1

Introduction

The theme of this volume centers on clustering methodologies for data which allow observations to be described by lists, intervals, histograms, and the like (referred to as "symbolic" data), instead of single point values (traditional "classical" data). Clustering techniques are frequent participants in exploratory data analyses when the goal is to elicit identifying classes in a data set. Often these classes are in and of themselves the goal of an analysis, but they can also become the starting point(s) of subsequent analyses. There are many texts available which focus on clustering for classically valued observations. This volume aims to provide one such outlet for symbolic data.

With the capabilities of the modern computer, large and extremely large data sets are becoming more routine. What is less routine is how to analyze these data. Data sets are becoming so large that even with the increased computational power of today, direct analyses through the myriad of classical procedures developed over the past century alone are not possible; for example, from Stirling's formula, the number of partitions of a data set of only 50 units is approximately 1.85×10^{47}. As a consequence, subsets of aggregated data are determined for subsequent analyses. Criteria for how and the directions taken in these aggregations would typically be driven by the underlying scientific questions pertaining to the nature and formation of the data sets at hand. Examples abound. Data streams may be aggregated into blocks of data or communications networks may have different patterns in phone usage across age groups and/or regions, studies of network traffic across different networks will inevitably involve symbolic data, satellite observations are aggregated into (smaller) sub-regional measurements, and so on. The list is endless. There are many different approaches and motivations behind the aggregations. The aggregated observations are perforce lists, intervals, histograms, etc., and as such are examples of symbolic data. Indeed, Schweizer (1984) anticipated this progress with his claim that "distributions are the numbers of the future".

In its purest, simplest form, symbolic data can be defined as taking values as hypercubes or as Cartesian products of distributions in p-dimensional space

Clustering Methodology for Symbolic Data, First Edition. Lynne Billard and Edwin Diday.

$\mathbb{R}^p$, in contrast to classical observations whose values are points in $\mathbb{R}^p$. Classical data are well known, being the currency of statistical analyses since the subject began. Symbolic data and their analyses are, however, relatively new and owe their origin to the seminal work of Diday (1987).

More specifically, observations may be multi-valued or lists (of categorical values). To illustrate, consider a text-mining document. The original database may consist of thousands or millions of text files characterized by a number (e.g., 6000) of key words. These words or sets of words can be aggregated into categories of words such as "themes" (e.g., telephone enquiries may be aggregated under categories of accounts, new accounts, discontinued service, broken lines, and so forth, with each of these consisting of its own sub-categories). Thus, a particular text message may contain the specific key words $Y = \{\text{home-phone, monthly contract, } \ldots\}$ from the list of possible key words $\mathcal{Y} = \{\text{two -party line, billing, local service, international calls, connections, home, monthly contract, } \ldots\}$. Or, the color of the bird species rainbow lorikeet is $Y = \{\text{green, yellow, red, blue}\}$ with Y taking values from the list of colors $\mathcal{Y} = \{\text{black, blue, brown, green, white, red, yellow, } \ldots, \text{(possible colors), } \ldots\}$. An aggregation of drivers by city census tract may produce a list of automobile ownership for one particular residential tract as $Y = \{\text{Ford, Renaullt, Volvo, Jeep}\}$ from $\mathcal{Y} = \{\ldots, \text{(possible car models), } \ldots\}$. As written, these are examples of non-modal observations. If the end user also wants to know proportional car ownership, say, then aggregation of the census tract classical observations might produce the modal list-valued observation $Y = \{\text{Holden, .2; Falcon, .25; Renault, .5; Volvo, .05}\}$ indicating 20% of the drivers own a Holden car, 50% own a Renault, and so forth.

Interval-valued observations, as the name suggests, are characterized as taking values across an interval $Y = [a, b]$ from $\mathcal{Y} \equiv \mathbb{R}$. There are endless examples. Stock prices have daily low and high values; temperatures have daily (or monthly, or yearly, …) minimum and maximum values. Observations within (or even between adjacent) pixels in a functional magnetic resonance imaging (fMRI) data set (from measurements of p different stimuli, say) are aggregated to produce a range of values across the separate pixels. In their study of face recognition features, Leroy et al. (1990) aggregated pixel values to obtain interval measurements. At the current time, more methodology is available for interval-valued data sets than for other types of symbolic observations, so special attention will be paid to these data.

Another frequently occurring type of symbolic data is the histogram-valued observation. These observations correspond to the traditional histogram that pertains when classical observations are summarized into a histogram format. For example, consider the height (Y) of high-school students. Rather than retain the values for each individual, a histogram is calculated to make an analysis of height characteristics of school students across the 1000 schools in the state. Thus, at a particular school, it may be that the heights, in inches,

are $Y = \{[50, 60), 0.12; [60, 65), 0.33; [65, 72), 0.45; [72, 80], 0.1\}$, where the relative frequency of students being 60–65 inches tall is 0.33 or 33%. More generally, rather than the sub-interval having a relative frequency, as in this example, other weights may pertain.

These lists, intervals, and histograms are just some of the many possible formats for symbolic data. Chapter 2 provides an introduction to symbolic data. A key question relates to how these data arrive in practice. Clearly, many symbolic data sets arise naturally, especially species data sets, such as the bird colors illustrated herein. However, most symbolic data sets will emerge from the aggregation of the massively large data sets generated by the modern computer. Accordingly, Chapter 2 looks briefly at this generation process. This chapter also considers the calculations of basic descriptive statistics, such as sample means, variances, covariances, and histograms, for symbolic data. It is noted that classical observations are special cases. However, it is also noted that symbolic data have internal variations, unlike classical data (for which this internal variation is zero). Bock and Diday (2000a), Billard and Diday (2003, 2006a), Diday and Noirhomme-Fraiture (2008), the reviews of Noirhomme-Fraiture and Brito (2011) and Diday (2016), and the non-technical introduction in Billard (2011) provide a wide coverage of symbolic data and some of the current methodologies.

As for classical statistics since the subject began, observations are realizations of some underlying random variable. Symbolic observations are also realizations of those same (standard, so to speak) random variables, the difference being that realizations are symbolic-valued instead of numerical or categorical point-valued. Thus, for example, the parameters of a distribution of the random variable, such as $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, are still points, e.g., $\boldsymbol{\mu} = (0, \ldots, 0)$ and $\boldsymbol{\Sigma} = \mathbf{I}$. This feature is especially evident when calculating descriptive statistics, e.g., the sample mean of interval observations (see section 2.4). That is, the output sample mean of intervals is a point, and is not an interval such as might be the case when interval arithmetic is employed. Indeed, as for classical statistics, standard classical arithmetic is in force (i.e., we do not use intervals or histograms or related arithmetics). In that same vein, aggregated observations are still distributed according to that underlying distribution (e.g., normally distributed); however, it is assumed that those normally distributed observations are uniformly spread across the interval, or sub-intervals for histogram valued data. Indeed, this is akin to the "group" data histogram problems of elementary applied statistics courses. While this uniform spread assumption exists in almost all symbolic data analytic procedures, relaxation to some other form of spread could be possible.

The starting premise of the clustering methodologies presupposes the data are already in a symbolic format, therefore the philosophical concepts involved behind the formation of symbolic data are by and large not included in this volume. The reader should be aware, however, that there are many issues that

might be considered between the initial data values (classical or otherwise) and the symbolic values of the data set to be analysed. This is particularly applicable when symbolic data emerge as a result of aggregating larger data sets. These principles are covered extensively in Bock and Diday (2000a), Billard and Diday (2006a), and Diday (2016).

Most clustering methodologies depend in some manner on dissimilarity and/or distance measures. The basic concepts underlying dissimilarity and/or distance measures are described in Chapter 3 along with some of their properties. Chapter 3 also presents dissimilarity/distance measures for non-modal symbolic data, i.e., for non-modal list multi-valued data and for interval-valued data. Chapter 4 considers such measures for modal observations, i.e., for modal list multi-valued data and for modal interval-valued (better known as histogram-valued) data. In most of the relevant literature, it is assumed that all variables are of the same type, e.g., all interval-valued. However, that is not always a necessary restriction. Therefore, the case of mixed type valued variables is illustrated on occasions, mainly in Chapters 6–8.

Chapter 5 reviews clustering procedures in general, with the primary focus on classical approaches. Clustering procedures are heavily computational and so started to emerge for classical data sets in the 1950s with the appearance of computers. Contemporary computers ensure these methods are even more accessible and even more in demand.

Broadly, clustering procedures can be categorized as organizing the entire data set Ω into non-overlapping but exhaustive partitions or into building hierarchical trees. The most frequent class of partitioning algorithm is the k-means algorithm or its variants usually based on cluster means or centroid values, including versions of the k-medoids algorithm which is typically based on dissimilarity or distance measures, and the more general dynamical partitioning method. Mixture distributions are also part of the partitioning paradigm.

There are two types of hierarchical tree constructions. The first approach is when the hierarchical tree is constructed from the top down divisively whereby the first cluster contains the entire data set Ω. At each step, a cluster is divided into two sub-clusters, with the division being dictated by some criteria, such as producing new clusters which attain a reduced sum of squares of the observations within clusters and/or between the clusters according to some collective measure of the cluster diagnostics. Alternatively, hierarchical trees can be built from the bottom up when the starting point consists of clusters of one only observation which are successively merged until reaching the top of the tree, with this tree-top cluster containing all observations in Ω. In this case, several criteria exist for the selection of which clusters are to be merged at each stage, e.g., nearest neighbor, farthest neighbor, Ward's minimum variance, among other criteria. An extension of the standard non-overlapping clusters of hierarchies is the agglomerative pyramidal methodology, which allows observations to belong to at most two distinct clusters.

These methods are extended to symbolic data in Chapter 6 for partitioning methods. Chapter 7 considers divisive hierarchies using either a monothetic algorithm or a polythetic algorithm. Because of the unique structure of symbolic data (e.g., they are not points), it becomes necessary to introduce new concepts (e.g., association measures) in order to develop algorithms for symbolic data. In Chapter 8, agglomerative methods are described for both hierarchies and pyramids. In each chapter, these constructions are illustrated for the different types of symbolic data: modal and non-modal list multi-valued data, interval-valued data, histogram-valued data and sometimes for mixed-valued data. As for classical methods, what becomes evident very quickly is that there is a plethora of available algorithms. These algorithms are in turn based on an extensive array of underlying criteria such as distance or dissimilarity matrices, with many different ways of calculating said matrices, and a further array of possible reallocation and starting and stopping rules.

At the end of a clustering process – be this a partitioning, a divisive hierarchy, an agglomerative hierarchy, or even a principal component analysis or aggregation of observations in some form – it is often the case that calculation of the cluster profile into a summarizing form is desired. Indeed, the field of symbolic data owes its origins to output data sets of clustering classical data when Diday (1987) recognised that summarizing obtained clusters by single point values involved a loss of critical details, especially a loss of information relating to the variation of the observations within a given cluster. This loss is especially significant if the clustering procedure is designed to produce outputs for future analyses. For example, suppose a process produces a cluster with two interval observations for a given variable of $Y = [1, 5]$ and $Y = [2, 8]$. One summary of these observations might be the interval $[1, 8]$ or the interval $[1.5, 6.5]$, among other possibilities. Chapters 5–8 contain numerous examples of output clusters. Rather than calculate a cluster representation value for each cluster for all these examples, the principle of this representation calculation is illustrated for some output clusters obtained in Chapter 6 (see section 6.7).

Likewise, by the same token, for all the hierarchy trees – those built by divisive algorithms or those built by agglomerative methods be these pure hierarchies or pyramidal hierarchies – tree heights (as measured along the y-axis) can be calculated. How this is done is illustrated for some trees in Chapter 7 (see section 7.4). An additional feature of Chapter 8 is the consideration of logical rules applied to a data set. This is particularly relevant when the data set being analysed was a result of aggregation of large classical data sets, though this can also be a factor in the aggregation of symbolic data sets. Thus, for example, apparent interval-valued observations may in fact be histogram-valued after appropriate logical rules are invoked (as in the data of Example 8.12).

All of these aspects can be applied to all examples in Chapters 5–8. We leave as exercises for the reader to establish output representations, tree heights, and rules where applicable for the data sets used and clusters obtained throughout.

Finally, all data sets used herein are available at <http://www.stat.uga.edu/faculty/LYNNE/Lynne.html>. The source reference will identify the table number where the data were first used. Sometimes the entire data set is used in the examples, sometimes only a portion is used. It is left as exercises for the reader to re-do those examples with the complete data sets and/or with different portions of them. Some algorithms used in symbolic analyses are contained in the SODAS (Symbolic Official Data Analysis System) package and can be downloaded from <www.ceremade.dauphine.fr/%7Etouati/sodas-pagegarde.htm>; an expanded package, SODAS2, can be downloaded from <http://www.assoproject.be>. Details of the use of these SODAS packages can be found in Bock and Diday (2000a) and Diday and Noirhomme-Fraiture (2008), respectively.

Many researchers in the field have indirectly contributed to this book through their published work. We hope we have done justice to those contributions. Of course, no book, most especially including this one, can provide an extensive detailed coverage of all applicable material; space limitations alone dictate that selections have of necessity had to be made.