Clustering for Histogram-valued Observations - A Perspective

Billard, L.
University of Georgia, Department of Statistics
Athens Georgia 30602 USA
E-mail: lynne@stat.uga.edu

Kim, Jaejik
Georgia Health Sciences University, Department of Biostatistics
Augusta Georgia 30912 USA
E-mail: jaejik@gmail.com

1. Introduction

Contemporary computer capacity produces massively large datasets; yet those same computers can have difficulty analyzing these datasets because of their size. One way to handle this is to aggregate the data according to some meaningful scientific question(s). The resulting datasets are perforce symbolic-valued (such as intervals, histograms, multi-modal-valued), thus necessitating new methodology for their analyses. For example, a census collects information (e.g., age, income, housing costs, gender, race/ethnicity, etc.) on individuals. These individual values are aggregated according to some social-scientific question(s) of interest, such as by region, city, state, and so on. 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. Many such datasets in the form of population pyramids by country (e.g.) can be found at "census.gov/ipc/www/idb/informationGateway.php".

We focus on histogram-valued observations. We first consider some distance measures. By our count, there are about 15 such measures, most developed in the last year or two; see Section 3. Then, we consider monothetic and polythetic divisive clustering algorithms for clustering histograms in Section 4. The methodologies are illustrated through the well-known Fisher (1936) Iris dataset. This gives us a benchmark against which to compare our results using histogram-valued methods.

2. The Data

We have the random variable $\mathbf{Y} = (Y_1, \ldots, Y_p)$ taking values in $\mathcal{R}^p$. Suppose we have a sample of size n observations taking histogram values of the form

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

where the histogram Y_{uj} consists of s_{uj} subintervals $[b_{ujk}, b_{uj,k+1})$ occurring with relative frequency p_{ujk}. Typically, across observations and variables, the length and number of subintervals will vary. However, without loss of generality, we can transform the observations into histograms with common subintervals and the same number of subintervals. Therefore, for all terms in the right-side of (1), except for the relative frequency term p_{ujk}, the u subscript can be dropped; see Kim and Billard (2011a) for details.

The Fisher (1936) Iris dataset consists of 50 observations from each of three species (*setosa, versicolor, virginica*). There are $p = 4$ variables, viz., Y_1 = Sepal Length, Y_2 = Sepal Width, Y_3 = Petal Length, and Y_4= Petal Width. The observations were aggregated into seven groups of 20 and the eighth group of 10 observations; histograms were constructed for each variable and each group.

The resulting histograms for group 1 are provided in Table 1. Groups 1 and 2 comprised irises from the *setosa* species, groups 4 and 5 from the *versicolor* species and groups 6, 7 and 8 were from the *virginica* species. Group 3 consisted of 10 observations from each of the *setosa* and *versicolor* species. We know from classical studies of the 150 classical observations that observations from the *setosa* species are quite distinct from the other two species which overlap. It might be particularly interesting to see what happens with group 3.

Table 1 - Iris Histogram Observations

Group	Variable	Histogram: $\{[b_{jk}, b_{j,k+1}), p_{jk}; k = 1, \ldots, s_{uj}\}$
1	Y_1	{[4.2, 4.4), .10; [4.4, 4.6), .10; [4.6, 4.8), .15; [4.8, 5.0), .20; [5.0, 5.2), .15; [5.2, 5.4), .15; [5.4, 5.6). .00; [5.6, 5.8] .15}
	Y_2	{[2.8, 3.0), .20; [3.0, 3.2), .15; [3.2, 3.4) .15; [3.4, 3.6), .15; [3.6, 3.8), .15; [3.8, 4.0), .15; [4.0, 4.2) .00; [4.2, 4.4], .05}
	Y_3	{[1.1, 1.2), .10; [1.2, 1.3), .10; [1.3, 1.4) .35; [1.4, 1.5), .30; [1.5, 1.6), .05; [1.6, 1.7], .10}
	Y_4	{[0.10, 0.15), .15; [0.15, 0.20), .50; [0.20, 0.25), .00; [.25, .30), .20; [.30, .35), .00; [.35, .40], .15}
...	...	...

3. Distance-Dissimilarity Measures

As the name suggests, a dis/similarity or distance measure between two observations, u and v (say, $u, v = 1, \ldots, n$), describes the distance between them in some way; or equivalently describes how dis/similar they might be. A fundamental distance is the Minkowski distance of order q

$$(2) \quad d^{(q)}(u,\ v) = \left(\sum_{j=1}^{p} w_j [d_j(u,\ v)]^q \right)^{1/q}$$

where $d_j(u,\ v)$ is a distance measure between u and v for the variable Y_j and $w_j \geq 0$ is a weight associated with Y_j, $j = 1, \ldots, p$. When $q = 2$, we have the familiar Euclidean distance. Many different formulations exist for the distance between two classical observations, i.e., points in $\mathcal{R}^p$; see, e.g., Gordon (1999) for a comprehensive review.

Since histogram-valued observations are hypercubes in $\mathcal{R}^p$, any two observations can overlap. In order to obtain distance measures between these observations, we first need to define the sample mean and sample variances of an observation u, the union $u \cup v$ and intersection $u \cap v$ of the histograms u and v. Thus, for observations (1), we have that the sample means are, respectively, for Y_j, $j = 1, \ldots, p$,

$$(3) \qquad M_{uj} = \sum_{k=1}^{s_j} p_{ujk}(b_{ujk} + b_{uj,k+1})/2,$$

$$(4) \quad M_{(u\cup v)j} = \sum_{k=1}^{t_j} p^*_{(u\cup v)jk}(b_{jk} + b_{j,k+1})/2,\ p^*_{(u\cup v)jk} = \frac{p_{(u\cup v)jk}}{\sum_{k=1}^{s_j} p_{(u\cup v)jk}},\ p_{(u\cup v)jk} = \max\{p_{ujk}, p_{vjk}\},$$

$$(5) \quad M_{(u\cap v)j} = \sum_{k=1}^{s_j} p^*_{(u\cap v)jk}(b_{jk} + b_{j,k+1})/2,\ p^*_{(u\cap v)jk} = \frac{p_{(u\cap v)jk}}{\sum_{k=1}^{t_j} p_{(u\cap v)jk}},\ p_{(u\cap v)jk} = \min\{p_{ujk}, p_{pvjk}\};$$

and the sample variances for Y_j are, respectively,

$$(6) \qquad S^2_{uj} = \sum_{k=1}^{s_j} p_{ujk}[(b_{jk} - M_{uj})^2 + (b_{jk} - M_{uj})(b_{j,k+1} - M_{uj}) + (b_{j,k+1} - M_{uj})^2]/3,$$

$$(7) \qquad S^2_{(u\cup v)j} = \sum_{k=1}^{s_j} p_{(u\cup v)jk}[(b_{jk} - M_{\cup})^2 + (b_{jk} - M_{\cup})(b_{j,k+1} - M_{\cup}) + (b_{j,k+1} - M_{\cup})^2]/3,$$

$$(8) \qquad S^2_{(u\cap v)j} = \sum_{k=1}^{s_j} p_{(u\cap v)jk}[(b_{jk} - M_{\cap})^2 + (b_{jk} - M_{\cap})(b_{j,k+1} - M_{\cap}) + (b_{j,k+1} - M_{\cap})^2]/3,$$

with M_{uj}, $M_{\cup} \equiv M_{(u\cup v)j}$ and $M_{\cap} \equiv M_{(u\cap v)j}$ and $p_{(u\cup v)jk}$ and $p_{(u\cap v)jk}$ as given in (3)-(5).

The extended Gowda-Diday dissimilarity measure between histograms u and v for the variable Y_j, $d_j^{DG}(u,\ v)$, (obtained by extending the Gowda and Diday, 1991, result for intervals) is defined by

$$(9) \qquad d_j^{DG}(u,\ v) = \frac{|S_{uj} - S_{vj}|}{S_{uj} + S_{vj}} + \frac{S_{uj} + S_{vj} - 2S_{(u\cap v)j}}{S_{uj} + S_{vj}} + \frac{|M_{uj} - M_{vj}|}{\Psi_j}, \quad j = 1, \ldots, p,$$

with $M_{(.)}$ and $S_{(.)}$ as in (3)-(8) and $\Psi_j = b_{j,s_j+1} - b_{j1}$ as the span of the observed histograms for Y_j. Then, the extended Gowda-Diday dissimilarity measure for $\mathbf{Y}$ is obtained from $\sum_{j=1}^{p} d_j^{DG}(u, v)$.

In the interests of space limitations, we illustrate these distances for the observations $u = 1, 3, 5, 7$ only from the Iris dataset. Hence, the extended Gowda-Diday distance matrix, $\mathbf{D}_1^{GD}$, with elements $d_1^{GD}(u, v)$ obtained from (9), for Y_1, and the overall distance matrix, $\mathbf{D}^{GD}$, are, respectively,

$$(10) \qquad \mathbf{D}_1^{GD} = \begin{pmatrix} . & 0.81 & 0.71 & 1.44 \\ & . & 0.66 & 1.08 \\ & & . & 0.99 \end{pmatrix}, \quad \mathbf{D}^{GD} = \begin{pmatrix} . & 5.29 & 6.04 & 7.05 \\ & . & 4.51 & 5.34 \\ & & . & 4.18 \end{pmatrix}.$$

The extended Ichino-Yaguchi dissimilarity measure between histograms u and v for the variable Y_j, $d_j^{DG}(u,\ v)$, (by extending the Ichino and Yaguchi, 1994, result for intervals) is defined by

$$(11) \quad d_j^{IY}(u,\ v) = S_{(u\cup v)j} - S_{(u\cap v)j} + \gamma(2S_{(u\cap v)j} - S_{uj} - S_{vj}), \quad j = 1, \ldots, p,$$

for a preassigned constant $0 < \gamma < 0.5$ and with $S_{(.)}$ as defined in (6)-(8). Substituting $d_j^{IY}(u, v)$, $j = 1, \ldots, p$, into (2) with $q = 2$, we obtain the Euclidean extended Ichino-Yaguchi distance matrix, $\mathbf{D}^{IY}$, for the observations $u = 1, 3, 5, 7$, when $\gamma = 0.25$, as

$$(12) \quad \mathbf{D}^{IY} = \begin{pmatrix} . & 0.49 & 0.72 & 1.08 \\ & . & 0.39 & 0.73 \\ & & . & 0.44 \end{pmatrix}.$$

DeCarvalho (1994) proposed extensions of the Ichino-Yaguchi dissimilarity for intervals, by introducing several comparison functions involving agreement and disagreement indexes. These ideas can be expanded to the extended Ichino-Yaguchi dissimilarity measure of (11) for histograms. Hence, one such extended deCarvalho dissimilarity measure between histograms u and v, for Y_j, is given by,

$$(13) \quad d_j^{cf}(u, v) = 1 - cf_j = 1 - (\alpha_j - \delta_j)/(\alpha_j + \beta_j + \chi_j + \delta_j), \quad j = 1, \ldots, p,$$

where $\alpha_j = S_{(u\cap v)j}$, $\beta_j = S_{uj} - S_{(u\cap v)j}$, $\chi_j = S_{vj} - S_{(u\cap v)j}$ and $\delta_j = S_{(u\cup v)j} + S_{u\cap v)j} - S_{uj} - S_{vj}$.

Then, by substituting into (2) with $q = 2$, for the observations $u = 1, 3, 5, 7$, the Euclidean extended deCarvalho distance matrix, $\mathbf{D}^{cf}$, becomes

$$(14)\quad \mathbf{D}^{cf} = \begin{pmatrix} . & 1.67 & 2.69 & 2.82 \\ & . & 1.42 & 1.92 \\ & & . & 2.00 \end{pmatrix}.$$

Several other distance matrices have been derived in Kim and Billard (2011b), q.v. These include a cumulative distribution (cdf) based distance, as well as extending other distances from deCarvalho (1994, 1998) to histogram data. Irpino and Verde (2006) develops a type of Wasserstein distance using inverse cdf's. Kim (2009) reviews these among others. Any one of these distances can then be substituted into (2) to obtain relevant Minkowski distances.

4. Monothetic and Polythetic Clustering

There are many clustering procedures for classical data and for interval data; see, e.g., Gordon (1999) and Kim (2009) for a review. The current focus is on divisive clustering for histogram data. Kim and Billard (2011c,a), respectively, introduced a monothetic algorithm and a polythetic algorithm for histogram observations. While both methods use the dissimilarity matrices $\mathbf{D}$, they differ in the strategy used to find the optimal partition. The monothetic method considers the order of the mean values of the histograms for each variable (in one sense, it is an extention and adaptation of Chavent's, 1998, method for intervals). In contrast, the polythetic algorithm does not depend on the orders of single variables but uses all variables simultaneously. In particular, it starts with a 'seed' observation that is the farthest away from other observations (in terms of dissimilarity measures), and then iteratively determines if an observation is closer to, and therefore moved into, so-called 'splinter' or 'main' subclusters. More complete details and a comparison of both methods are in Kim (2009).

Suppose at the r^{th} stage, the original set of observations $\Omega \equiv P_1$ has been partitioned into r clusters, $P_r = \{C_1, \ldots, C_r\}$. Suppose cluster C_w contains n_w observations, $w = 1, \ldots, r$. The divisive algorithm then selects that cluster C_w to be partitioned into $C_w = (C_w^1, C_w^2)$ which minimizes the total within-cluster variation $W(P_r) = \sum_{w=1}^{r} I(C_w)$, or equivalently which maximizes

$$(15)\quad \Delta_w = I(C_w) - I(C_w^1) - I(C_w^2), \quad I(C_w) = \frac{1}{2\tau}\sum_{u=1}^{n_w}\sum_{v=1}^{n_w} g_u g_v [d(u,v)]^2, \quad w = 1, \ldots, r,$$

where $d(u, v)$ is a dissimilarity or distance measure (e.g., those considered in Section 3) between the observations u and v in the cluster C_w, g_u is the weight for observation u, and $\tau = \sum_{u=1}^{n_u} g_u$.

The monothetic and polythetic algorithms were applied to the Iris histogram data using all eight observations. The dendograms were somewhat similar for each algorithm though the validity indexes differed (not defined/shown herein; but see Kim and Billard, 2011a) with perturbations in the cluster structures being driven by the group 3 observation (which plays a role as a link between two species/clusters). We give the dendograms for the monothetic algorithm. Thus, Figure 1 shows the clusters that emerged when using the extended Gowda-Diday distance matrix. Figure 2 gives the results for the Euclidean distance matrix based on the Ichino-Yaguchi distances. The algorithm based on the Euclidean extended deCarvalho distance produced the dendogram of Figure 3.

It is immediately clear that the dendograms shown in the Figures 1 and 2 ultimately comprise the same clusters, although they arrive there by different routes and with different cutting variables.

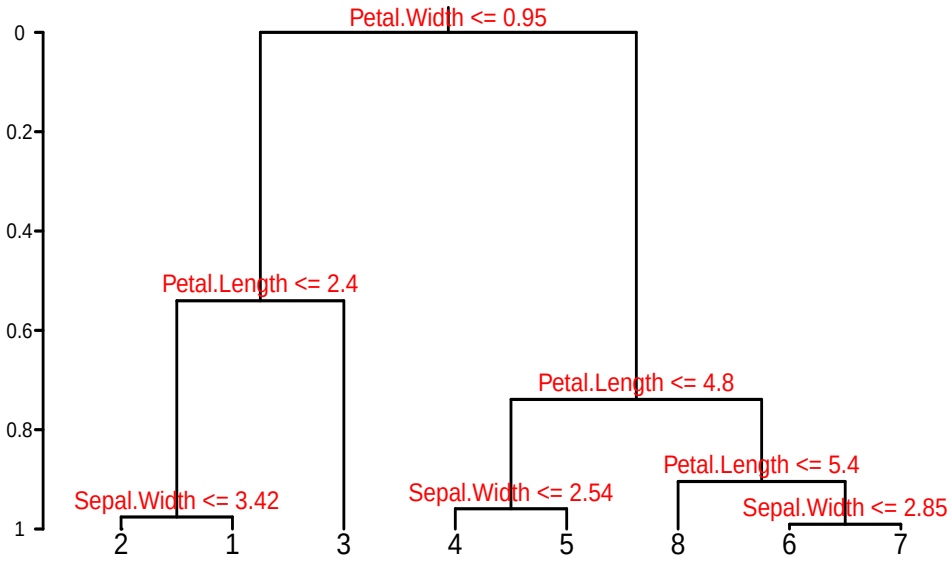

Figure 1 - Clusters based on extended Gowda-Diday Distances

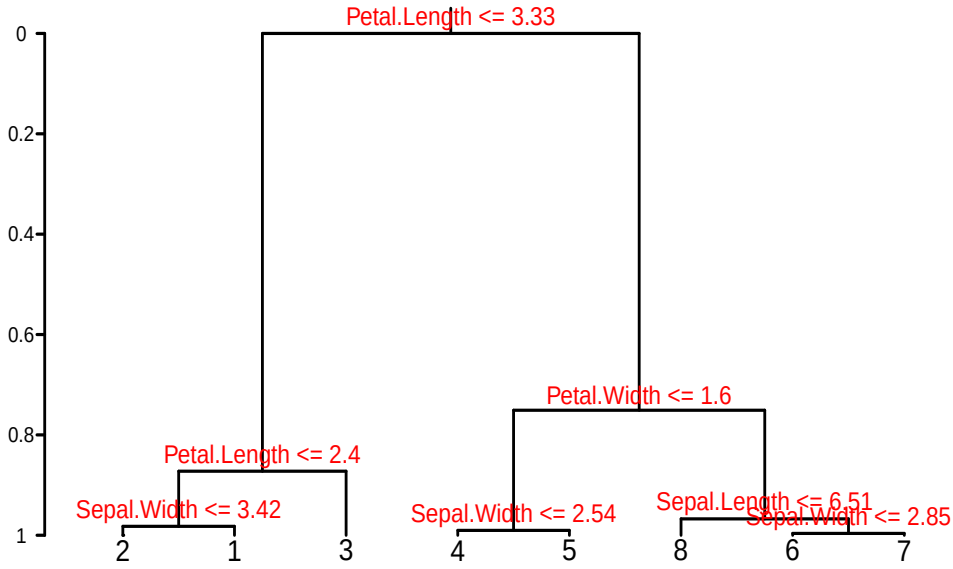

Figure 2 - Clusters based on Euclidean extended Ichino-Yaguchi Distances

For example, for Figure 1, the first cut variable is "Petal Width $= Y_4 \leq 0.95$", whereas for Figure 2, the first cut is made on "Petal Length $= Y_3 \leq 3.33$"; also, in Figure 1, the second stage partitions the subcluster with observations $C_1^1 = \{1, 2, 3\}$, while in Figure 2, the second stage partitions the subcluster of observations $C_1^2 = \{4, \ldots, 8\}$, and so on.

In contrast, Figure 3 is quite different primarily because of the group 3 observation, which immediately at the first partition is placed with the subcluster $C_1^2 = \{3, \ldots, 8\}$ and not the observations 1 and 2 as in the first two dendograms. Subsequent partitioning stages revolve around partitioning this larger subcluster C_1^2 before returning to the C_1^1 to partition it into its constituent components.

This distinction for the extended deCarvalho distances is a consequence of the fact that the this distance measure gives larger weight to overlapping areas compared to the extended Gowda-Diday and extended Ichino-Yaguchi distance measures.

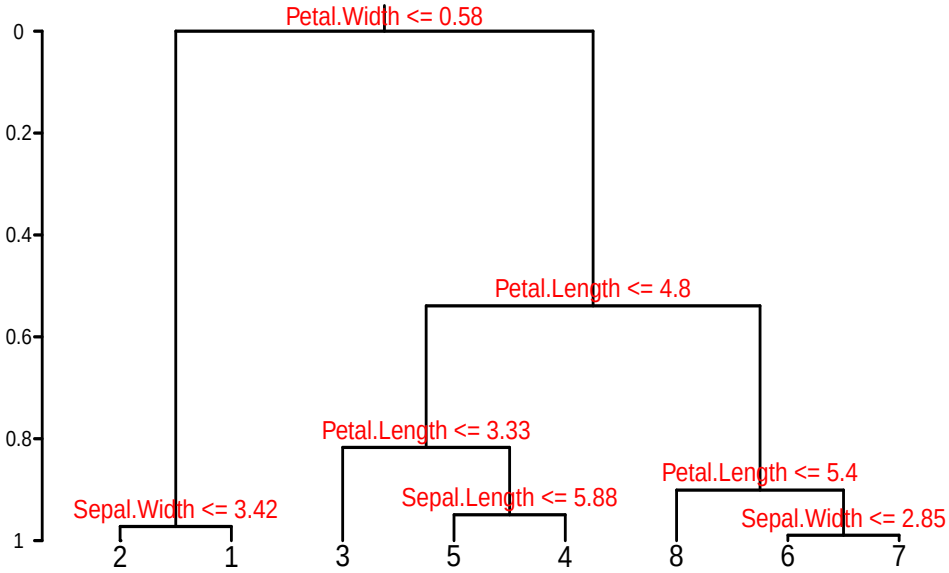

Figure 3 - Clusters based on Euclidean extended DeCarvaldo Distances

REFERENCES

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 Organisation*, 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.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179-188.

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

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

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.

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

Kim, J. and Billard L. (2011b). Dissimilarity measures for histogram-valued observations. *Communications in Statistics: Theory and Methods*, in press.

Kim, J. and Billard L. (2011c). Divisive clustering for histogram-valued data. Technical report.

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