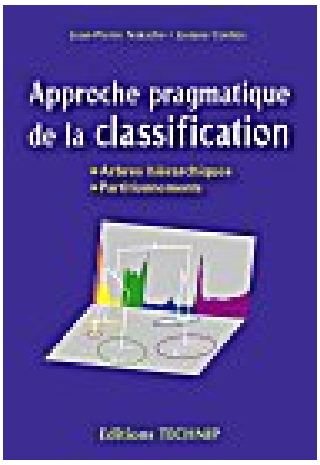
Note de lecture



Approche pragmatique de la classification.
par Josiane Confais et Jean-Pierre Nakache
Technip, 2004, 272 p., 49 €.

Les techniques de classification figurent parmi les méthodologies les plus sollicitées par des approches exploratoires tant pour les sciences du vivant que pour celles de la société. Les technologies pour l'extraction d'information s'appuient désormais largement sur de nouvelles approches algorithmiques développées pour la fouille d'entrepôts de données recelant des ensembles d'objets dont la cardinalité est très grande.

L'ouvrage que nous proposent Jean-Pierre Nakache et Josiane Confais, est le fruit de leurs expériences de formateurs, tant en formation initiale au sein de l'Institut de Statistique de l'Université Pierre et Marie Curie (ISUP), qu'en formation continuée dans le cadre des stages organisés par le CISIA-CERESTA. L'approche de l'ouvrage est effectivement pragmatique, s'appuyant sur de nombreux exemples, les logiciels les plus couramment utilisés en la matière ainsi qu'une bibliographie très riche à la fois au plan méthodologique mais également au plan applicatif. Les auteurs passent bien sûr en revue les techniques classiques proposées par les logiciels commerciaux (SAS, Splus, SPSS, Statistica) ou libres (R), mais donnent également un aperçu sur les développements les plus récents, par exemple certains algorithmes de classification conduisant à des classes non convexes de forme arbitraire (*shrinking*).

Après un chapitre introductif consacré aux mesures de la ressemblance ou de la dissemblance entre objets taxonomiques, la structure de l'ouvrage reprend la distinction communément admise pour les techniques de classification entre méthodes hiérarchiques et méthodes de partitionnement. La partie de l'ouvrage consacrée aux méthodes hiérarchiques traite de façon approfondie les algorithmes de type agglomératif (classification ascendante

hiérarchique) puis aborde de façon substantielle les algorithmes de segmentation utilisés dans l'approche divisive.

Après avoir présenté les concepts et algorithmes fondamentaux des méthodes agglomératives, l'ouvrage étudie les critères fondés sur la notion de métrique puis ceux basés celle de densité. La comparaison des résultats fournis par les méthodes hiérarchiques ascendantes s'appuie sur le concept d'ordonnance pour proposer une mesure de l'écart entre deux arbres hiérarchiques. Les développements récents concernant les méthodes agglomératives (algorithmes CURE, ROCK et BIRCH) et l'analyse des données spatiales concluent le premier chapitre.

Le chapitre suivant détaille la construction d'une hiérarchie par méthode agglomérative en privilégiant deux critères, celui de la perte d'inertie minimale (Ward) et celui du saut minimal (ultramétrique sous-dominante), Le premier critère est très utilisé en complément des méthodes factorielles, le second est plus connu pour ses propriétés mathématiques.

Le chapitre consacré aux méthodes divisives distingue les algorithmes non supervisés de classification, où toutes les variables jouent le même rôle, des méthodes de classification dite « supervisée » où à chaque niveau le processus de segmentation privilégie une variable en fonction d'une mesure d'homogénéité qu'il s'agisse du critère d'impureté de Gini (CART) ou du Chi-Deux (CHAID). Sont également présentés dans ce chapitre, les méthodes de classification de grandes collections de documents (algorithmes *PDDP*)

Le chapitre sur les méthodes de partitionnement, fournissant une partition déterminée des éléments à classer, présente les différentes variantes des méthodes de type *k-means* (centres mobiles et nuées dynamiques) et leurs extensions aux données catégorielles et mixtes. Sont également étudiées des adaptations de ces algorithmes, substituant à la notion de barycentre celle de représentant d'une classe (*k-medoids*) mais au prix d'une complexité algorithmique plus élevée.

Le cinquième chapitre est consacré aux méthodes de classification mixte (*hybryd clustering*) utilisant conjointement différentes techniques classifiantes pour proposer une stratégie unique de classification adaptée aux grands ensembles de données hétérogènes du point de vue du niveau de mesure (continu, ordinal ou nominal) de l'information statistique. L'ouvrage détaille l'utilisation des procédures de classification correspondante pour deux logiciels (SAS et SPAD) différant sensiblement dans leur conception.

Les procédures de fouille dans les entrepôts de données, regroupées sous le vocable générique de *Data Mining*, puisent largement dans les deux catégories de techniques de classification déjà évoquées. Cependant , elles ont également suscité des développements spécifiques à ce contexte applicatif qui sont fondés sur des concepts soit de densité (DBSCAN), de modèle d'organisation (approches neuronales) ou de quadrillage de l'espace (CLIQUE). Le chapitre 6 rend compte de ces développements en présentant rapidement les principaux algorithmes utilisés.

Le septième chapitre traite du nombre de classes à spécifier, question critique en classification tant du point de vue des méthodes hiérarchiques (choix du bon niveau de coupure) que des méthodes de partitionnement (choix du nombre de populations à identifier). Les indicateurs présentés permettent de déterminer la valeur de ce paramètre en optimisant un critère qui peut parfois s'interpréter en termes de compacité ou de séparabilité des classes au niveau de chaque partition.

Les techniques de caractérisation des classes, sont présentées dans le chapitre 8 tant du point de vue graphique que numérique. Y figurent les adaptations de techniques exploratoires développées en analyse des données multidimensionnelles pour les aides à l'interprétation, en particulier celles fondées sur l'approche empirique de valeur-test.

Le dernier chapitre présente quelques techniques de classification de variables parmi celles les plus couramment usitées avec un exemple permettant de comparer les résultats de la procédure VARCLUS (SAS) avec la variante (VARCAH) proposée par Quannari et Vigneau.

Des références logicielles et bibliographiques complètent un ouvrage que son approche effectivement pragmatique et sa pédagogie permettent de recommander à un large public, y compris aux étudiants.

Dominique Desbois