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 $[x_{ij}]$		
	Variable scale level	Ratio	Ratio	Interval
2	Selection of variable normalization formula	n6 – quotient transformation n6a – positional quotient transformation n7 – quotient transformation n8 – quotient transformation n9 – quotient transformation n9a – positional quotient transformation n10 – quotient transformation n11 – quotient transformation	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] n12 – normalization n12a – positional normalization n13 – normalization with zero being the central point	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] n12 – normalization n12a – positional normalization n13 – normalization with zero being the central point
	Transformed variable scale level	Ratio	Interval	Interval

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(n12) $$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

(n12a) $$z_{ij} = \frac{x_{ij} - med_j}{\sqrt{\sum_{i=1}^{n}(x_{ij} - med_j)^2}}$$

(n13)[1] $$z_{ij} = \frac{x_{ij} - m_j}{r_j/2}$$

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,

$med_j = \underset{i}{med}(x_{ij})$ – median for j-th variable,

$mad_j = \underset{i}{mad}(x_{ij})$ – median absolute deviation for j-th variable,

$r_j = \underset{i}{max}\{x_{ij}\} - \underset{i}{min}\{x_{ij}\}$ – range for j-th variable,

$m_j = \frac{\underset{i}{max}\{x_{ij}\} + \underset{i}{min}\{x_{ij}\}}{2}$ – mid-range for j-th variable.

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. DOI: https://doi.org/10.1007/978-3-642-57280-7_11.

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.

Walesiak, M. (2014), *Przegląd formuł normalizacji wartości zmiennych oraz ich własności w statystycznej analizie wielowymiarowej [Data normalization in multivariate data analysis. An overview and properties]*, "Przegląd Statystyczny" (Statistical Review), vol. 61, no 4, 365-374.

[1] http://www.benetzkorn.com/2011/11/data-normalization-and-standardization/ (1.06.2014).