data.Normalization (clusterSim)

Types of variable normalization formulas

A. Variable (column) normalization

Variable (column) normalization can be applied to any data matrix.

1	Selection of objects and variables	data matrix $[x_{ij}]$		
2	Variable scale level	Ratio	Ratio	Interval
	Selection of variable normalization formula	n6 – quotient transformation (x/sd) n6a – positional quotient transformation (x/mad) n7 – quotient transformation (x/range) n8 – quotient transformation (x/max) n9 – quotient transformation (x/mean) n9a – positional quotient transformation (x/median) n10 – quotient transformation (x/sum) n11 – quotient transformation x/sqrt(SSQ)	n1 – standardization n2 – positional standardization n3 – unitization n3a – positional unitization n4 – unitization with zero minimum n5 – normalization in range [–1, 1] n5a – positional normalization in range [–1, 1]	n1 – standardization n2 – positional standardization n3 – unitization n3a – positional unitization n4 – unitization with zero minimum n5 – normalization in range [–1, 1] n5a – positional normalization in range [–1, 1]
	Transformed variable scale level	Ratio	Interval	Interval

(n1) $$z_{ij} = (x_{ij} - \bar{x}_j) / s_j$$

(n2) $$z_{ij} = (x_{ij} - med_j) \big/ mad_j$$

(n3) $$z_{ij} = (x_{ij} - \bar{x}_j) / r_j$$

(n3a) $$z_{ij} = (x_{ij} - med_j) \big/ r_j$$

(n4) $$z_{ij} = \left[x_{ij} - \min_i \{x_{ij}\} \right] \Big/ r_j$$

(n5) $$z_{ij} = (x_{ij} - \bar{x}_j) \Big/ \max_i \left| x_{ij} - \bar{x}_j \right|$$

(n5a) $$z_{ij} = (x_{ij} - med_j) \Big/ \max_i \left| x_{ij} - med_j \right|$$

(n6) $$x_{ij} \big/ s_j$$

(n6a) $$z_{ij} = x_{ij} \big/ mad_j$$

(n7) $$x_{ij} \big/ r_j$$

(n8) $$x_{ij} \Big/ \max_i \{x_{ij}\}$$

(n9) $$x_{ij} \big/ \bar{x}_j$$

(n9a) $$z_{ij} = x_{ij} \big/ med_j$$

(n10) $$x_{ij} \Big/ \sum_{i=1}^{n} x_{ij}$$

(n11) $$x_{ij} \Big/ \sqrt{\sum_{i=1}^{n} x_{ij}^2}$$

where: $x_{ij}\,(z_{ij})$ – i-th observation on j-th variable (i-th normalized observation on j-th variable),

$\bar{x}_j\,(s_j)$ – mean (standard deviation) for j-th variable,

med_j (mad_j) – median (median absolute deviation) for j-th variable,

$r_j = \max_i \{x_{ij}\} - \min_i \{x_{ij}\}$.

B. Object (row) normalization

The same normalization procedures can be applied as for variable (column) normalization. Object (row) normalization makes sense only when all variables are expressed in the same unit. This is often the case for instance with structural data.

References

Anderberg, M.R. (1973), *Cluster analysis for applications*, Academic Press, New York, San Francisco, London.

Gatnar, E., Walesiak, M. (Eds.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych [Multivariate statistical analysis methods in marketing research]*, Wydawnictwo AE, Wroclaw, 35-38.

Jajuga, K., Walesiak, M. (2000), *Standardisation of data set under different measurement scales*, In: R. Decker, W. Gaul (Eds*.), Classification and information processing at the turn of the millennium*, Springer-Verlag, Berlin, Heidelberg, 105-112.

Milligan, G.W., Cooper, M.C. (1988), *A study of standardization of variables in cluster analysis*, "Journal of Classification", vol. 5, 181-204.

Młodak, A. (2006), *Analiza taksonomiczna w statystyce regionalnej*, Difin, Warszawa.