Cluster analysis

Clustering variables, assessing clustering algorithms and biclustering

Miguel Rodo

2024-02-12

Outline

- Clustering variables using ClustOfVar
- Clustering observationals and variables using biclustering

M. Chavent et al. (2011)

Journal of Statistical Software

September 2012, Volume 50, Issue 13. *http://www.jstatsoft.org/*

ClustOfVar: An R Package for the Clustering of Variables

Marie Chavent
University of Bordeaux

Vanessa Kuentz-Simonet
Irstea

Benoît Liquet
University of Bordeaux

Jérôme Saracco
University of Bordeaux

- The algorithm `ClustOfVar` clusters a mix of quantitative and qualitative variables

Synthetic variables

- Clustering may be both hierarchical and non-hierarchical (k-means)
- The key novelty is the introduction of a synthetic variable for each cluster, used to guide clustering
  - For a given cluster, the synthetic variable is the first principal component of the variables in the cluster
  - As the data may have qualitative and quantitative variables, the algorithm PCAMix (Marie Chavent, Kuentz-Simonet, and Saracco 2012) is used.

Homogeneity

- The synthetic variable is used to define the homogeneity ("togetherness") of a cluster
- Definition of homogeneity:
  - Sum of R^2 between the synthetic variable and each variable in the cluster
  - R^2 is the sum proportion of variation in the dependent variable (the synthetic variable in this case) explained by the independent variables (the variables in the cluster)

Hierarchical clustering algorithm

- An agglomerative hierarchical clustering procedure is employed
- The choice of which two clusters to merge is based on the homogeneity of the original and resulting clusters. This is the only novelty.
- Specifically, for $H(\cdot)$ the homogeneity of a given cluster and A and B two clusters, the algorithm merges two clusters such that $d(A, B) = H(A) + H(B) - H(A \cup B)$ is minimised.

Partitioning algorithm

- As with hierarchical clustering, the partitioning algorithm uses the synthetic variables to guide the clustering.
- For a given cluster with synthetic variable, the association with an actual variable is measured by the canonical correlation coefficient
  - As we are only considering the first canonical variate and the synthetic variable is quantitative (not categorical), this is equal to the R^2 of a linear regression of the synthetic variable on the actual varialble
- As before, variables are allocated to clusters for which the dissimilarity is minimised (canonical correlation with synthetic variable is maximised).

Choosing the number of clusters

Cluster stability

- Assessed using cluster stability under resampling
- Essentially, this is the procedure:
  - Bootstrap B samples of the n observations
  - Apply the clustering algorithm to each bootstrap sample
  - Calculate the Rand index (Rand 1971) between the clusters obtained from the original sample and the clusters obtained from the bootstrap sample
- The average Rand index is the cluster stability.

Rand index

- The Rand index is a rather odd measure to assess the similarity of the two clusters.
  - For a given pair observations, they are regarded as having been clustered the same way if they either are clustered together in both clusterings or are clustered separately in both clusterings.
  - The Rand index is the proportion of pairs of observations that are clustered the same way in both clusterings.
- The adjusted Rand index (Hubert and Arabie 1985) corrects for chance agreement.

Assessment of `ClustOfVar`

- Essentially performing regression or the SVD each time the dissimilarity needs to be calculated is time-consuming.
  - They note that performance is slow when there are many variables.
  - At least at the time of writing, a parallel version of the algorithm was planned.
  - Looking at their GitHub repository, this does not seem to have been done.
    - Package is on CRAN, but has not been updated in years.
- Good that they made an attempt to help guide the number of clusters selected
  - Unusual that they ignored per-cluster stability assessments (Hennig 2007)
- Appropriateness in particular domains would need to be asssessed (i.e. how does it do on particular kinds of datasets, e.g. economic, biological)

Selection of clustering algorithms

- Part of the purpose of showing `ClustOfVar` is to show the flexibility in coming up with new clustering algorithms
  - The first set of slides introduce basic clustering approaches
  - These may be combined with other techniques (such as PCAMix) to create new algorithms
- The problme afterwards is - which algorithm to use?
- Typically, there are two main considerations:
  - Theoeretical considerations
    - For example: computational complexity (affecting run time, memory constraints), assumptions about the data (e.g. normality), whether the number of clusters is pre-specified, etc.
  - Empirical considerations
    - Performance in real world datasets: correspondence with manual labels, stability, ability to identify predictive variables, etc.

Labelling cells

- Modern experimental techniques measure tens of variables on millions of cells rapidly
- The cells need to be labelled, e.g. as T cells, B cells, etc., which first requires clustering them.
- Traditionally, this was done by hand (as below), but this is very slow at scale:

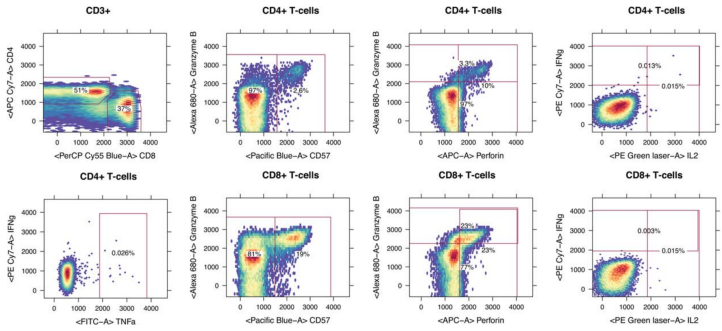

Figure 1: Finak (2014)

Aghaeepour et al. (2013)

- Whilst manual clustering of cells is slow, it is relatively trusted.
- It was not clear how well automated algorithms would perform, in terms of reproducing manual clusterings or identifying biologically relevant clusters.
- Aghaeepour et al. (2013) therefore constructed an empirical comparison, assessing algorithm performance on multiple datasets in terms of the following criteria:
  - Ability to reproduce manual clusterings
  - Ability to identify cell types associated with disease

Several automated algorithms typically identified manual clusters well

	F-measure[a]							
	GvHD	DLBCL	HSCT	WNV	ND	Mean	Runtime h:mm:ss[b]	Rank score[c]
Challenge 1: completely automated								
ADICyt	**0.81 (0.72, 0.88)**	**0.93 (0.91, 0.95)**	**0.93 (0.90, 0.96)**	**0.86 (0.84, 0.87)**	**0.92 (0.92, 0.93)**	0.89	4:50:37	52
flowMeans	**0.88 (0.82, 0.93)**	**0.92 (0.89, 0.95)**	**0.92 (0.90, 0.94)**	**0.88 (0.86, 0.90)**	0.85 (0.76, 0.92)	0.89	0:02:18	49
FLOCK	**0.84 (0.76, 0.90)**	**0.88 (0.85, 0.91)**	0.86 (0.83, 0.89)	**0.83 (0.80, 0.86)**	0.91 (0.89, 0.92)	0.86	0:00:20	45
FLAME	**0.85 (0.77, 0.91)**	**0.91 (0.88, 0.93)**	**0.94 (0.92, 0.95)**	0.80 (0.76, 0.84)	0.90 (0.89, 0.90)	0.88	0:04:20	44
SamSPECTRAL	**0.87 (0.81, 0.93)**	0.86 (0.82, 0.90)	0.85 (0.82, 0.88)	0.75 (0.60, 0.85)	**0.92 (0.92, 0.93)**	0.85	0:03:51	39
MMPCA	**0.84 (0.74, 0.93)**	0.85 (0.82, 0.88)	**0.91 (0.88, 0.94)**	0.64 (0.51, 0.71)	0.76 (0.75, 0.77)	0.80	0:00:03	29
FlowVB	**0.85 (0.79, 0.91)**	0.87 (0.85, 0.90)	0.75 (0.70, 0.79)	0.81 (0.78, 0.83)	0.85 (0.84, 0.86)	0.82	0:38:49	28
MM	**0.83 (0.74, 0.91)**	**0.90 (0.87, 0.92)**	0.73 (0.66, 0.80)	0.69 (0.60, 0.75)	0.75 (0.74, 0.76)	0.78	0:00:10	28
flowClust/Merge	0.69 (0.55, 0.79)	0.84 (0.81, 0.86)	0.81 (0.77, 0.85)	0.77 (0.74, 0.79)	0.73 (0.58, 0.85)	0.77	2:12:00	24
L2kmeans	0.64 (0.57, 0.72)	0.79 (0.74, 0.83)	0.70 (0.65, 0.75)	0.78 (0.75, 0.81)	0.81 (0.80, 0.82)	0.74	0:08:03	20
CDP	0.52 (0.46, 0.58)	0.87 (0.85, 0.90)	0.50 (0.48, 0.52)	0.71 (0.68, 0.75)	0.88 (0.86, 0.90)	0.70	0:00:57	19
SWIFT	0.63 (0.56, 0.70)	0.67 (0.62, 0.71)	0.59 (0.55, 0.62)	0.69 (0.64, 0.74)	0.87 (0.86, 0.88)	0.69	1:14:50	15

Figure 2: Aghaeepour (2013)

Biclustering

- Goal is to identify subgroups of observations and variables that are highly correlated
- For example:
  - In gene expression data, we may want to identify genes that are co-expressed in a subset of samples
    - Several genes may only be expressed (made/increased/elevated) in patients with, say, flu, but these genes are not expressed by healthy individuals or patients with other diseases
  - Attempting to cluster genes and samples separately may miss these patterns

ANOVA model for biclustering I

- We assume that the expression level of gene i in sample j is given by:

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

- where μ is average expression level, α_i is the effect of gene i, β_j is the effect of sample j, and ϵ_{ij} is the error term.
  - Note that, in this case, samples are along the columns and variables along the rows.
- A cluster is a subset of genes and samples for which the α_i and β_j are similar.