Clustering Continuous or Categorical Data with the Forward Search

Cerioli, Andrea
Università di Parma, Dipartimento di Economia
Via J. Kennedy 6
43100 Parma, Italy
E-mail: andrea.cerioli@unipr.it

Riani, Marco
Università di Parma, Dipartimento di Economia
Via J. Kennedy 6
43100 Parma, Italy
E-mail: mriani@unipr.it

Atkinson, Anthony C.
The London School of Economics, Department of Statistics
Houghton Street
London WC2A 2AE, UK
E-mail: a.c.atkinson@lse.ac.uk

1. Introduction

The normal distribution, perhaps after data transformation, is the most widely used model for continuous multivariate data. Models for multivariate categorical data are more problematic. In this paper we give examples of the use of the forward search for clustering data that are either all continuous or all discrete.

For normal data our procedure uses a series of robust Mahalanobis distances and is related to outlier detection. Rousseeuw and van Zomeren (1990) and Rousseeuw and Van Driessen (1999) give examples of the calculation and use of robust Mahalanobis distances in the detection of outliers. The robust estimates of means and covariances that we use in the calculation of Mahalanobis distances in this paper come from the forward search, in which subsamples of increasing size are used for parameter estimation. The forward search in the analysis of multivariate data is given book-length treatment by Atkinson, Riani and Cerioli (2004), with material on transformations of multivariate data in Chapter 4.

Our analyses rely on forward plots of robust distances. In order to provide sensitive inferences about the existence of clusters it is necessary to augment such graphs with envelopes of the distributions of the statistics being plotted. Examples of such envelopes and their use in the forward search for clustering moderate sized data sets are presented by Atkinson and Riani (2006). Here we present the detailed analysis of a larger example.

Mahalanobis distances and the forward search are defined in §2, with the extended analysis of an example with continuous data in §3. Section 4 gives the details for clustering categorical variables, the main difference being the adoption of Euclidean instead of Mahalanobis distances. Application to data on soybeans and comparison with other clustering algorithms for categorical data are in §5.

2. Mahalanobis Distances and the Forward Search

The squared Mahalanobis distances for a sample of n v-dimensional normal observations are defined as

$$d_i^2 = \{y_i - \hat{\mu}\}^T \hat{\Sigma}^{-1} \{y_i - \hat{\mu}\}, , \qquad i = 1, \ldots, n, \tag{1}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are the unbiased moment estimators of the mean and covariance matrix of the n

observations and y_i is $v \times 1$.

In the forward search the parameters μ and Σ are estimated from a subset of m observations. From this subset we obtain n squared Mahalanobis distances

$$(2) \qquad d_i^2(m) = \{y_i - \hat{\mu}(m)\}^T \hat{\Sigma}^{-1}(m)\{y_i - \hat{\mu}(m)\}, \qquad i = 1, \ldots, n.$$

To start the search when the observations are assumed to come from a single multivariate normal population with some outliers, Atkinson, Riani and Cerioli (2004) pick a starting subset $S^*(m_0)$ that excludes any two-dimensional outliers. One search is run from this unique starting point, with successive searches using larger subsets of observations. When a subset $S(m)$ of m observations is used in fitting we order the squared distances and take the observations corresponding to the $m+1$ smallest as the new subset $S(m+1)$.

To detect outliers we examine the minimum Mahalanobis distance amongst observations not in the subset

$$(3) \qquad d_{\min}(m) = \min d_i(m) \quad i \notin S(m).$$

If this observation is an outlier relative to the other m observations, this distance will be "large" compared to the maximum Mahalanobis distance of observations in the subset. All other observations not in the subset will, by definition, have distances greater than $d_{\min}(m)$ and will therefore also be outliers.

For small datasets we can use envelopes from bootstrap simulations to determine the threshold of our statistic during the forward search. For moderate sized datasets we can instead use the polynomial approximations of Atkinson and Riani (2007). For large samples Atkinson, Riani and Laurini (2007) rescale a paradigmatic curve obtained by simulation to have the correct sample size and number of variables.

For cluster definition, as opposed to outlier identification, several searches are needed, the most informative being those that start in individual clusters and continue to add observations from the cluster until all observations in that cluster have been used in estimation. There is then a clear change in the Mahalanobis distances as units from other clusters enter the subset used for estimation. This strategy seemingly requires that we know the clusters, at least approximately, before running the searches. But we instead use many searches with random starting points to provide information on cluster existence and definition. Atkinson, Riani and Cerioli (2006) discuss random starts and argue that one set of simulation envelopes is appropriate whether the membership of $S(m_0)$ is robustly or randomly chosen.

3. Example: The Detection of Three Clusters

To illustrate our methodology we look at a large synthetic example with three clusters. We show how random start forward searches combined with envelope plots of forward Mahalanobis distances lead to the indication of clusters. We then interrogate the forward plots to obtain an initial definition of the clusters. We indicate how the information gained from these initial forward searches can be used as the basis of a definitive clustering. In the example we start with $m_0 = v + 1$, the smallest possible size and that which gives the highest probability of getting a subset consisting solely of observations from one cluster.

We analyse simulated data with three clusters of four-dimensional correlated observations, which is available at `www.riani.it/ISI07`. There are 1,000 observations in all. The scatter plot of y_1 against y_2 is in the left-hand panel of Figure 1, with the same symbol for each group. In this view the data seem to be a single homogeneous sample. The other panels of the scatterplot matrix are similar and so are not given here.

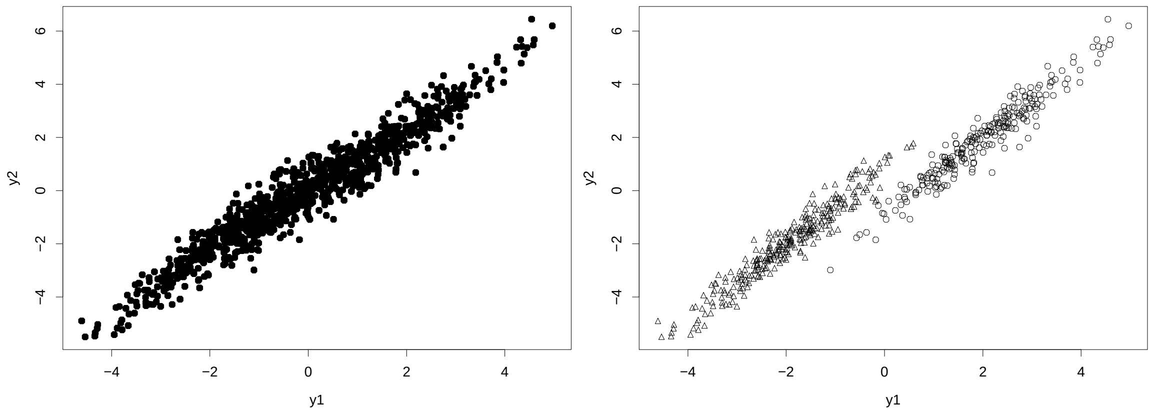

Figure 1: Three clusters of correlated normal variables; scatterplots of y_2 against y_1. Left-hand panel, all three groups; right-hand panel, Groups 1 and 2

There are 250 observations in the first group, 300 in the second and 450 in the third. The right-hand panel of Figure 1 repeats the scatter plot matrix of the left-hand panel with the third group omitted. It is clear that the observations in each group are highly correlated and that Group 3 lies between the other two. Of course, in our analysis we ignore the information about group structure, or even about the number of groups.

We now run 200 random start forward searches. The resulting forward plot of minimum distances is in Figure 2.

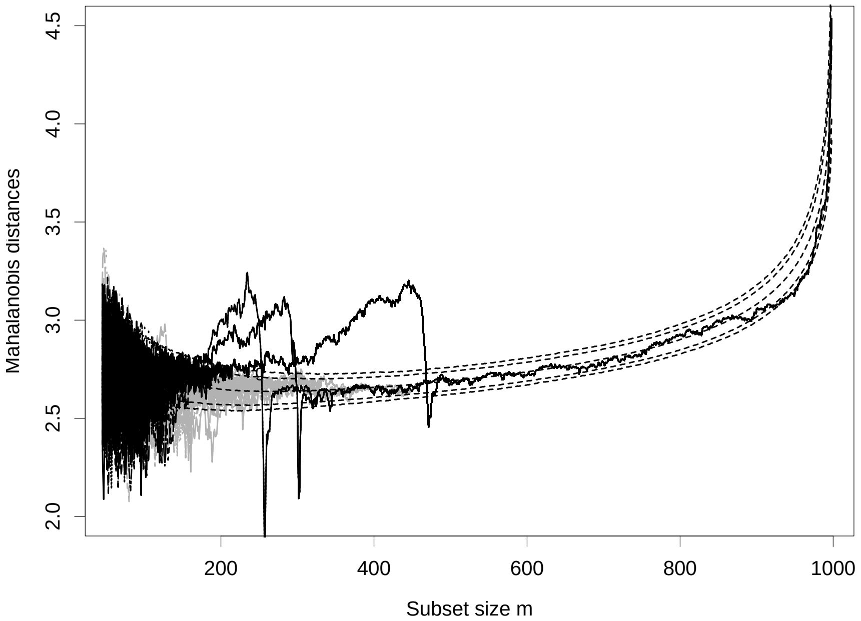

Figure 2: Three clusters of correlated normal variables: forward plot of minimum Mahalanobis distances from 200 random starts with 1%, 5%, 50%, 95% and 99% envelopes. Three clusters are evident. Trajectories in grey always include units from both groups

The forward searches fall into four classes: those that start in each of the three groups and those that, from the beginning of the search, include observations from at least two groups. These are shown in grey on the plot. From around $m = 150$ the searches with observations from only one group start to lie outside the envelopes. In general, the plots and envelopes for smaller n and given m rise above those with larger n as m increases. The curves here are behaving much as those from samples of slightly less than the known but ignored population values of 250, 300 and 450. However, at around

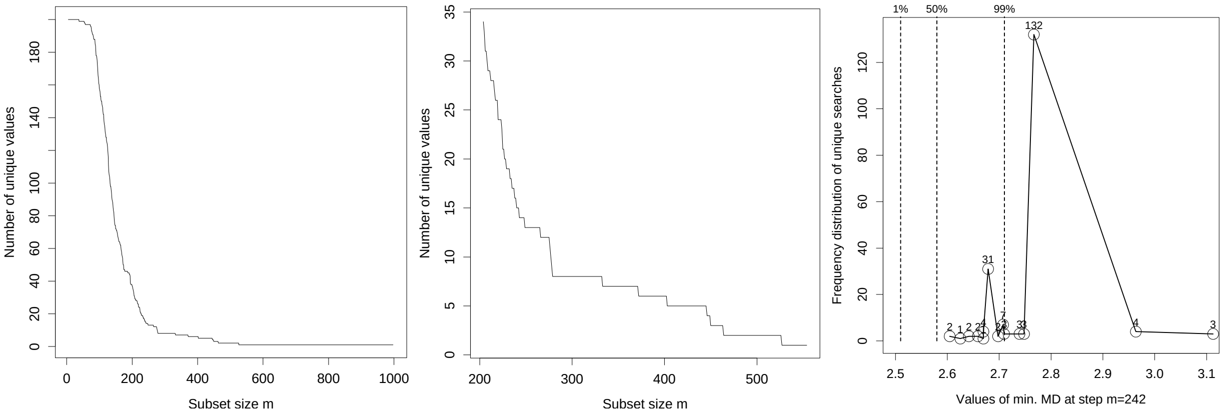

Figure 3: Three clusters of correlated normal variables. Left-hand panel: forward plot of number of unique Mahalanobis distances from 200 random starts; central panel, zoom of plot where clusters become apparent. Right-hand panel: frequency distribution of values of $d_{\min}(242)$

these values the curves suddenly dip below the envelopes as relatively remote observations from the other groups enter the subset used in fitting. From a little after $m = 500$ there is a single forward plot, in which a common mean and common covariance matrix are calculated from the observations in more than one group. As a result the estimated covariance matrix is slightly too large and the plotted distances are slightly small; they lie rather low in the envelope in Figure 2, the effect being most noticeable around $m = 900$. The approximate values of m at the three peaks (and total number in the group) are: 230 (250), 290 (300) and 450 (450). Despite the unpromising appearance of the scatter plots of Figure 1 our method has initially indicated clusters for 97% of the observations.

We now move from Figure 2 to cluster membership. Figure 2 shows that, in the second half of the search, all 200 searches have converged in the sense that, for each m, there is one common set of observations $S(m)$ and one value of $d_{\min}(m)$. Once two searches have the same value of $S(m)$ they will continue to follow the same path, wherever they started from, producing identical values of $d_{\min}(m)$. The left-hand panel of Figure 3 shows how the number of different values of $d_{\min}(m)$ decreases with m. Up to $m = 37$ there are 200 trajectories. The number of trajectories then decreases rapidly, reaching 1 at $m = 527$. The central panel is a zoom of the central part of this plot.

We are interested in the subsets $S(m)$ for those trajectories where there is evidence of a cluster structure. From Figure 2 this is a little less than $m = 250$. To find the clusters we interrogate Figure 2 near this value to find the subsets giving rise to the larger values of $d_{\min}(m)$ in the centre of the search.

The right-hand panel of Figure 3 is a plot of the frequency distribution of the values of $d_{\min}(242)$. There are two large values, which are achieved by respectively 3 and 4 searches; the third largest value is appreciably smaller, but already contains 132 trajectories. Figure 2 shows that these trajectories continue to give the third peak around $m = 450$. The membership of the subsets for these three groups can be illustrated using 'entry' plots.

The entry plot is a way of representing the membership of $S(m)$ as a search progresses. For each m those observations included in the subset are plotted with a heavy symbol, so that the plot becomes darker as m increases. Such plots for a single search are discussed in Atkinson et al. (2004, §7.3.3). Here we need to present information from several searches.

The left-hand panel of Figure 4 is the entry plot for one of the 3 searches with the most extreme value of $d_{\min}(242)$ in Figure 3. The plot has been truncated at $m = 600$ since, by this point, units from all clusters are in the subset and the plot provides no further information on clustering. Since the searches have converged at $m = 242$, all will have the same trajectory thereafter, so it does not matter which of the 3 we choose to plot and we selected randomly from one of these 3 searches to obtain a typical plot. The panel shows clearly that we have found the trajectories that include observations

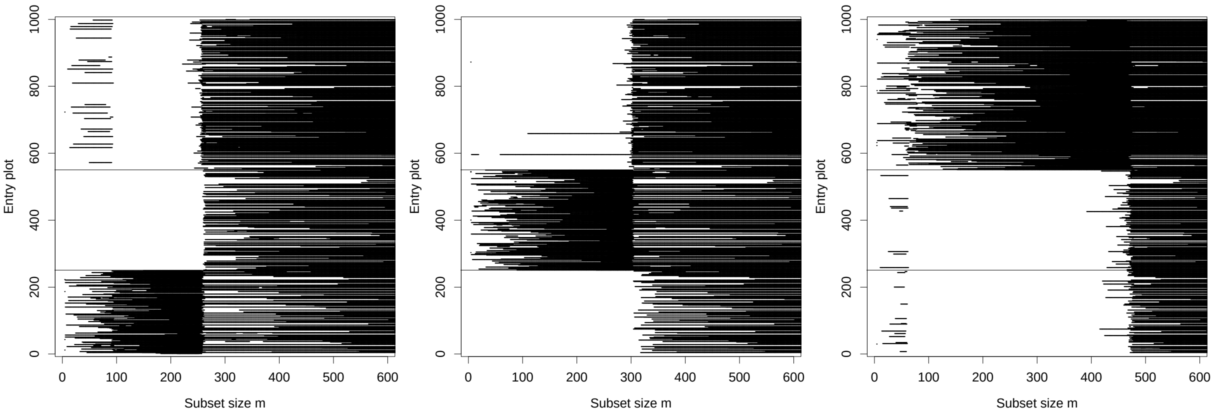

Figure 4: Three clusters of correlated normal variables; entry plots. Left-hand panel, the highest value of $d_{\min}(242)$ (3 trajectories) in Figure 2; centre panel, the second-highest value of $d_{\min}(242)$ (4 trajectories) and, right-hand panel, the next highest value of $d_{\min}(242)$ (132 trajectories) in Figure 2

from those numbered 1 - 250. Interestingly, the plot shows that up to $m = 100$ this particular search also included a number of units from the intermediate Group 3. The centre panel of Figure 4 shows the entry plot for one of the four trajectories corresponding to the second highest value of $d_{\min}(242)$ in Figure 3. Here we have detected the units numbered 251-550 belonging to the second cluster; two of units, 597 and 659, from the intermediate Group 3 have also been included in the subset. Likewise the right-hand panel reveals the third cluster of observations numbered 551 - 1,000.

The results shown in Figure 4 clearly indicate how well our cluster identification procedure works. Since the clusters correspond to the natural order of the units, we are able to check that our method has achieved what is required. But permutation of the labels of the units, corresponding to not knowing the cluster membership, makes the plots noninformative.

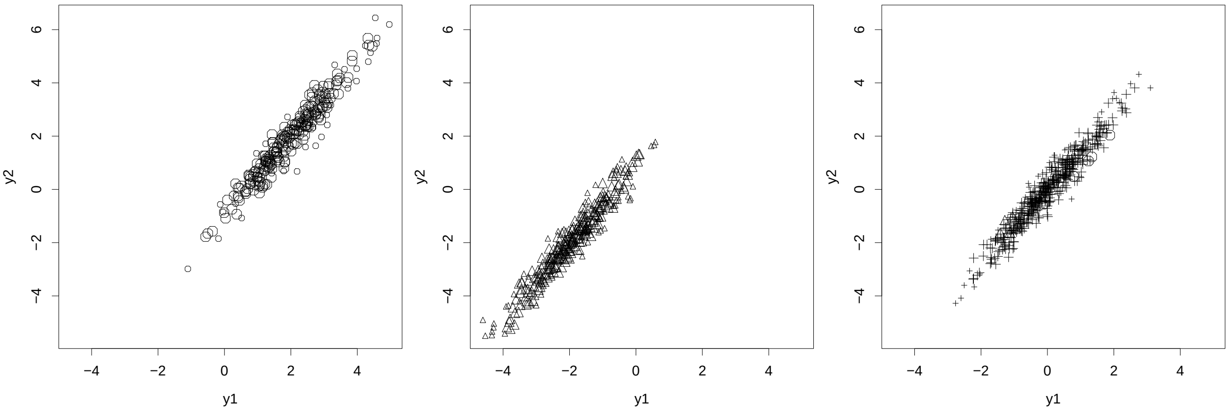

Figure 5: Three clusters of correlated normal variables. Three tentative clusters; scatterplots of y_2 against y_1; small symbols, units not included. Right-hand panel, the intermediate cluster 2 includes 13 units from other groups

A label-free indication of how well our clustering procedure has done can be obtained by looking once more at the data. The three panels of Figure 5 repeat the scatterplots of Figure 1 for the three groups of observations identified by the three peaks in Figure 2. In each panel we give the units in the cluster with the same symbols as before. Units from the group that are not included in the cluster are shown with small symbols.

Cluster 1 includes 218 of the 250 units of Group 1 and no other units. Cluster 2 contains 233 of the 300 units in Group 2 and the two units 597 and 659 from Group 3 that we saw in Figure 4. The intermediate Group 3 is the most misclassified; 420 of the 450 units are clustered, but together with six from Group 1 and seven from Group 2. The slightly small size of the clusters is due to the overlapping nature of the groups.

This forward process has tentatively identified three clusters. One way to confirm this identification is to run searches on individual clusters. If the peak for a particular cluster in the forward plot analogous to Figure 2 occurs when $m = n_c$ we include the next few units to enter and then run a search on these n_c^+ units, superimposing envelopes for a sample of size n_c^+. If no outliers are found, we have a homogenous cluster and increment n_c^+ to check whether we have failed to include some units. If outliers are detected, we delete the last observation to enter, reduce the sample size by one and superimpose envelopes for this reduced sample size. Eventually we obtain the largest group of homogenous observations containing no outliers. Examples of this procedure are given by Atkinson and Riani (2007). Atkinson, Riani and Cerioli (2004, Cap. 7) describe a combined forward search in which each unit is free to enter any one of the, here three, clusters. However, their procedure does not use envelopes for testing the cluster allocations. Atkinson and Riani (2007) give comparisons with other clustering methods for smaller data sets. Here k-means completely fails.

4. The Forward Search for Categorical Variables

In this section we show how the random-start forward search algorithm can overcome in a robust way most of the shortcomings of clustering methods for categorical data. As we have seen, in a clustering context unusual observations can either be multivariate outliers or inliers that do not belong to any of the main groups. The latter typology occurs frequently and is of particular concern with categorical data. Definition (1) is clearly meaningless when we have a sample of n v-dimensional 'labels'. However, we can use dummy coding to process this categorical information. Write $y_{ij} \in \mathcal{C}^{(j)}$ for the observed class of variable j in unit i, where $\mathcal{C}^{(j)}$ is the set of possible classes for variable j. The number of such classes is c_j. Let $y_{ij}^{(c)} = 1$ if $y_{ij} = c$ and $y_{ij}^{(c)} = 0$ otherwise. Unit i is now represented by the $C \times 1$ vector $y_i = [x_{i1}^{(1)}, \ldots, x_{i1}^{(c_1)}, \ldots, x_{iv}^{(1)}, \ldots, x_{iv}^{(c_v)}]'$, with $C = \sum_{j=1}^{v} c_j$.

We make the mild assumption that y_i is a random observation from a population with class probabilities $\mu = [\mu_1^{(1)}, \ldots, \mu_1^{(c_1)}, \ldots, \mu_v^{(1)}, \ldots, \mu_v^{(c_v)}]'$, so that $\mathrm{E}(y_{ij}^{(c)}) = \mu_j^{(c)}$, for $c \in \mathcal{C}^{(j)}$ and $j = 1, \ldots, v$. Computation of a fully-parametrized Mahalanobis distance like (1) has two main shortcomings in the present context. First, second-order moments are not able to capture all the dependence structure among indicator variables. Second, $\hat{\Sigma}$ is now of dimension $C \times C$, with C typically much larger than v. A much simpler option is then the squared Euclidean distance

$$d_i = d(y_i, \hat{\mu}) = \sum_{j=1}^{v} \sum_{c \in \mathcal{C}^{(j)}} (y_{ij}^{(c)} - \hat{\mu}_j^{(c)})^2. \tag{4}$$

It is easily seen that (4) is equivalent to the simple matching dissimilarity index when computed on pais of observations. Furthermore, it can easily be generalized to encompass differential weighting and correlation among the classes of the same variable (Kurczynski, 1970; Friedman and Meulman, 2004).

The forward search for categorical variables is run following the same rules outlined in §2. At each step m we compute the n dissimilarities

$$d_i(m) = d(u_i, \hat{\mu}(m)) = \sum_{j=1}^{v} \sum_{c \in \mathcal{C}^{(j)}} (x_{ij}^{(c)} - \hat{\mu}_j^{(c)}(m))^2 \qquad i = 1, \ldots, n, \tag{5}$$

which are used both for progressing and for monitoring the search. We start with a randomly selected subset of $m_0 = 2$ observations, to enhance the probability of starting from only one cluster. To detect potential clusters, we again look at the forward plot of $d_{\min}(m)$. Other valuable tools for detecting important cluster features include the forward plots of individual dissimilarities $d_i(m)$, $i = 1, \ldots, n$, the entry plot, and the forward plots of sample proportions $\hat{\mu}_j^{(c)}(m)$, which can be useful for the interpretation of how clusters evolve along the search.

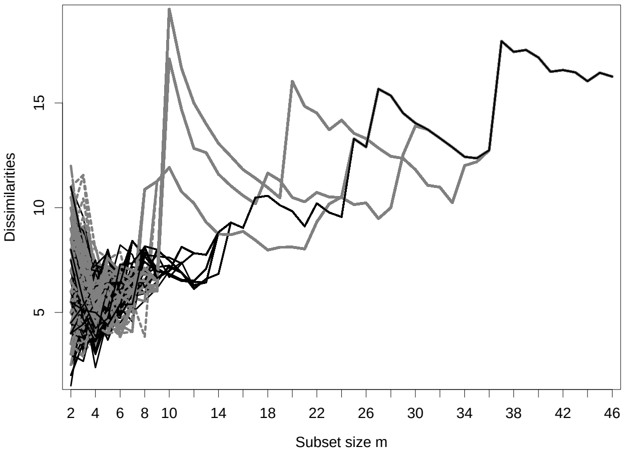

Figure 6: Soybean data. Forward plots of $d_{\min}(m)$ for 1000 searches with random starting points.

5. Comparison with the k-modes algorithm: the soybean data set

Some simulation evidence of the good performance of the forward search algorithm for clustering categorical observations is provided by Cerioli, Riani and Atkinson (2006). Here we compare the forward search with the k-modes algorithm on a data set that has frequently been used to test clustering algorithms for categorical data. We analyze the soybean data set, available from the UCI Machine Learning Repository. It has 47 observations on 21 categorical variables. Each observation is labelled as one of four soybean diseases: Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot, and Phytophthora Rot. Except for the last disease, which has 17 observations, all other diseases have 10 instances each. We discard the disease information for clustering the soybean data set. Huang (1998) reports the results for the k-modes algorithm. This technique is sensitive to the order in which the units are listed, so repeated application to random permutations of the original records is required. Even in the unrealistic situation where k is known, the algorithm is able to attain complete recovery of the four disease classes in only 14% of the permutations of the soybean data set. In 36% of these permutations the result is less than good, with more than 6 misclassified units.

Figure 6 contains the results of 1000 forward searches from randomly selected starting subsets. For each search the dissimilarity $d_{\min}(m)$ is plotted. Most of the features of this plot are similar to those depicted in Figure 2. From $m = 15$ onwards, the searches follow only four different trajectories, regardless of their starting point. The four distinct trajectories provide clear evidence about the group structure, information that is not available from the k-modes algorithm. The three peaks at $m = 10$, pictured in grey in Figure 6, are very pronounced and correspond to well separated clusters of the same size. The fourth trajectory, shown in boldface in Figure 6, increases at a more stable pace, an evidence of a cluster without sharp boundaries. The largest peak at $m = 10$ is for searches including units from the second group (Charcoal Rot), numbered from 11 to 20 in the data set. Figure 7 depicts the structure emerging from the viewpoint provided by this cluster. The left-hand panel shows the forward plots of individual dissimilarities. The effect of adding units from a different group to a homogeneous subset is evident just after $m = 10$, when the first observation from the Diaporthe Stem Canker group enters into $S(m)$. The right-hand panel is the corresponding entry plot for $m = 2, \ldots, 15$.

The general structure of the data is revealed by the different trajectory shapes in the left-hand panel of Figure 7. A more accurate dissection can be obtained by looking at the searches that give rise to the other peaks in Figure 6. Proceeding as in §3, we find that the second largest peak at $m = 10$ is for the units in the Diaporthe Stem Canker group, while the third largest peak at $m = 10$ identifies the observations with the Rhizoctonia Root Rot disease. The fourth disease corresponds to the remaining

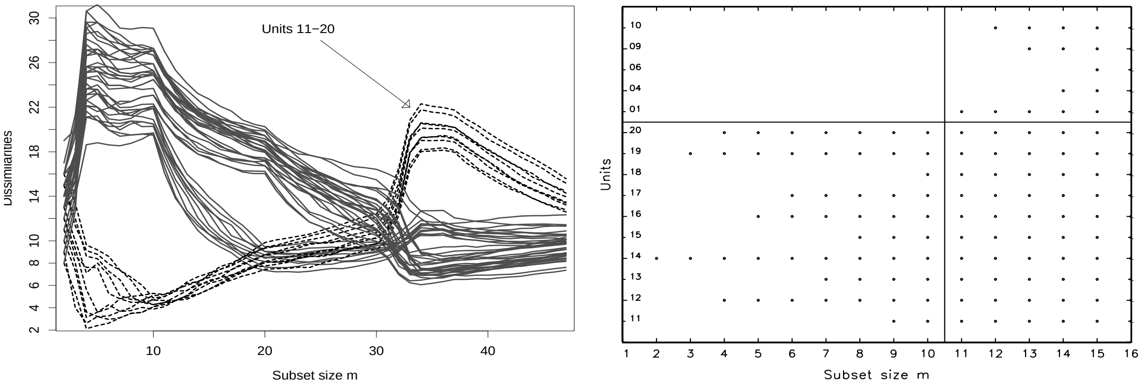

Figure 7: Soybean data: searches that yield the largest peak in $d_{\min}(m)$ at $m = 10$. Left-hand panel: forward plots of individual dissimilarities; right-hand panel: entry plot for $m = 2, \ldots, 15$.

stable trajectory in Figure 6. This trajectory has its first peak at $m = 18$, although $d_{\min}(17)$ is only slightly smaller. The corresponding units in $S(18)$ belong to the Phytophthora Rot group, with the exception of unit 28, which is the last to join the subset. Observation 28 also belongs to the subset originating one of the peaks at $m = 10$, so it may be viewed as a "borderline" unit and need not be firmly classified, at least at an exploratory stage. We conclude that also in this example our forward process has identified the number of clusters and has correctly allocated all but one observation, for which classification is uncertain.

REFERENCES

Atkinson, A. C. and M. Riani (2006). Distribution theory and simulations for tests of outliers in regression. *Journal of Computational and Graphical Statistics 15*, 460–476.

Atkinson, A. C. and M. Riani (2007). Exploratory tools for clustering multivariate data. *Computational Statistics and Data Analysis*. doi:10.1016/j.csda.2006.12.034.

Atkinson, A. C., M. Riani, and A. Cerioli (2004). *Exploring Multivariate Data with the Forward Search*. New York: Springer–Verlag.

Atkinson, A. C., M. Riani, and A. Cerioli (2006). Random start forward searches with envelopes for detecting clusters in multivariate data. In S. Zani, A. Cerioli, M. Riani, and M. Vichi (Eds.), *Data Analysis, Classification and the Forward Search*, pp. 163–171. Berlin: Springer-Verlag.

Atkinson, A. C., M. Riani, and F. Laurini (2007). Approximate envelopes for finding an unknown number of multivariate outliers in large data sets. (Submitted),.

Cerioli, A., M. Riani, and A. C. Atkinson (2006). Robust classification with categorical variables. In A. Rizzi, and M. Vichi (Eds.),*COMPSTAT 2006*, pp. 507–519. Heidelberg: Physica-Verlag.

Friedman, J. H. and J. J. Meulman (2004). Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society B 66*, 815–849.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery 2*, 283–304.

Kurczynski, T. W. (1970). Generalized distance and discrete variables. *Biometrics 26*, 525–534.

Rousseeuw, P. J. and K. Van Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41*, 212–223.

Rousseeuw, P. J. and B. C. van Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association 85*, 633–9.