Clustering in R

Vladimir Batagelj

IMFM Ljubljana, IAM UP Koper and University of Ljubljana

7th Internatonal Summer School
THEORY AND METHODS OF NETWORK ANALYSIS
HSE, Moscow, Russia, 19-23 June, 2017

Outline

Dissimilarities on $\mathbb{R}^m$ / examples 1

n	measure	definition	range	note
1	Euclidean	$\sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$	$[0, \infty)$	$M(2)$
2	Sq. Euclidean	$\sum_{i=1}^{m}(x_i - y_i)^2$	$[0, \infty)$	$M(2)^2$
3	Manhattan	$\sum_{i=1}^{m}\lvert x_i - y_i\rvert$	$[0, \infty)$	$M(1)$
4	rook	$\max_{i=1}^{m}\lvert x_i - y_i\rvert$	$[0, \infty)$	$M(\infty)$
5	Minkowski	$\sqrt[p]{\sum_{i=1}^{m}(x_i - y_i)^p}$	$[0, \infty)$	$M(p)$

Dissimilarities on $\mathbb{R}^m$ / examples 2

n	measure	definition	range	note
6	Canberra	$\sum_{i=1}^{m} \frac{\lvert x_i - y_i \rvert}{\lvert x_i + y_i \rvert}$	$[0, \infty)$	
7	Heincke	$\sqrt{\sum_{i=1}^{m} (\frac{\lvert x_i - y_i \rvert}{\lvert x_i + y_i \rvert})^2}$	$[0, \infty)$	
8	Self-balanced	$\sum_{i=1}^{m} \frac{\lvert x_i - y_i \rvert}{\max(x_i, y_i)}$	$[0, \infty)$	
9	Lance-Williams	$\frac{\sum_{i=1}^{m} \lvert x_i - y_i \rvert}{\sum_{i=1}^{m} x_i + y_i}$	$[0, \infty)$	
10	Correlation c.	$\frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}$	$[1, -1]$	

(Dis)similarities on $\mathbb{B}^m$ / examples

Let $\mathbb{B} = \{0, 1\}$. For $X, Y \in \mathbb{B}^m$ we define $a = XY$, $b = X\overline{Y}$, $c = \overline{X}Y$, $d = \overline{X}\overline{Y}$. It holds $a + b + c + d = m$. The counters a, b, c, d are used to define several (dis)similarity measures on binary vectors.

In some cases the definition can yield an indefinite expression $\frac{0}{0}$. In such cases we can restrict the use of the measure, or define the values also for indefinite cases. For example, we extend the values of Jaccard coefficient such that $s_4(X, X) = 1$. And for Kulczynski coefficient, we preserve the relation $T = \frac{1}{s_4} - 1$ by

$$s_4 = \begin{cases} 1 & d = m \\ \frac{a}{a+b+c} & \text{otherwise} \end{cases} \qquad s_3^{-1} = T = \begin{cases} 0 & a = 0,\ d = m \\ \infty & a = 0,\ d < m \\ \frac{b+c}{a} & \text{otherwise} \end{cases}$$

We transform a similarity s from $[1, 0]$ into dissimilarity d on $[0, 1]$ by $d = 1 - s$.

For details see Batagelj, Bren (1995).

(Dis)similarities on $\mathbb{B}^m$ / examples 1

n	measure	definition	range
1	Russel and Rao (1940)	$\frac{a}{m}$	$[1,0]$
2	Kendall, Sokal-Michener (1958)	$\frac{a+d}{m}$	$[1,0]$
3	Kulczynski (1927), T^{-1}	$\frac{a}{b+c}$	$[\infty,0]$
4	Jaccard (1908)	$\frac{a}{a+b+c}$	$[1,0]$
5	Kulczynski	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$	$[1,0]$
6	Sokal & Sneath (1963), un_4	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$	$[1,0]$
7	Driver & Kroeber (1932)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	$[1,0]$
8	Sokal & Sneath (1963), un_5	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$[1,0]$

(Dis)similarities on $\mathbb{B}^m$ / examples 2

n	measure	definition	range
9	Q_0	$\frac{bc}{ad}$	$[0, \infty]$
10	Yule (1927), Q	$\frac{ad-bc}{ad+bc}$	$[1, -1]$
11	Pearson, ϕ	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$[1, -1]$
12	– bc –	$\frac{4bc}{m^2}$	$[0, 1]$
13	Baroni-Urbani, Buser (1976), S^{**}	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	$[1, 0]$
14	Braun-Blanquet (1932)	$\frac{a}{\max(a+b,a+c)}$	$[1, 0]$
15	Simpson (1943)	$\frac{a}{\min(a+b,a+c)}$	$[1, 0]$
16	Michael (1920)	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	$[1, -1]$

Dissimilarities between sets

Let $\mathcal{F}$ be a finite family of subsets of the finite set U; $A, B \in \mathcal{F}$ and let $A \oplus B = (A \setminus B) \cup (B \setminus A)$ denotes the symmetric difference between A and B.

The 'standard' dissimilarity between sets is the *Hamming distance*:

$$d_H(A, B) := \mathrm{card}(A \oplus B)$$

Usually we normalize it $d_h(A, B) = \frac{1}{M}\mathrm{card}(A \oplus B)$. One normalization is $M = \mathrm{card}(U)$; the other $M = m_1 + m_2$, where m_1 and m_2 are the first and the second largest value in $\{\mathrm{card}(X) : X \in \mathcal{F}\}$.

Other dissimilarities

$$d_s(A, B) = \frac{\mathrm{card}(A \oplus B)}{\mathrm{card}(A) + \mathrm{card}(B)} \qquad d_u(A, B) = \frac{\mathrm{card}(A \oplus B)}{\mathrm{card}(A \cup B)}$$

$$d_m(A, B) = \frac{\max(\mathrm{card}(A \setminus B), \mathrm{card}(B \setminus A))}{\max(\mathrm{card}(A), \mathrm{card}(B))}$$

For all these dissimilarities $d(A, B) = 0$ if $A = B = \emptyset$.

Problems with dissimilarities

Functions in R: `dist`, `cluster/daisy`

What to do in the case of *mixed units* (with variables measured in different types of scales)?

- conversion to a common scale
- compute the dissimilarities on homogeneous parts and combine them (Gower's dissimilarity)

Fairness of dissimilarity – all variables contribute equally. Approaches: use of normalized variables, analysis of dependencies among variables.

Gower's dissimilarity

the Gower dissimilarity coefficient for a mix of variables
$d_{ij} = \sum_{v=1}^{m} \frac{\delta_{ijv} d_{ijv}}{\sum_{i=1}^{m} \delta_{ijv}}$
where δ_{ijv} is a binary indicator equal to one whenever both observations i and j are nonmissing for variable v, and zero otherwise. Observations with missing values are not included.
For binary and nominal variables v, $d_{ijv} = 0$ if $x_{iv} = x_{jv}$; and $d_{ijv} = 1$ otherwise.
Ordinal variables v are considered as categorical ordinal variables and the values are substituted with the corresponding position index, r_{iv} in the factor levels. These position indexes are transformed in the following manner $z_{iv} = \frac{r_{iv}-1}{\max_k r_{kv}-1}$ These new values, z_{iv}, are treated as observations of an interval scaled variable.
For continuous variables v,
$d_{ijv} = \frac{|x_{iv}-x_{jv}|}{\max_k(x_{kv})-\min_k(x_{kv})}$
d_{ijv} is set to 0 if $\max_k(x_{kv}) = \min_k(x_{kv})$.

Package `StatMatch`.

Solving the clustering problem

Finite - solution always exists, but in most cases algorithmically hard problem.

Heuristics:

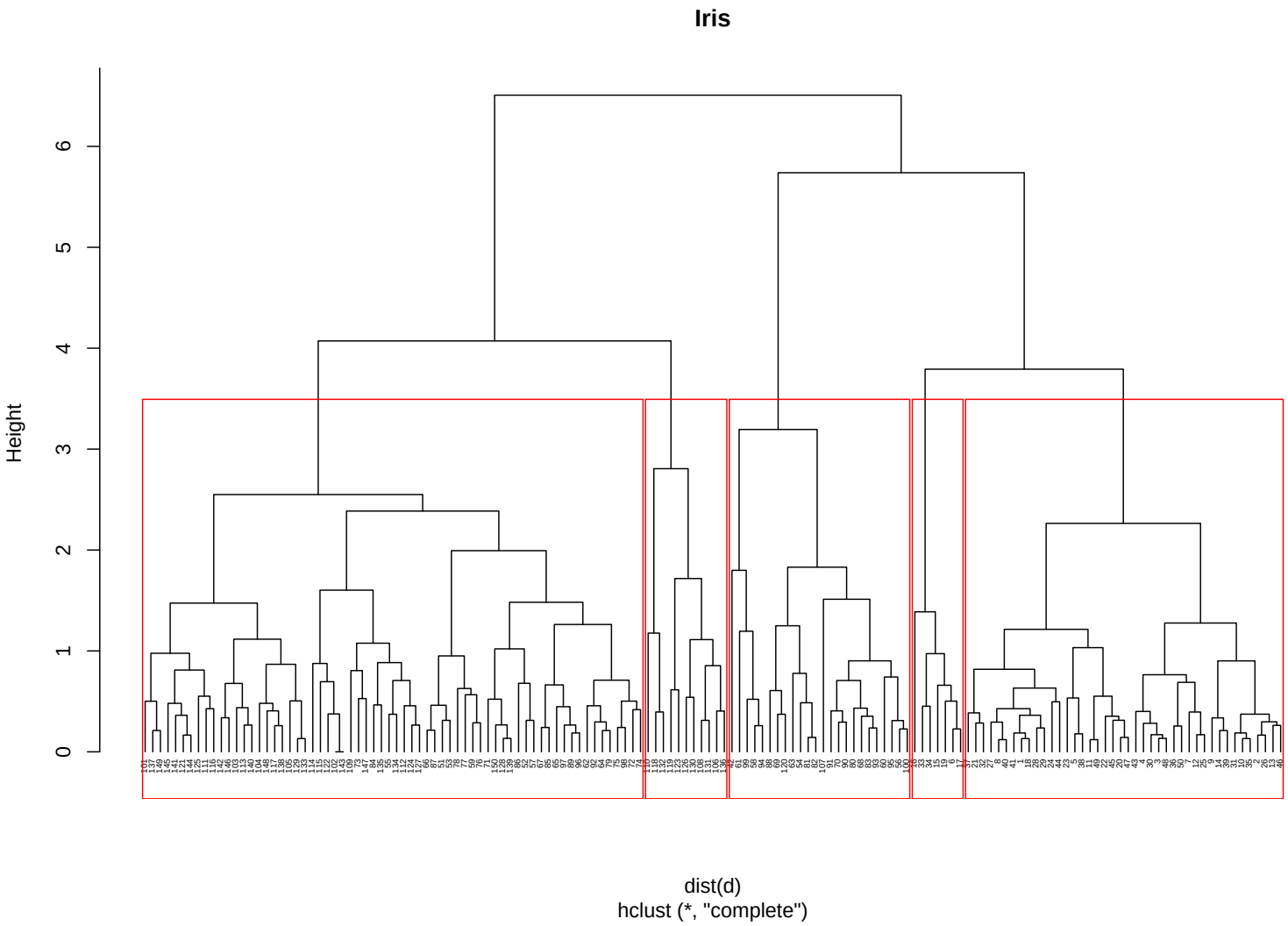
- hierarchical
  - agglomerative methods (`hclust`, `cluster/agnes`, `amap/hcluster`, `amap/hclusterpar`)
  - divisive methods (`cluster/diana`, `cluster/mona`)
  - adding methods
- local optimization (leaders method) (`kmeans`, `cluster/pam`, `cluster/clara`, `cluster/fanny`)
- linear algebra methods
- graph theory methods
- other methods (`mclust/Mclust`)

Fisher's irises

Anderson 1935 / Fisher 1936

```
> help(iris)
> attach(iris)
> z <- function(x){(x-mean(x))/sd(x)}
> d <- cbind(z(Sepal.Length),z(Sepal.Width),z(Petal.Length),z(Petal.Width))
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
...
150          5.9         3.0          5.1         1.8  virginica
> d
              [,1]        [,2]        [,3]          [,4]
  [1,] -0.89767388  1.01560199 -1.33575163 -1.3110521482
  [2,] -1.13920048 -0.13153881 -1.33575163 -1.3110521482
...
[150,]  0.06843254 -0.13153881  0.76021149  0.7880306775
> t <-hclust(dist(d))
> pdf("iris.pdf",width=11.7,height=8.3,paper="a4r")
> plot(t,hang=-1,cex=0.4,main="Iris")
> rect.hclust(t,k=5,border="red")
> dev.off()
> p <- cutree(t,k=5)
> iris$Species[p==1]
 [1] setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa
> library(cluster)
> r <- agnes(d,method="ward")
> plot(r,which.plots=2,main="iris",cex=0.2)
```

Irises dendrogram



Places

http://lib.stat.cmu.edu/datasets/places.data

```
> places <- read.csv2("places.txt",skip=2)
> P <- as.matrix(places[,-10])
> rownames(P) <- places$Place
> P[1:10,]
> Q <- apply(P,2,z)
> R <- scale(P)   # R = Q
> s <- kmeans(Q,centers=10,iter.max=30)
> ps <- s$cluster
> s$centers
   Climate.Terrain    Housing HealthCare.Environment        Crime Transportation   Education
1       -0.17190445 -0.2544117            -0.03706850  0.088538118      0.6135215  0.15364718 -0.
2       -0.03083485 -0.4961185            -0.40320613 -0.897595905     -0.6807537  0.15272723 -0.
3       -1.72307932  2.6168447             0.61683070 -0.188189327     -0.1523102 -0.22507766  0.
4       -1.33628400  2.1346807             4.16126912  1.809056089      2.1577288  1.69407024  5.
5       -0.11037444  0.5466795             1.54904444  0.381941359      1.1044245  1.55545539  1.
6       -0.35109872  0.1903805             0.07994732  1.403431919      0.4379780 -0.14489137  0.
7        1.09502512  0.5134756            -0.37461400 -0.057706916      0.3031451 -0.20469688 -0.
8       -0.02806038 -0.6116676            -0.58742411  0.008760816     -0.8750241 -1.19534723 -0.
9        0.50175579 -0.1640700            -0.50260187  0.257568871     -0.4700833 -0.00443408 -0.
10       1.94737379 -0.2680192            -0.27580536 -1.088287090      0.1997809 -0.39439494 -0.
> rownames(P)[ps==4]
[1] " Boston, MA"                  " Chicago, IL"                 " Los Angeles, Long Beach, CA"
[4] " New York, NY"                " San Francisco, CA"           " Washington, DC-MD-VA"
```