CLASSIFICATION AUTOMATIQUE (K9)
(24 / 12 / 2019, © Monfort, Dicostat2005, 2005-2019)

On appelle **classification automatique**, ou **classification formelle**, une méthode dont l'objet est de structurer, de façon systématique, formelle et légère, un **ensemble** fini (**population** ou **échantillon**) A, constitué d'éléments (**unités statistiques**) a_n ($n \in N_N^*$), en un ensemble de classes homogènes aussi réduit que possible (cf **homogénéité**, **partition**). Cet ensemble est souvent de cardinalité élevée (cf **grande base de données**).

Il s'agit donc de regrouper (ou d'éclater) un ensemble, souvent complexe, d'unités (ou d'**observations** effectuées sur elles) dans des classes tq toutes les unités d'une même classe possèdent des caractéristiques semblables (cf **caractère**) : l'algorithme mis en oeuvre tient compte des différenciations existant entre les unités à travers ces caractéristiques. Les classes sont ainsi définies (ou construites) à l'aide des **variables** (ou caractères) mesuré(e)s sur les unités statistiques. Ces variables peuvent être numériques ξ_1 ,..., ξ_K (cf **variables quantitatives**), qualitatives (**variables qualitatives** ou attributs) κ_1 ,..., κ_G ou mixtes : elles permettent de définir des **statistiques** utilisables pour la définition des classes.

(i) Dans la mise en oeuvre des nombreuses méthodes existantes, le choix des unités statistiques, ou celui des observations effectuées sur ces unités, est central.

Les **données** se présentent souvent sous la forme d'un (N,I)-**tableau statistique** (matrice) $T = (t_{ni})_{(n,i)}$, où t_{ni} désigne l'**observation** de la variable i sur l'unité n.

Un cas fréquent est celui où $t_{ni} \in \{0, 1\}$ (**variable binaire**).

Un problème important de choix des **unités de mesure** (étalonnage, **échelle de mesure**) se pose souvent. Des **échelles de mesure** non linéaires sont couramment utilisées : on remplace la mesure t_{ni} par la valeur $Z_{ni} = \varphi_i (t_{ni})$ si l'échelle φ_i est commune à toutes les observations, ou par la valeur $Z_{ni} = \varphi_n (t_{ni})$ si l'échelle φ_n est commune à toutes les variables (cf **transformation des données**).

On peut encore utiliser des **données centrées** (T est remplacé par Z = P T, où P est la **matrice de centrage par rapport à la moyenne**) ou des **données normées** (ie centrées-réduites ou « standardisées » : T est remplacé par $Z = P T S^{-1}$, ou par $Z = S^{-1} T P$, où S est la **matrice diagonale** de format ad hoc dont les éléments sont les **écarts-types** empiriques des variables ou des observations), ou encore la **matrice des covariances** empiriques (cf **centrage**, **normalisation**, **variable centrée**, **variable normée**, **variable réduite**).

(ii) Trois grandes familles de **méthodes de classification** existent :

(a) les méthodes **hiérarchiques**, qui définissent des suites de partitions (ou chaînes) emboîtées dans l'ensemble des unités (ou des observations) (cf **analyse hiérarchique**, **hiérarchie**) ;

(b) les méthodes **non hiérarchiques**, qui définissent d'emblée une seule partition (dont le nombre de classes est donné a priori) ;

(c) les méthodes **par transferts successifs des unités entre classes** (algorithmes).

Ces méthodes sont le plus souvent de type algorithmique.

(iii) Un **algorithme** de classification peut être :

(a) soit (situation assez courante) un **algorithme ascendant**, ou **algorithme en montée**, ou **algorithme par agrégation**, auquel cas les unités statistique de base sont regroupées, à chaque étape, en classes de moins en moins nombreuses (passage de la partition la plus fine vers la partition triviale) (cf aussi **agrégation selon la variance**) ;

(b) soit un **algorithme descendant**, ou **algorithme en descente**, ou **algorithme par désagrégation**, auquel cas la population initiale est partitionnée, à chaque étape, en classes de plus en plus nombreuses (passage de la partition triviale vers la plus fine).

(iv) La **mesure de proximité**, ou **mesure de similarité** (ressemblance ou dissemblance), entre les unités (ou entre groupes d'unités) peut être de nature diverse. Elle peut être :

(a) un **indice de similarité** ou un **indice de dissimilarité** ;

(b) un **coefficient de corrélation** : **coefficient de corrélation linéaire** (de BRAVAIS-PEARSON), **coefficient de corrélation des rangs**, **coefficient de corrélation ponctuel**, etc ;

(c) une **distance**.

Si les variables sont numériques (ou non numériques mais codifiées) et éventuellement transformées au préalable, la distance euclidienne usuelle entre deux variables d'indices k et l s'écrit (cf **espace euclidien**) :

(1) $d^2(k, l) = ||x_k - x_l||_2^2 = \sum_{n=1}^{N} (x_{nk} - x_{nl})^2, \quad \forall (k, l) \in (N_K^*)^2.$

Cette distance euclidienne peut être pondérée, eg sous la forme :

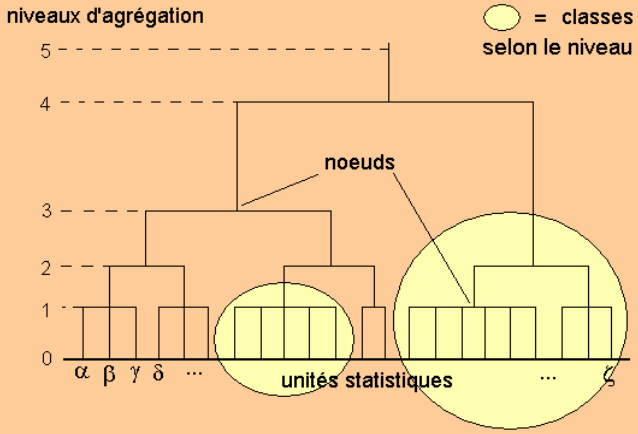
(2) $d_\omega^2(k, l) = ||x_k - x_l||_\omega^2 = \sum_{n=1}^{N} \omega_n (x_{nk} - x_{nl})^2, \quad \forall (k , l) \in (N_K^*)^2,$

où $\omega = (\omega_1 ,..., \omega_N)' \in S_N$ (**simplexe** de $\mathbf{R}^N$).

On utilise aussi d'autres distances (eg la **distance de MAHALANOBIS**) ou des **quasi-distances** (ou **pseudo-distances**, ou **semi-distances**).

Ces formules se transposent au cas, dual, des distances entre observations d'indices $(\alpha, \beta) \in (N_K^*)^2$.

Selon la méthode de classification, la mesure de proximité utilisée sert soit de **critère d'éclatement**, soit de **critère d'agglutination** (ou **critère d'agrégation**), des unités (cf aussi **agrégation selon la variance**). La deuxième situation caractérise surtout les méthodes hiérarchiques. Ces dernières aboutissent ainsi, au bout d'un nombre fini d'étapes, à un **arbre de classification**, ou **arbre d'agrégation**, dont la longueur des branches n'est autre que la valeur de la mesure de proximité entre les classes formées à chaque étape (ou lui est proportionnelle) (cf shéma ci-après).



(v) Les procédures précédentes soulèvent plusieurs problèmes pratiques :

(a) celui de la transposition des formules définissant les mesures de proximité entre unités en formules définissant les mesures de proximité entre classes d'unités (notamment à travers une « **unité représentative** » de chaque classe) ;

(b) celui du niveau de l'arbre d'agrégation auquel on admet que la classification opérée est une **classification optimale**. Les méthodes d'**analyse des données** (eg les **analyses factorielles**) permettent souvent d'associer à la structure de la partition obtenue dans une classification des facteurs (ou **axes factoriels**) issus du tableau T.

Ces méthodes sont mises en oeuvre en **analyse exploratoire des données**.

Par ailleurs, elles sont susceptibles d'extension au cas des partitions floues (cf **théorie des parties floues**), dans lesquelles certaines unités peuvent appartenir « plus ou moins » à des classes différentes (donc à des classes non « disjointes »), ie se voient attribuer divers « degrés d'appartenance ».

(vi) D'un point de vue formel, soit Ω un **ensemble** dont les éléments ω représentent des **unités statistiques** élémentaires.

On appelle **classe**, ou **groupe**, ou encore **grappe**, de Ω toute **partie** non vide $\mathcal{G} \subset \Omega$ (ie $\mathcal{G} \in \mathcal{P}(\Omega) \setminus \{\emptyset\}$). Une **classification** sur Ω n'est autre qu'un ensemble $C \subset \mathcal{P}(\Omega)$ de classes tq C.

Si l'on utilise eg un **indice de dissimilarité** d sur $\mathcal{G}$ (ie une fonction d : $E^2 \mapsto \mathbf{R}_+$), la valeur d (C', C") représente la **dissimilarité** entre les classes C' et C". On appelle alors **critère de classification** une fonction $\Phi : \mathcal{G}_E \mapsto \mathbf{R}_+$, où $\mathcal{G}_E$ désigne l'ensemble des classifications $\mathcal{G}$ sur Ω.

Un **problème de classification automatique** est alors un **problème d'optimisation** du type suivant :

(3) $\min \Phi(\mathcal{G})$, sous $\mathcal{G} \in \mathcal{G}_E$.

Dans ce qui précède, d et Φ sont supposés donnés. En pratique, ceci n'est pas toujours le cas. On étudie souvent le comportement de la solution $\tilde{\mathcal{G}}$ de (3) lorsque d ou Φ varie (**robustesse** d'une classification automatique).

Deux exemples classiques de fonctions F sont :

(4)
$$\Phi(\mathcal{G}) = \Sigma_{C \in \mathcal{G}} \varphi(C),$$
$$\Phi(\mathcal{G}) = \max_{C \in \mathcal{G}} \varphi(C),$$

où φ est une fonction donnée, ne dépendant pas de $C \in \mathcal{G}$.

En pratique, comme précédemment, E est un échantillon A d'unités sur lesquelles on peut observer diverses **variables d'intérêt** (descripteurs, etc).