Table 1). We calculate the Euclidian distances $\{s_{ij}^{(init)};\ i < j\}(i, j = 1, 2, \ldots, 9)$ (Table 2). With the mean of distributions, we generate gamma distributions where the each mean is fixed by $\{s_{ij}^{(init)}\}$, *i.e.,* we put shape parameters $\{\alpha_{ij};\ i < j\}(i, j = 1, 2, \ldots, 9)$ and scale parameter $\{\beta_{ij};\ i < j\}(i, j = 1, 2, \ldots, 9)$, where means are fixed as $\alpha_{ij}\beta_{ij} = s_{ij}^{(init)}$ (Table 3).

We apply the conventional hierarchical clustering, single linkage, complete linkage, average linkage and centroid linkage with $\{s_{ij}^{(init)};\ i < j\}(i, j = 1, 2, \ldots 9)$. The dendrograms are shown in Figure 4.

We generate α_{ij} from Uniform distribution U(1,3) and then we get $\beta_{ij} = s_{ij}^{(init)}/\alpha_{ij}$. We use 1000 samples. For the distribution valued dissimilarities $\{s_{ij,b};\ b = 1, 2, \ldots, 1000\}$, we apply the proposed method (Figure 5).

The differences of 4 dendrograms indicate the dispersions of the dissimilarities. When we consider only means with the conventional methods, the results should be similar. However, when we take the dispersions of dissimilarities such as variance and skewness into the analysis, the result show that constitution of the clusters would be varied even when statistical expectations are identical.

For instance, object 5 and 9 are merged in simulation 1, but in simulation 2, object 5 and 7 are merged. To explain this difference, we investigate quartile of $s_{5,9}$ and $s_{5,7}$: $(q_{0.25}, q_{0.5}, q_{0.75})$ where suffix indicates percentages. In simulation 1, quartile of $s_{5,9}$ is $(1.787, 3.266, 5.415)$ and that of $s_{5,7}$ is $(2.210, 3.536, 5.315)$. Then it seems reasonable to merge object 5 and 9 since they are more similar

Table 3: Parameters of gamma distributions in 4 simulations

	Simulation 1		Simulation 2		Simulation 3		Simulation 4	
	α_{ij}	β_{ij}	α_{ij}	β_{ij}	α_{ij}	β_{ij}	α_{ij}	β_{ij}
$s_{1,2}$	2.581	0.341	2.748	0.320	1.673	0.526	1.257	0.700
$s_{1,3}$	1.543	2.236	2.214	1.559	2.595	1.330	2.674	1.290
$s_{1,4}$	1.025	4.257	1.476	2.956	2.127	2.051	1.012	4.311
$s_{1,5}$	1.002	4.929	1.743	2.834	2.008	2.460	2.467	2.003
$s_{1,6}$	1.913	1.049	1.303	1.540	1.513	1.326	2.908	0.690
$s_{1,7}$	1.173	3.043	1.955	1.827	1.951	1.830	2.214	1.613
$s_{1,8}$	1.085	2.339	2.220	1.143	2.130	1.191	2.506	1.012
$s_{1,9}$	2.402	3.670	2.216	3.979	2.660	3.315	1.832	4.812
$s_{2,3}$	2.647	1.471	1.319	2.951	1.703	2.286	1.479	2.632
$s_{2,4}$	2.346	1.710	1.480	2.711	2.513	1.596	2.879	1.393
$s_{2,5}$	2.223	2.617	2.661	2.187	1.294	4.496	1.243	4.682
$s_{2,6}$	2.963	0.932	1.993	1.386	1.084	2.548	2.527	1.092
$s_{2,7}$	2.754	1.526	2.885	1.457	1.027	4.092	1.704	2.467
$s_{2,8}$	2.384	1.354	1.104	2.924	2.607	1.238	2.969	1.088
$s_{2,9}$	2.913	3.322	2.975	3.252	1.723	5.616	1.672	5.786
$s_{3,4}$	1.422	2.521	2.162	1.658	2.232	1.606	2.802	1.279
$s_{3,5}$	2.158	2.064	2.797	1.592	1.802	2.470	1.166	3.818
$s_{3,6}$	1.781	1.097	1.281	1.526	1.234	1.583	2.645	0.739
$s_{3,7}$	1.257	4.702	2.605	2.269	2.472	2.391	1.874	3.154
$s_{3,8}$	2.614	0.553	2.247	0.644	2.982	0.485	1.529	0.946
$s_{3,9}$	1.341	5.423	1.509	4.817	1.238	5.870	1.284	5.662
$s_{4,5}$	2.549	3.066	1.777	4.397	1.190	6.570	1.129	6.923
$s_{4,6}$	2.513	1.808	2.963	1.533	2.078	2.187	1.750	2.596
$s_{4,7}$	1.403	5.624	1.834	4.300	2.341	3.369	2.951	2.673
$s_{4,8}$	1.239	3.583	1.282	3.464	2.553	1.739	1.970	2.254
$s_{4,9}$	1.769	6.137	1.436	7.559	1.946	5.576	2.086	5.204
$s_{5,6}$	2.090	1.632	1.895	1.800	1.981	1.722	1.800	1.895
$s_{5,7}$	2.654	1.518	2.490	1.618	1.815	2.219	2.172	1.854
$s_{5,8}$	1.914	1.765	2.657	1.272	1.411	2.394	1.101	3.069
$s_{5,9}$	1.763	2.263	2.972	1.343	1.048	3.807	1.943	2.054
$s_{6,7}$	2.291	1.729	2.874	1.378	2.132	1.858	1.747	2.267
$s_{6,8}$	1.747	0.344	2.755	0.218	2.815	0.213	1.726	0.348
$s_{6,9}$	1.650	4.258	2.887	2.434	1.097	6.404	1.387	5.068
$s_{7,8}$	1.862	2.406	1.174	3.816	1.027	4.364	2.122	2.111
$s_{7,9}$	2.213	3.524	1.789	4.359	2.217	3.516	2.356	3.309
$s_{8,9}$	1.306	5.179	2.368	2.858	2.538	2.666	1.351	5.009

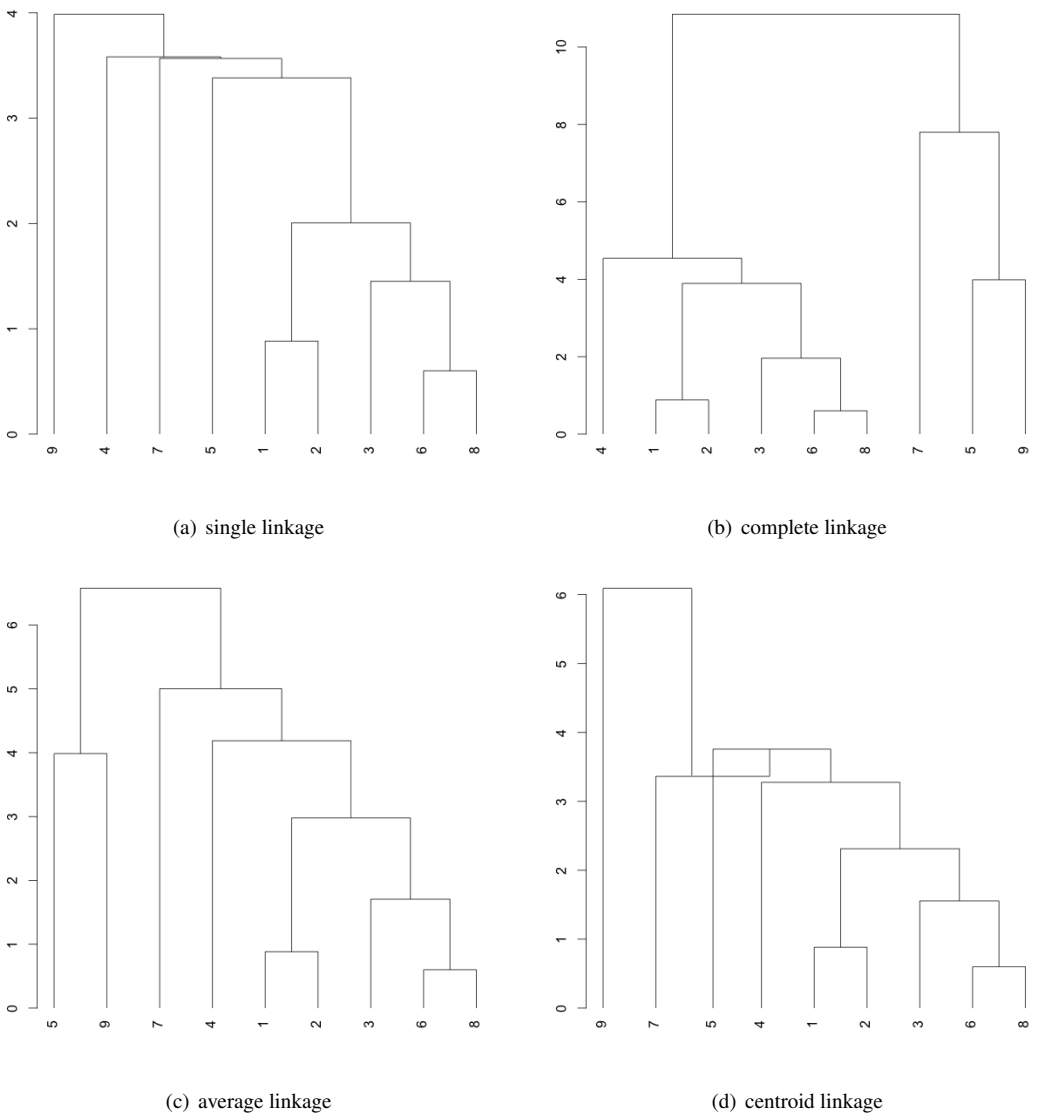
than object 5 and 7 in the sense of quartile. In simulation 2, quartile of $s_{5,9}$ is $(2.290, 3.554, 5.220)$ and that of $s_{5,7}$ is $(2.151, 3.504, 5.340)$, then it appears good to merge object 5 and 7.

In this way, the results of the proposed method reflect the variations of distributions and we could obtain feasible clusters from distribution valued dissimilarities.

6. Concluding Remarks

We study the hierarchical clustering for distribution valued dissimilarities. The actual example indicated that our method provides a feasible interpretation than conventional hierarchical clustering with statistical summaries. We assume that input data are distributional dissimilarities, but in real situations, we cannot obtain a set of values, not distributions. In order to deal with them as distributions, we need many observations.

SDA is quite effective for Big Data analysis since it has various data descriptions and a main feature of Big Data is *variety*; therefore, the proposed method is suitable to Big Data. The idea



(a) single linkage

(b) complete linkage

(c) average linkage

(d) centroid linkage

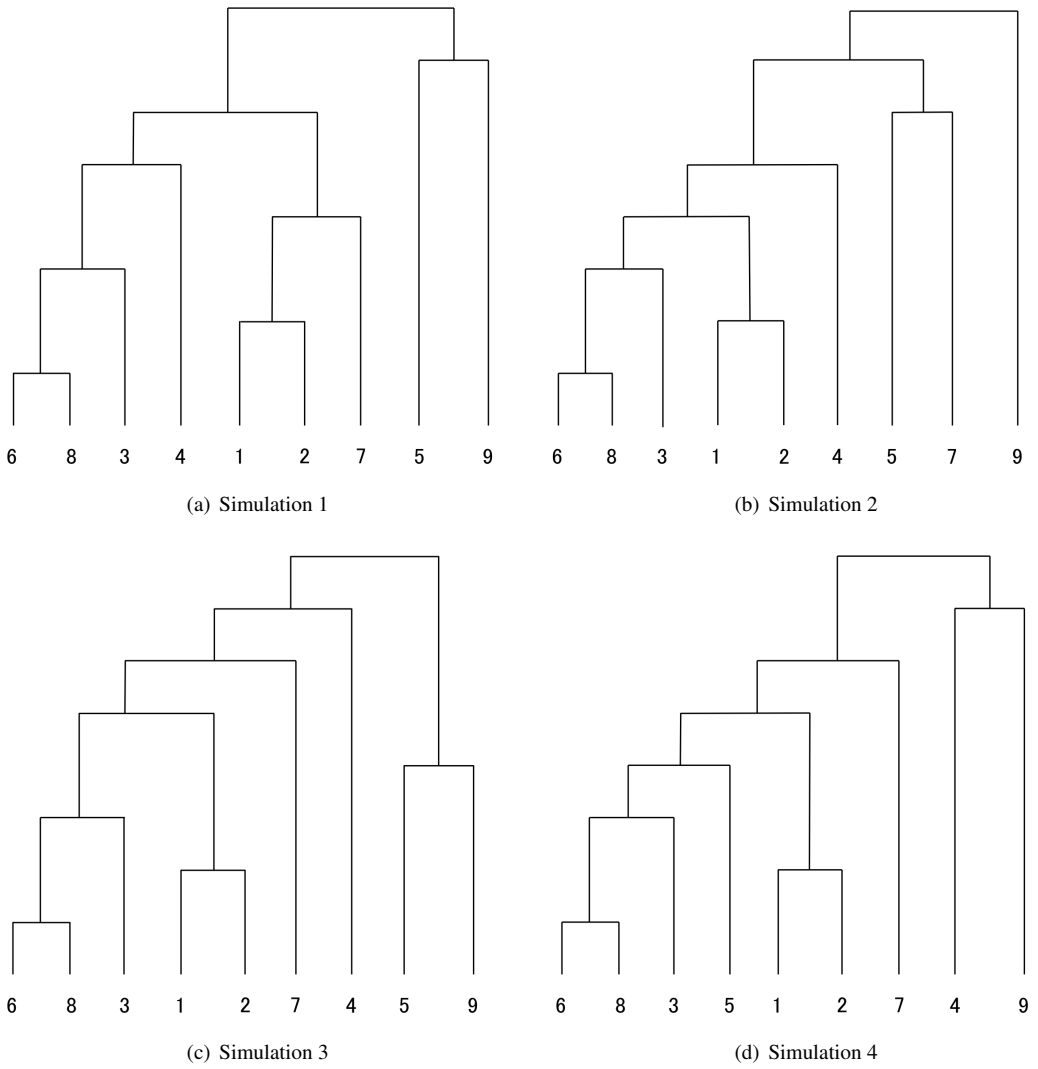
Figure 4: *Conventional hierarchical clustering with means of dissimilarities*

can be extended to other methods to deal with distribution valued dissimilarities such as symbolic multidimensional scaling.

References

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley, Chichester.

Bock, H. H. and Diday, E. (2000). *Analysis of Symbolic Data*, Springer, Berlin Heidelberg.



(a) Simulation 1 (b) Simulation 2

(c) Simulation 3 (d) Simulation 4

Figure 5: *Simulation results of the proposed method*

Diday, E. and Brito, M. P. (1989). Symbolic Cluster Analysis, In: Optiz, Otto (eds.), *Conceptual and Numerical Analysis of Data*, 45–84, Springer, Berlin Heidelberg.

Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*, Wiley-Interscience.

Diday, E. and Vrac, M. (2005). Mixture decomposition of distributions by copulas in the symbolic data analysis framework, *Discrete Applied Mathematics*, **147**, 27–41, Elsevier Science Publishers B. V, Amsterdam.

Huh, M. H. (2002). Setting the Number of Clusters in K-Means Clustering, In: Baba, Y., Hayter, A. J., Kanefuji, K. and Kuriki, S. (eds.), *Recent Advances in Statistical Research and Data Analysis*, 115–124, Springer, Tokyo.

Katayama, K., Minami, H. and Mizuta, M. (2009). Hierarchical symbolic clustering for distribution valued data, *Journal of the Japanese Society of Computational Statistics*, **22**, 83–89 (In Japanese).

Matsui, Y., Komiya, Y., Minami, H. and Mizuta, M. (2013). Comparison of Two Distribution Valued Dissimilarities and Its Application for Symbolic Clustering, In: Gaul, W., Geyer-Schulz, A., Baba, Y. and Okada, A. (eds.), *German-Japanese Interchange of Data Analysis Results*. Studies in Classification, Data Analysis, and Knowledge Organization. (to appear), Springer, Heidelberg.

Matsui, Y., Minami, H. and Mizuta, M. (2013). Symbolic Cluster Analysis for Distribution Valued Data, In: Cho, S. H. (eds.), *Proceedings of Joint Meeting of the IASC Satellite Conference and the 8th Conference of the Asian Regional Section of the IASC*, 305–310, Aug. 22–23, 2013, Yonsei University, Seoul, Korea.

Mizuta, M. and Minami, H. (2012). Analysis of Distribution Valued Dissimilarity Data. In: Gaul, W. A., Geyer-Schulz, A., Schmidt-Thieme, L. and Kunze, J., *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, Studies in Classification, Data Analysis, and Knowledge Organization, 23–28, Springer, Heidelberg.

Schweizer, B. (1968). Distributions are the numbers of the future, *Proceedings section Napoli Meeting on "The mathematics of fuzzy systems"*, 137–149, Instituto di Mathematica delle Faculta di Achitectura, Universita degli studi di Napoli.

Terada, Y. and Yadohisa, H. (2010). Non-hierarchical clustering for distribution-valued data, *COMPSTAT 2010: Proceedings in Computational Statistics*, 1653–1660, Psysica-Verlag, Heidelberg.

Received January 9, 2014; Revised April 17, 2014; Accepted May 7, 2014