INDICE DE SIMILARITÉ (K1, K9, K10)

(03 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

La notion d'**indice de similarité** est propre à l'**analyse des données : données catégorielles** ou données classées, **classifications automatiques**.

(i) La version suivante est courante en analyse des données classées. Soit $(\Omega, \mathcal{F})$ un **espace fondamental** et $X : \Omega \mapsto (N_1)^K$ un **vecteur aléatoire** dont les coordonnées X_k sont des **variables indicatrices** (ie des va à valeurs dans $N_1 = \{0, 1\}$). On note resp $x_1 = X(\omega_1)$ et $x_2 = X(\omega_2)$ les **observations** correspondant aux éléments $\omega_1 \in \Omega$ et $\omega_2 \in \Omega$.

On appelle alors :

(a) **indice de similarité**, ou **indice de similitude**, le nombre :

$$(1) \quad s_{12} = K^{-1} x_1' x_2 = x_1' x_2 / (e_K' e_K),$$

qui mesure la **proportion** de coordonnées de X égales à 1 ;

(b) **indice de dissimilarité**, ou **indice de différenciation**, le nombre :

$$(1)' \quad d_{12} = 1 - s_{12}.$$

En pratique, on suppose que, $\forall k \in N_K^*$:

$$(2) \quad X_k(\omega) = \begin{cases} 1 & \text{si l'élément } \omega \text{ vérifie une propriété } \mathcal{P}_k, \\ 0 & \text{sinon.} \end{cases}$$

L'indice s_{12} mesure alors la proportion du nombre de fois où x_1 et x_2 vérifient la propriétés $\mathcal{P}_k$.

(ii) En analyse des données, on définit, plus généralement, la notion d'indice de similarité comme suit. Soit E un **ensemble** quelconque (eg une **population** ou un **échantillon**).

On appelle (**indice de**) **similarité** sur E toute fonction $s : E^2 \mapsto \mathbf{R}_+$ vérifiant, $\forall (x, y) \in E^2$, les deux propriétés suivantes :

$$(a) \quad s(x, y) = s(y, x) \text{ (symétrie) ;}$$

$$(b) \quad s(x, x) = s(y, y) > s(x, y).$$

L'inégalité $s(x, y) > s(z, t)$ signifie que x et y sont davantage semblables (ie sont plus « proches » entre eux) que ne le sont z et t.

Cette notion, substituée à celle de **distance** (cf **espace métrique**), est utilisée en **classification automatique**.

A titre d'exemple, si $T \in M_{mn}(\{0,1\})$ est une **matrice** (ou un **tableau statistique**) tq $t_{ij} \in \{0, 1\}, \forall (i, j) \in N_m^* \times N_n^*$, on dit que l'**unité statistique** (individu, etc) i possède le **caractère** j lorsque $t_{ij} = 1$ et ne le possède pas sinon.

On peut alors comparer des unités α et β en définissant les entiers suivants :

$$(3) \quad \begin{aligned} S_{\alpha\beta} &= \sum_{j=1}^n t_{\alpha j} t_{\beta j}, \\ T_{\alpha\beta} &= \sum_{j=1}^n (1 - t_{\alpha j})(1 - t_{\beta j}), \\ U_{\alpha\beta} &= \sum_{j=1}^n t_{\alpha j} (1 - t_{\beta j}), \\ V_{\alpha\beta} &= \sum_{j=1}^n (1 - t_{\alpha j}) t_{\beta j}, \end{aligned}$$

qui représentent resp le nombre de caractères possédés par α et β , non possédés par α et β , possédés par α mais non par β , non possédés par α mais possédés par β . On a donc $S_{\alpha\beta} + T_{\alpha\beta} + U_{\alpha\beta} + V_{\alpha\beta} = n, \forall (\alpha, \beta)$.

L'**indice de similarité** s entre éléments de T est alors basé sur le quadruplet $(S_{\alpha\beta}, T_{\alpha\beta}, U_{\alpha\beta}, V_{\alpha\beta})$.

Ainsi, en notant T_α et T_β deux lignes de T , on définit :

(a) l'**indice de L.R. DICE - T. SORENSEN** :

$$(4) \quad s(T_\alpha, T_\beta) = 2 S_{\alpha\beta} / (2 S_{\alpha\beta} + U_{\alpha\beta} + V_{\alpha\beta}) \in [0, 1];$$

(b) l'**indice de P. JACCARD** :

$$(5) \quad s(T_\alpha, T_\beta) = S_{\alpha\beta} / (S_{\alpha\beta} + U_{\alpha\beta} + V_{\alpha\beta}) \in [0, 1];$$

(c) l'**indice de T.W. KULCZYNSKI** :

$$(6) \quad s(T_\alpha, T_\beta) = S_{\alpha\beta} / (U_{\alpha\beta} + V_{\alpha\beta}) \in \mathbf{R}_+;$$

(d) l'**indice de D.J. ROGERS - T.T. TANIMOTO** :

$$(7) \quad s(T_\alpha, T_\beta) = (S_{\alpha\beta} + T_{\alpha\beta}) / (S_{\alpha\beta} + T_{\alpha\beta} + U_{\alpha\beta}) \in [0, 1];$$

(e) l'**indice de T.R. RAO - P.F. RUSSELL** :

$$(8) \quad s(T_\alpha, T_\beta) = n^{-1} S_{\alpha\beta} \in [0, 1];$$

(f) l'**indice de C.D. MICHENER - R.R. SOKAL** :

$$(9) \quad s(T_\alpha, T_\beta) = n^{-1} (S_{\alpha\beta} + T_{\alpha\beta}) \in [0, 1];$$

(g) les **indices de P.H.A. SNEATH - R.R. SOKAL** :

$$(10) \quad s(T_\alpha, T_\beta) = (S_{\alpha\beta} + T_{\alpha\beta}) / (U_{\alpha\beta} + V_{\alpha\beta}) \in \mathbf{R}_+,$$

$$s(T_\alpha, T_\beta) = S_{\alpha\beta} / \{S_{\alpha\beta} + 2(U_{\alpha\beta} + V_{\alpha\beta})\} \in [0, 1].$$

(iii) Si s est un indice de similarité (fini) sur E , le nombre :

$$(11) \quad s^*(x, y) = s(x, y) / s(x, x) \in [0, 1]$$

est appelé **indice de similarité normé**.

Si d est une distance sur E , la fonction s définie par :

$$(12) \quad s(x, y) = \{1 + d(x, y)\}^{-1} \in [0, 1], \quad \forall (x, y) \in E^2,$$

est un indice de similarité sur E .

A tout indice de similarité on peut associer un **indice de dissimilarité**.