

Lösung zu Kapitel 12: Beispiel 4

In einer Umfrage im Mai 2008 wurden 229 Personen (mit Kabel-TV- oder Satelliten-TV-Empfang) im Raum Wien zu ihrem TV-Sehverhalten befragt. Ein Teil dieser Umfrage zielte darauf ab, Eigenschaften (informativ, sensationslüstern etc.) von Fernsehsendern herauszufiltern. Die Antworten zu den Fragen nach Aktualität, kritischer Berichterstattung und politischer Unabhängigkeit sind im Datenfile `tvimage.csv` für die drei Sender ORF1, Pro7 und RTL enthalten.

- Gibt es unter den Befragten Gruppen, die das TV-Angebot ähnlich einschätzen?

Die Variablen enthalten Antworten auf reine Ja-Nein-Fragen. Eine Möglichkeit mit diesen 0-1-Angaben umzugehen, ist, sie wie metrische Variable zu behandeln und etwa ein gewohntes hierarchisches Clustern oder ein Centroid-Verfahren wie im Buch aufzurufen. Da lauter Ja-Nein-Antworten vorliegen, demonstrieren wir die im Buch nur in einem Nebensatz erwähnte Funktion `mona()`, die speziell für binäre Variablen entwickelt worden ist.

Im Datenfile sind einige wenige Werte (für die Einschätzung der politischen Unabhängigkeit des ORF) fehlend. Nach dem Einlesen werden die Beobachtungen mit den fehlenden Werten eliminiert.

R

```
> library("cluster")
> tvimage <- read.csv2("tvimage.csv", header = TRUE)
> attach(tvimage)
> tvimage1 <- subset(tvimage, polunabh_ORF1 != 9)
> detach(tvimage)
> clust_tvi <- mona(tvimage1)
> plot(clust_tvi)
```

Für einen übersichtlichen Bannerplot (► Abbildung 1) liegen zu viele Beobachtungen vor. Wir erkennen aber leichtere und tiefere Einschnitte im Bannerplot. Die tiefen Einschnitte entsprechen starken Unterschieden zwischen den benachbarten Clustern; weniger tiefe Einschnitte deuten weniger starke Unterschiede zwischen den Clustern an.

Die Tiefe der Einschnitte (Skalierung auf der x-Achse des Bannerplots) ist in einem numerischen Outputteil von `mona()` enthalten.

R

```
> table(clust_tvi$step)
```

```
 0  1  2  3  4  5  6  7  8  9
153 1  2  4  8 16 19 16  6  2
```

Es liegt also eine ganz tiefe Trennung (*step* = 1) des Datensatzes in zwei Cluster vor. Welche Variable dafür verantwortlich ist, kann ebenfalls ermittelt werden.

R

```
> step1 <- which(clust_tvi$step == 1)
> clust_tvi$variable[step1]
```

Banner of `mona(x = timage1)`

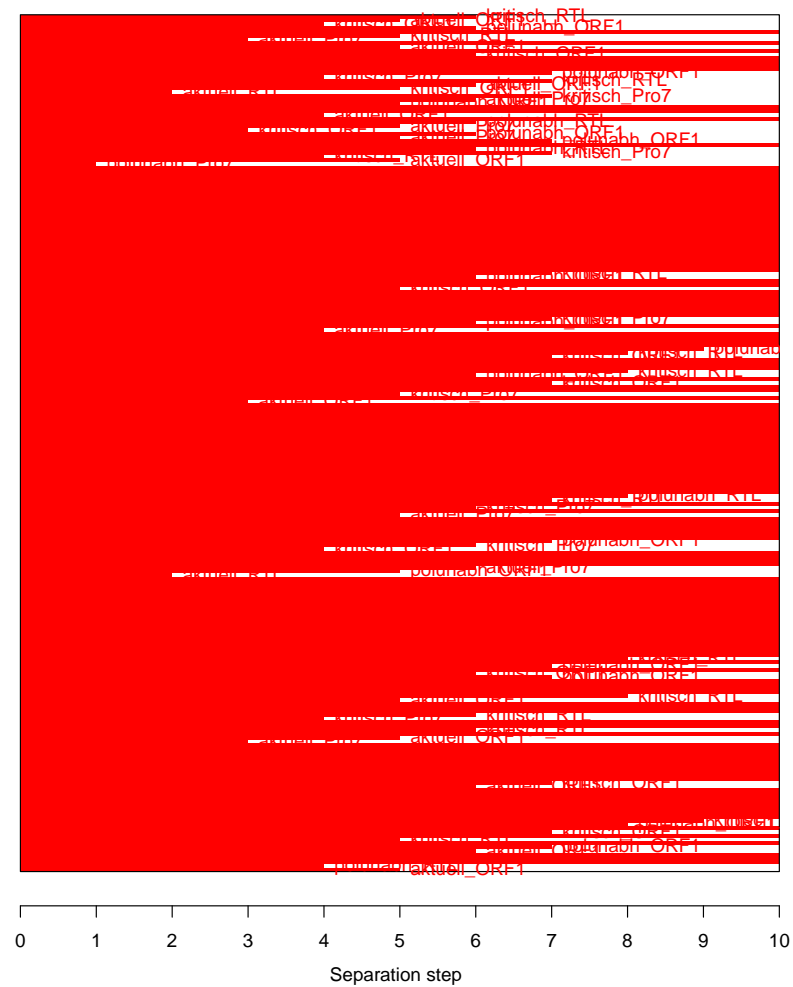


Abbildung 1: Bannerplot: TV-Image

[1] "polunabh_Pro7"

Welche Beobachtungen liegen in diesen beiden Grobclustern?

R

```
> clust_tvi$order[1:step1]
```

```
[1] 1 48 58 23 44 180 46 3 209 5 153 20 68 70 89 141 49 81 100
[20] 123 148 24 169 51 59 135 174 88 195 98 173 26 67 69 101 168 181 208
[39] 53 202
```

```
> von <- step1 + 1  
> bis <- dim(tvimage1)[1]  
> clust_tvi$order[von:bis]
```

```
[1] 2 6 30 38 39 52 55 60 65 76 87 93 95 97 107 111 127 133  
[19] 144 145 147 159 166 167 172 176 196 224 225 214 149 165 175 9 71 143  
[37] 191 201 210 213 228 73 50 183 18 36 91 104 160 105 198 22 80 94  
[55] 192 96 66 120 85 108 154 86 151 7 13 15 16 19 25 28 34 42  
[73] 64 83 84 113 118 125 132 163 194 199 211 219 220 221 223 226 99 62  
[91] 72 31 204 61 63 78 116 136 139 187 140 206 102 106 128 177 200 137  
[109] 110 4 11 14 21 37 45 54 57 82 90 109 130 134 146 156 164 170  
[127] 186 188 190 203 217 79 103 124 193 17 33 41 56 182 215 179 10 112  
[145] 117 126 115 92 162 227 205 212 184 8 35 43 75 158 171 185 197 207  
[163] 216 218 40 12 29 114 119 122 150 152 155 178 189 32 222 138 161 27  
[181] 74 129 77 121 142 157 47 131
```

Analog können feinere Cluster untersucht werden, der Aufwand dafür steigt natürlich stark an.