Range ta chambre ! De l'art de classer et grouper ses données

Nicolas Frerebeau, François-Xavier Le Bourdonnec, Stéphanie Leroy

Range ta chambre *!*

De l'art de classer et grouper ses données

Nicolas Frerebeau[1], François-Xavier Le Bourdonnec[2] & Stéphanie Leroy[1]

Introduction

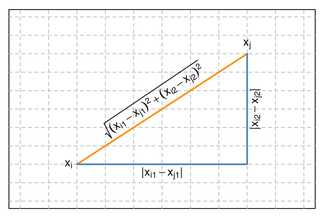
■ Pour la construction de classifications et de groupes au sein de jeux de données plus ou moins complexes, deux approches sont envisageables : selon que l'on cherche à attribuer des individus au sein de catégories existant *a priori* (un jeu de règles permet l'attribution à une classe prédéfinie – approche dite supervisée) ou que l'on cherche à organiser des individus sans système pré-établi (les groupes sont définis par les objets qu'ils contiennent – approche non supervisée).

■ De ces procédures de *clustering*, il découle une quasi infinité de possibilités d'arrangements, sans pour autant que l'une d'entre elles demeure plus valide qu'une autre : seuls comptent les objectifs poursuivis et le problème initial.

■ L'objectif de ce poster est d'explorer quelques problèmes, tant conceptuels que méthodologiques, susceptibles d'émerger selon les situations et de proposer quelques pistes de reflexion.

$$D_1 : \quad d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

$$D_2 : \quad d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{\frac{1}{2}}$$



$$D_3 : \quad d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r \right)^{\frac{1}{r}} \qquad (r \geq 1)$$

$$D_4 : \quad d_{ij}^2 = (x_i - x_j)^t C^{-1} (x_i - x_j)$$

Figure A : principales distances pour des données continues. D_1 : distance de Manhattan ; D_2 : distance euclidienne ; D_3 : distance de Minkowski ; D_4 : distance de Mahalanobis (où C est la matrice de covariance).

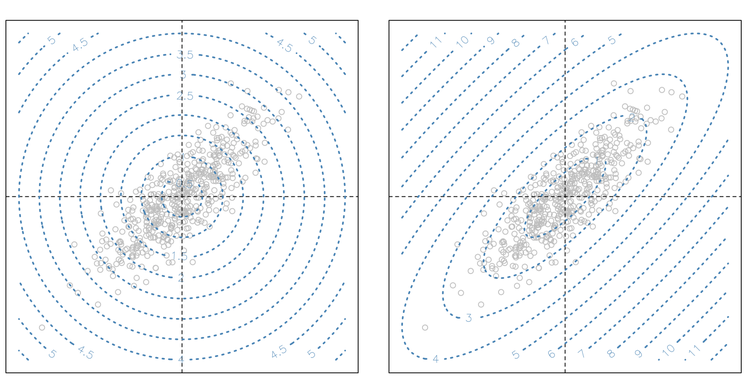


Figure B : calcul des distances euclidienne (gauche) et de Mahalanobis (droite) depuis l'origine.

La distance euclidienne, sans doute la plus intuitive, permet de calculer la distance absolue entre deux observations, mais elle apparaît particulièrement sensible à l'échelle des différentes variables et est fortement affectée par les redondances entre variables. Difficultés dont la distance de mahalanobis, qui repose sur la calcul de la matrice de covariance, peut s'affranchir.

Le choix de la métrique

■ Ces méthodes reposent sur une mesure de similarité (ou de dissemblance) entre plusieurs individus ou groupes d'individus dans un espace multidimensionnel.

■ Reste à choisir une mesure de la distance entre individus **(fig. A)**. Les différentes distances ont des propriétés particulières mais le choix de l'une ou l'autre fait rarement (sinon jamais) l'objet d'une discussion. Le choix de la mesure de proximité entre deux observations ne peut cependant se faire en routine et doit dépendre de la nature des données et du problème posé **(fig. B)**.

■ Par exemple, doit-on s'attacher à mesurer des distances absolues entre deux observations ou des distances construites sur la correlation entre ces mêmes variables ? Les différentes variables ont-elles la même échelle ? Existe-t-il des redondances dans le jeu de données ? Qu'en est-il si l'espace considéré n'est pas isotrope ?

L'absence de l'inconnu

■ Dans le cas des approches supervisées, l'objectif est d'associer des individus à des catégories pré-déterminées.

■ Cela suppose une connaissance exhaustive *a priori* rarement atteignable, avec comme conséquence un risque important de produire des faux-positifs et des classifications fluctuant au gré des découvertes.

La malédiction de la dimension

■ Les méthodes de classification peuvent être utilisées sur des jeux de données dont le nombre de dimensions tend à être de plus en plus important (cas des données de composition par exemple).

■ L'augmentation de la dimensionnalité s'accompagne d'une augmentation exponentielle du volume de l'espace considéré **(fig. C)**, si bien que les données sont éparpillées au point de rendre difficile l'identification de groupes.

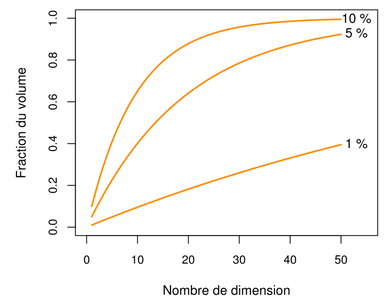


Figure C : pourcentage du volume de l'hypercube contenant 1 %, 5%, 10 % des données (cas d'une distribution uniforme).

A mesure que le nombre de dimension augemente, le volume nécessaire pour réaliser le même échantillonnage augmente de manière exponentielle.

Conclusion : une instruction à charge

■ Les méthodes de classification offrent un panel d'outils d'autant plus intéressant qu'ils relèvent d'un champ de recherche actif et que le volume de données disponibles tend à augmenter (malgré les problèmes d'échantillonnage propres à la donnée archéologique). Pour autant, les quelques problèmes soulevés ici ne constituent que le sommet de l'iceberg et de nombreux points demandent à être explorés, au-delà du seul choix de la mesure de proximité (ou de l'algorithme d'agglomération).

■ Enfin, la possibilité d'une non appartenance à une catégorie (et la prise en compte de l'inconnu) doit également être considéré (quelle serait la valeur d'une identification judiciaire construite sur la base du seul fichier des individus déjà connus ?).

[1] **Laboratoire Archéologie Prévention de l'Altération**
IRAMAT-LMC (UMR 5060) / NIMBE (UMR 3685)
Commissariat à l'Energie Atomique – Bât. 637
F-91191 Gif-sur-Yvette cedex
nicolas.frerebeau@cea.fr

[2] **IRAMAT-CRP2A (UMR 5060)**
Maison de l'Archéologie
Université Bordeaux Montaigne
F-33607 Pessac cedex

Les ouvrages suivants ont fortement alimentés cette reflexion :
- Banning, E. B. *The Archaeologist's Laboratory: The Analysis of Archaeological Data*. Springer, 2000.
- Bellman, R. E. *Adaptive Controle Processes: A Guided Tour*. Princeton University Press, 1961.
- Ginzburg, C. « Signes, traces, pistes – Racines d'un paradigme de l'indice. » *Le Débat*, vol. 6 (6), 1980.
- Lebart, L., Piron, M. & Morineau, A. *Statistiques exploratoire multidimensionnelle*. Dunod, 2006. Quatrième édition.
- S. Everitt, B., Landau, S., Leese, M. & Stahl, D. *Cluster Analysis*. Wiley, 2011. Cinquième édition.

Les auteurs remercient également M. Jean-Marc André, commandant à la Police Judiciaire de Bordeaux pour les échanges constructifs.

La présente étude a reçu un soutien financier de l'État géré par l'Agence Nationale de la Recherche au titre du programme *Investissements d'Avenir* portant la référence ANR-10-LABX-52.