Clustering in Contemporary Mixed-valued Data

L. Billard[1,3], and Jaejik Kim[2]
[1] *University of Georgia, Department of Statistics, Athens Georgia 30602 USA,*
[2] *Georgia Regents University, Department of Biostatistics Augusta Georgia 30912 USA*
[3] *Corresponding author: L. Billard, email: lynne@stat.uga.edu*

Abstract

This work considers clustering for mixed symbolic data whereby realizations of some random variables are interval-valued, some histogram-valued, and some are modal and/or non-modal multi-valued.
Keywords: extended Gowda-Diday and Ichino-Yaguchi dissimilarities, agglomerative clustering.

1. Introduction

Contemporary databases can be too large to be analysed by traditional methods. One approach is to aggregate the data according to whatever scientific question(s) are driving the analysis. The resulting data are perforce intervals, lists, histograms, and the like, which formats are but examples of symbolic data (Diday, 1987). This work will look at clustering for data sets in $\mathcal{R}^p$ of mixed types whereby some random variables take interval realizations, some histogram realizations, and/or some are modal or non-modal multi-valued realizations. For example, a census collects information on individuals. These individual values are aggregated according to some geographical-social-scientific question(s) of interest, such as by region, city, state, age×gender, income, etc. Depending on the nature of the original observations and the nature of the aggregation, the data are subsequently recorded as, e.g., histograms over a range of suitable subintervals, lists of possible categorical entities, and so forth.

Our focus is on data which contain a mixture of types of symbolic realizations. This will first involve the calculation of distance/dissimilarity measures. Such measures have been introduced for interval and non-modal categorical data by Gowda and Diday (1991) and Ichino and Yaguchi (1994) with adaptations of the Ichino-Yaguchi distances for intervals developed by de Carvalho (1994, 1998). Chavent (1998) used the Hausdorff (1937) distance in a divisive clustering for interval data, while Kim and Billard (2012) developed a divisive clustering for multi-modal data. More recently, Kim and Billard (2013) introduced dissimilarity measures for histogram-valued realizations and for modal categorical data. In Section 3, we combine these concepts so as to obtain a distance/dissimilarity matrix for a set of data that has interval, histogram and categorical data. Then, in Section 4, we obtain clusters using an agglomerative clustering algorithm. The methodology is illustrated through the a mixed-valued data set obtained from census records, described in Section 2.

2. The Data

We have a random sample of m realizations of the random variable $\mathbf{Y} = (Y_1, \ldots, Y_p)$ taking values in $\mathcal{R}^p$. Modal categorical realizations take the form

$$(1) \quad Y_{uj} = \{\xi_{ujk}, p_{ujk};\ k = 1, \ldots, s_{uj}\}, \quad \sum_{k=1}^{s_{uj}} p_{ujk} = 1, \quad j = 1, \ldots, p, \quad u = 1, \ldots, m,$$

where ξ_{ujk}, $k = 1, \ldots, s_{uj}$, is the list of categorical values from the set of possible categorical values $\mathcal{Y}_j$, $j = 1, \ldots, p$, that actually occurred for observation $u = 1, \ldots, m$, and where p_{ujk} is the relative frequency or probability associated with the value ξ_{ujk}. Without loss of generality, we can write

$s_{uj} = s_j$ and set $p_{ujk} \equiv 0$ for those categories k that do not occur in Y_{uj}.

When realizations are histogram-valued, they take the form

$$(2) \quad Y_{uj} = \{[b_{ujk}, b_{uj,k+1}), p_{ujk};\ k = 1, \ldots, s_{uj}\}, \quad j = 1, \ldots, p, \quad u = 1, \ldots, m,$$

where the histogram Y_{uj} consists of s_{uj} subintervals $[b_{ujk}, b_{uj,k+1})$ (with $b_{ujk} \leq b_{uj,k+1}$) occurring with relative frequency p_{ujk}. Typically, the length and number of subintervals will vary across observations and variables. However, without loss of generality, the observations can be transformed into histograms with common subintervals and the same number of subintervals. Therefore, for all terms in the right-side of (2), except for the relative frequency term p_{ujk}, the u subscript can be dropped; see Kim and Billard (2013) for details. While not necessary theoretically, this transformation allows for easier computational efficiency.

Non-modal realizations are special cases of (1) and (2). Thus, a list of categories from $\mathcal{Y}_j$, $j = 1, \ldots, p$, has the same format as in (1) but where now each of the possible values that actually occur is assumed to be equally likely and so has in effect a relative frequency of $1/s_{uj}$. Interval data are special cases of histograms in (2) with $s_{uj} \equiv 1$ and hence $p_{ujk} = 1$ for all k, j, u.

The data set considered herein consists of realizations of household characteristics for $m = 10$ counties (Fresno, Humboldt, Lassen, Mariposa, Merced, Napa, Orange, Riverside, San Joaquin, San Mateo). Specifically, we have Y_1 = Age (with $s_1 = 12$ histogram subintervals: {[0, 4), [4,17), [17, 20), [20,24), [24, 34), [34,44), [44,54), [54, 64), [64, 74), [74, 84), [84, 94), [94, 120]} years old), Y_2 = Home value (with $s_2 = 7$ histogram subintervals: {[0, 49), [49, 99), [99, 149), [149, 199), [199, 299), [299, 499), [499, 1000]} in \$1000's), Y_3 = Gender (with modal categories $\mathcal{Y}_3$ = {male, female}), Y_4 = Fuel type used in the home (with non-modal categorical values taken from $\mathcal{Y}_4$ = {gas, electricity, coal, oil}), Y_5 = Tenure (with modal categorical values taken from $\mathcal{Y}_5$ = {owner occupied, renter occupied, vacant}), and Y_6 = Income (with interval values from the real line $\mathcal{R}$). To illustrate realizations for Fresno County are shown in Table 1. The data were extracted from Census (2000).

Table 1 - Census Mixed-valued Observations

Variable	Symbolic realizations for Fresno					
Age	{[0,4),.085;	[4,17),.236;	[17,20),.051;	[20,24),.060;	[24,34),.140;	[34,44),145;
	[44,54),.115;	[54,64].039;	[64,74),.031;	[74,84),.052;	[84,94),.034;	[94,120],.012}
Home value	{[0,49),.037;	[49,99),.432;	[99,149),.290;	[149,199),.126;	[199,299),.080;	[299,499),.028;
	[499,1000],.009}					
Gender	{ male, .499;	female,.501}				
Fuel	{gas,	electricity}				
Tenure	{owner,.527;	renter,.407;	vacant,.066}			
Income	[23.7, 44.8]					

3. Distance-Dissimilarity Measures

Unlike classical realizations (which are points in $\mathcal{R}^p$,), symbolic realizations are hypercubes in $\mathcal{R}^p$ or lists in $\mathcal{Y}^p$, and so observations can overlap. One consequence is that in deriving distance/dissimilarity measures between any two observations, it is necessary to obtain meet and join terms when dealing with non-modal data and union and intersections terms for modal data (i.e., for modal categorical and histogram data). The Gowda-Diday dissimilarity between two observations Y_{u_1} and Y_{u_2} is

$$(3) \quad d(Y_{u_1}, Y_{u_2}) = \sum_{j \in H} d_j(Y_{u_1}, Y_{u_2}) + \frac{3}{2} \sum_{j \in C} d_j(Y_{u_1}, Y_{u_2}), \quad u_1, u_2 = 1, \ldots, m,$$

where H is the set of variables j which are interval or histogram valued and C is the set with modal or non-modal categorical values. Since the Gowda-Diday dissimilarities have two terms for categorical realizations and three for interval-histograms realizations, the 3/2 factor is necessary so that each component has the same range measure.

The appropriate dissimilarities for non-modal categorical and interval data are well known and were established by Gowda and Diday (1991). For observations with Y_j taking modal categorical values, Kim and Billard (2012) extended the original Gowda and Diday (1991) results to give the dissimilarity, $d_j(Y_{u_1}, Y_{u_2})$, between Y_{u_1} and Y_{u_2} as

$$(4) \quad d_j(Y_{u_1}, Y_{u_2}) = d_{1j}(Y_{u_1}, Y_{u_2}) + d_{2j}(Y_{u_1}, Y_{u_2}), \quad u_1, u_2 = 1, \ldots, m,$$

where

$$(5) \quad d_{1j}(Y_{u_1}, Y_{u_2}) = \frac{\sum_{k=1}^{s_j} |p_{u_1jk} - p_{u_2jk}|}{\sum_{k=1}^{s_j} p_{(u_1 \cup u_2)jk}}, \quad d_{2j}(Y_{u_1}, Y_{u_2}) = \frac{\sum_{k=1}^{s_j} (p_{u_1jk} + p_{u_2jk} - 2p_{(u_1 \cap u_2)jk})}{\sum_{k=1}^{s_j} p_{(u_1 \cup u_2)jk}}$$

where the union $p_{(u_1 \cup u_2)jk}$ and intersection probabilities $p_{(u_1 \cap u_2)jk}$ are, respectively,

$$(6) \quad p_{(u_1 \cup u_2)jk} = \max\{p_{u_1jk}, p_{u_2jk}\}, \quad p_{(u_1 \cap u_2)jk} = \min\{p_{u_1jk}, p_{u_2jk}\}.$$

When the realizations for Y_j are histogram valued, the extended Gowda-Diday dissimilarity is

$$(7) \quad d_j(Y_{u_1}, Y_{u_2}) = D_{1j}(Y_{u_1}, Y_{u_2}) + D_{2j}(Y_{u_1}, Y_{u_2}) + D_{3j}(Y_{u_1}, Y_{u_2})$$

where

(8)

$$D_{1j}(Y_{u_1}, Y_{u_2}) = \frac{|S_{u_1j} - S_{u_2j}|}{(S_{u_1j} + S_{u_2j})}, D_{2j}(Y_{u_1}, Y_{u_2}) = \frac{S_{u_1j} + S_{u_2j} - 2S_{(u_1 \cap u_2)j}}{(S_{u_1j} + S_{u_2j})}, D_{3j}(Y_{u_1}, Y_{u_2}) = \frac{|\bar{Y}_{u_1j} - \bar{Y}_{u_2j}|}{\Psi_j}$$

where $\bar{Y}_{uj}$ and S_{uj} are the mean and standard deviation, respectively, for a single observation (Y_u, $u = u_1, u_2$), $S_{(u_1 \cap u_2)j}$ is the standard deviation of the intersection of Y_{u_1} and Y_{u_2}, and $\Psi_j = (b_{js_j} - b_{j1})$ is the span of values across all $\{Y_u, u = 1, \ldots, m\}$. See Kim and Billard (2013) for details and examples.

The generalized Minkowski distance based on, e.g., the Ichino-Yaguchi dissimilarity, is

$$(9) \quad d^{(q)}(Y_{u_1}, Y_{u_2}) = \left[\sum_{j \in H} d_j^q(Y_{u_1}, Y_{u_2}) + \sum_{j \in C} d_j^q(Y_{u_1}, Y_{u_2}) \right]^{1/q}, \quad u_1, u_2 = 1, \ldots, m.$$

Ichino and Yaguchi (1994) developed dissimilarities between two non-modal, and two interval, observations. These were extended to their modal counterparts in Kim and Billard (2013). Thus, for a given variable Y_j, the dissimilarity between two modal categorical observations becomes

$$(10) \quad d_j(Y_{u_1}, Y_{u_2}) = \sum_{k=1}^{s_j} [p_{(u_1 \cup u_2)jk} - p_{(u_1 \cap u_2)jk} + \gamma(2p_{(u_1 \cap u_2)jk} - p_{u_1jk} - p_{u_2jk})]$$

with union and intersection probabilities as in (6) and with $0 \leq \gamma \leq 0.5$ a pre-specified constant.

For histogram realizations, the normalized extended Ichino-Yaguchi dissimilarity is, for Y_j,

$$(11) \quad d_j(Y_{u_1}, Y_{u_2}) = [S_{(u_1 \cup u_2)j} - S_{(u_1 \cap u_2)j} + \gamma(2S_{(u_1 \cap u_2)j} - S_{u_1 j} - S_{u_2 j})]/N_j$$

where $S_u, u = u_1, u_2$, is the standard deviation of the histogram Y_u and $S_{(u_1 \cup u_2)j}$ and $S_{(u_1 \cap u_2)j}$ are the standard deviations of the union and intersection of the Y_{u_1} and Y_{u_2} histograms, and where the normalization factor N_j is given by

$$(12) \quad N_j^2 = (5V_{1j} + 2V_{2j} - 6V_{3j})/24$$

$$(13) \quad V_{1j} = b_{j1}^2 + b_{j2}^2 + b_{j,s_j-1}^2 + b_{js_j}^2, \quad V_{2j} = b_{j1}b_{j2} + b_{j,s_j-1}b_{js_j}, \quad V_{3j} = (b_{j1} + b_{j2})(b_{j,s_j-1} + b_{js_j}).$$

See Kim and Billard (2013).

To illustrate, suppose we want to calculate the Gowda-Diday dissimilarity matrix for the Census county data. Then, we use (7)-(8) for the histogram variables Y_1 and Y_2; we use (4)-(5) for the modal categorical variables Y_3 and Y_5; we use the special case of non-modal categorical data from Gowda and Diday (1991) for the variable Y_4; and we use the special case from Gowda and Diday (1991) for the interval data of Y_6. Then, from (3), we obtain the Gowda-Diday dissimilarity matrix as

$$(14) \quad \mathbf{D} = \begin{bmatrix} 0 & . & . & . & . & . & . & . & . & . \\ 2.450 & 0 & . & . & . & . & . & . & . & . \\ 4.616 & 3.462 & 0 & . & . & . & . & . & . & . \\ 3.534 & 3.237 & 3.379 & 0 & . & . & . & . & . & . \\ 0.484 & 2.472 & 4.376 & 3.297 & 0 & . & . & . & . & . \\ 4.379 & 4.565 & 6.768 & 6.404 & 4.701 & 0 & . & . & . & . \\ 3.002 & 4.980 & 7.231 & 5.876 & 3.212 & 3.192 & 0 & . & . & . \\ 1.907 & 3.734 & 5.548 & 4.033 & 2.163 & 3.558 & 2.368 & 0 & . & . \\ 1.462 & 3.488 & 5.882 & 4.425 & 1.767 & 3.384 & 2.016 & 1.048 & 0 & . \\ 3.612 & 5.237 & 7.601 & 6.138 & 3.606 & 4.095 & 1.526 & 3.505 & 3.12 & 0 \end{bmatrix}.$$

If instead we want to calculate the Euclidean dissimilarity matrix based on the Ichino-Yaguchi dissimilarities, then we use (11)-(13) for Y_1 and Y_2, we use (10) for Y_3 and Y_5; and the special cases for non-modal categorical data Y_4 and interval data Y_6 from Ichino and Yaguchi (1994). Then, the respective dissimilarities are substituted into (9) with $q = 2$. The Euclidean dissimilarity matrix is

$$(15) \quad \mathbf{D} = \begin{bmatrix} 0 & . & . & . & . & . & . & . & . & . \\ 0.753 & 0 & . & . & . & . & . & . & . & . \\ 1.530 & 0.804 & 0 & . & . & . & . & . & . & . \\ 0.805 & 0.795 & 0.778 & 0 & . & . & . & . & . & . \\ 0.045 & 0.753 & 1.526 & 0.804 & 0 & . & . & . & . & . \\ 0.849 & 0.871 & 1.579 & 1.568 & 0.857 & 0 & . & . & . & . \\ 0.627 & 1.014 & 1.662 & 1.011 & 0.642 & 0.798 & 0 & . & . & . \\ 0.290 & 0.813 & 1.533 & 0.794 & 0.285 & 0.784 & 0.442 & 0 & . & . \\ 0.272 & 0.820 & 1.557 & 0.851 & 0.285 & 0.778 & 0.366 & 0.181 & 0 & . \\ 0.838 & 1.157 & 1.754 & 1.145 & 0.847 & 0.878 & 0.228 & 0.658 & 0.595 & 0 \end{bmatrix}.$$

The literature includes a number of other distance matrices. For example, Irpino and Verde (2006) developed a type of Wasserstein distance using inverse cumulative distributions. DeCarvalho (1994, 1998) derived extensions to the Ichino-Yaguchi distances. Kim and Billard (2013) extended the deCarvalho (1994, 1998) distances to histogram data. Kim (2009) reviews these among others. Any one of these distances can then be substituted into (9) to obtain relevant Minkowski distances.

4. Clustering

There are many hierarchical clustering procedures for classical data and for interval and histogram data. See, e.g., Gordon (1999) for a review of classical procedures. Chavent (1998) has developed a divisive procedure for interval data, and Kim and Billard (2011) has a polythetic divisive procedure for histogram data, among others. Others, such as de Carvalho et al. (2008), have considered partitioning techniques. All these methodologies assume that all variables are of the same type (e.g., all interval realizations).

In this work, we implement an agglomerative procedure to our mixed-value data. There are many possible agglomerative methods in the literature. These include single-link (or "nearest neighbor") methods, complete-link (or "farthest neighbor") methods, Ward's (or, "minimum variance") methods, average-link (group average and weighted average), median-link, flexible, and so on. See, e.g, Anderberg (1973), Jain and Dubes (1988), Gower (1971), Symons (1981). Since these methodologies developed for classical data are based on a distance/dissimilarity matrix, they can be easily extended to the corresponding matrices calculated from symbolic data.

In particular, Figure 1 shows the resulting hierarchy when the average-linkage agglomerative method is applied to the Gowda-Diday matrix of (14). In contrast, Figure 2 shows the corresponding hierarchy when using the Euclidean Ichino-Yaguci matrix of (15). It is immediately seen that the hierarchies are different, with Napa and Humbolt counties appearing in different parts of the respective trees. This is a not-so-surprising result since it is well known that different distances can produce different tree structures. Likewise, using different agglomerative methods (e.g., complete-link instead of average-link) can produce different trees.

References

Anderberg, M. R. (1973) *Cluster Analysis for Applications.* Academic Press, New York.

Census (2000). *California:2000 Summary Population and Housing Characteristics.* U. S. Department of Commerce.

Chavent, M. (1998) "A monothetic clustering method," *Pattern Recognition Letters*, 19, 989-996.

DeCarvalho, F. A. T. (1994) "Proximity coefficients between boolean symbolic objects," in: (Diday, E., Lechevallier, Y., Schader, M., and Bertrand, P., eds), *New Ap- proaches in Classification and Data Analysis, Series: Studies in Classification, Data Analysis, and Knowledge Organization*, 387-394 Springer-Verlag, Berlin.

DeCarvalho, F. A. T. (1998) "Extension based proximity coefficients between con- strained boolean symbolic objects," in: (Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.-H., and Baba, Y., eds.) *Data Science, Classification, and Related Methods*, 370-378. Springer-Verlag, Berlin.

Diday, E. (1987) "Introduction a e'Approche Symbolique en Analyse des Donnees," *Premiere Jouneles Symbolique-Numerique*, CEREMADE, Université Paris, Dauphine, 21-56.

Gordon, A. D. (1999) *Classification* (2nd. ed.), Chapman and Hall, Boca Raton.

Gowda, K. C. and Diday, E. (1991) "Symbolic clustering using a new dissimilarity measure," *Pattern Recognition*, 24, 567-578.

Gower, J. C. (1971) "Statistical methods of comparing different multivariate analyses of the same data," in: *Anglo-Romanian Conference on Mathematics in Archeology and Historical Sciences* (eds. F. R. Hodson, D. G. Kendall and P. Tautu), Edinburgh University Press, Edinburgh, 138-149.

Hausdorff, F. (1937) *Set Theory* (translated into English by J. R. Aumann, 1957), Chelsey, New York.

Ichino, M. and Yaguchi, H. (1994) " Generalized Minkowski metrics for mixed feature type data analysis," *IEEE Transactions on Systems, Man and Cybernetics*, 24, 698-708.

Irpino, A. and Verde, R. (2006) "A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data," in: (Batagelj, V., Bock, H.-H., Ferligoj, A., and Ziberna, A., eds.) *Data Science and Classification*, 185-192. Springer-Verlag, Berlin.

Jain, A. K. and Dubes, R. C. (1988) *Algorithms for Clustering Data*, Prentice Hall, New Jersey.

Kim, J. (2009) *Dissimilarity Measures for Histogram-valued Data and Divisive Clustering of Symbolic Objects*, Doctoral Dissertation, University of Georgia.

Kim, J. and Billard, L. (2011) "A polythetic clustering process for symbolic observations and cluster validity indexes," *Computational Statistics and Data Analysis*, 55, 2250-2262.

Kim, J. and Billard, L. (2012) "Dissimilarity measures and divisive clustering for symbolic multimodal-valued data," *Computational Statistics and Data Analysis*, 56, 2795-2808.

Kim, J. and Billard L. (2013) "Dissimilarity measures for histogram-valued observations," *Communications in Statistics: Theory and Methods*, 42, 283-303.

Symons, M. J. (1981) "Clustering criteria and multivariate normal mixtures," *Biometrics*, 37, 35-43.

Ward, J. H. (1963) "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, 58, 236-244.

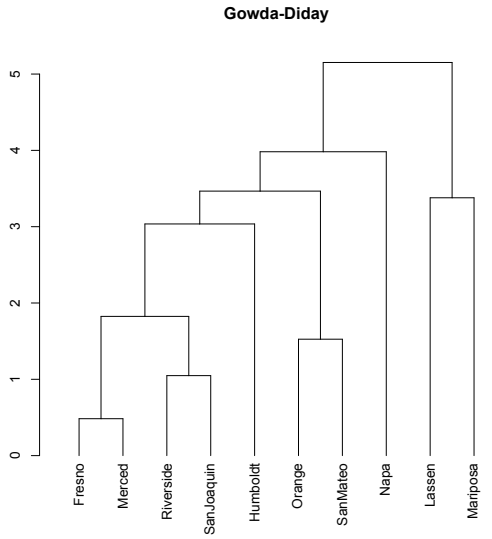

Figure 1 - Gowda-Diday Distances

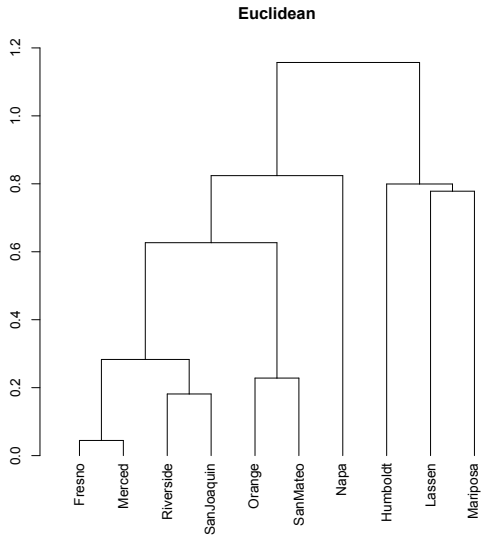

Figure 2 - Ichino-Yaguchi Distances