

ANALYSE BIBLIOGRAPHIQUE

Jean-Pierre Barthélemy, François Brucker, *Éléments de classification*, Londres, Hermès, 2007, 438p.

Les ouvrages francophones consacrés à la classification ne sont pas légion ! Le livre de J.-P. Barthélemy et F. Brucker, *Éléments de classification*, contribue à combler cette lacune.

Le domaine n'est pourtant pas nouveau en France, notamment depuis les travaux de G. L. Leclerc, comte de Buffon, comme le rappellent certaines références citées dans le chapitre 1 (lesquelles remontent même au XVII^e siècle, avec la classification des plantes proposée par J. Pitton de Tournefort, précurseur de C. von Linné). Comme l'annonce l'avant-propos, ce livre est consacré « aux approches relevant des mathématiques discrètes (combinatoire : graphes et hypergraphes, ensembles ordonnés et treillis finis, dénombrements, espaces métriques discrets), de l'algorithmique (NP-complétude, recherche d'instances polynomiales et problèmes polynomiaux), corrélativement à l'optimisation combinatoire (conception et codage en pseudo-langage d'algorithmes exacts ou heuristiques pour des problèmes d'optimisation, stratégies locales, reconnaissance de structures classificatoires, utilisation d'algorithmes d'optimisation classiques pour des problèmes de classification) ».

Ces différents thèmes constituent en effet les 438 pages de l'ouvrage, regroupées en huit chapitres :

- « Une introduction à la classification », 14 pages ;
- « Espaces de représentation », 58 pages ;
- « Classes et partitions », 50 pages ;
- « Modèles généraux », 44 pages ;
- « Hiérarchies », 50 pages ;
- « Hiérarchies faibles », 56 pages ;
- « Systèmes rigides sur un chemin ou un cycle », 64 pages ;
- « Représentations arborées », 58 pages.

Ces huit chapitres, de volumes très homogènes comme on le constate, sont précédés d'un avant-propos (6 pages) et suivis d'une bibliographie fournie (326 références, 14 pages) et d'un utile index (6 pages). Chaque chapitre contient des exercices (non corrigés et diversement répartis) permettant de mieux assimiler les notions introduites et les résultats obtenus.

Introduit par un savoureux extrait de *L'Enfance de Bécassine*, l'avant-propos précise, et on peut rendre hommage à cette honnêteté, que ces *Éléments* n'épuisent pas le sujet. En particulier, les sujets suivants ne sont pas traités dans le livre, du moins de manière exhaustive : la théorie du consensus, les approches probabilistes et statistiques, les problématiques et les pratiques issues de l'intelligence artificielle, les arbres de décision, les réseaux connexionnistes, le raisonnement par cas et les classifications incrémentales, les approches symboliques, l'application des métaheuristiques aux problèmes de classification.

De nature non technique, le chapitre 1 rappelle l'objectif global de la classification, résumé par une phrase de Buffon extraite de son *Histoire naturelle* de 1749 :

« Le seul moyen de faire une méthode instructive et naturelle est de mettre ensemble les choses qui se ressemblent et de séparer celles qui diffèrent les unes des autres ». Et les auteurs de montrer que cet objectif double est plus difficile à atteindre qu'il n'y paraît peut-être de prime abord. Ce chapitre retrace aussi, et on s'en réjouira, les principales étapes jalonnant le développement de la classification depuis l'Antiquité et évoque les nombreux domaines d'application de cette discipline : sciences de la nature, statistique, archéologie, sciences sociales, philosophie...

Avec le chapitre 2, on entre dans des considérations plus mathématiques et plus algorithmiques. Cinq points y sont essentiellement examinés, introduisant les outils qui seront exploités dans le reste du livre :

- les descriptions d'objets par des variables (binaires, numériques, ordinales) et les espaces de représentation associés ;
- les descriptions d'objets caractérisés par des mots définis sur un alphabet fini ;
- les comparaisons qualitatives d'objets et les espaces de représentation associés en faisant appel à des graphes ;
- les comparaisons quantitatives d'objets et les espaces de représentation associés en faisant appel à des dissimilarités ;
- les recodages non élémentaires dans un espace de dissimilarité.

On y trouvera en particulier les éléments nécessaires concernant les structures ordonnées (telles les relations d'équivalence), la théorie des graphes et les algorithmes de graphes de base, les dissimilarités, la théorie de la complexité algorithmique et de la NP-complétude (rappelons que l'un des auteurs, trop modeste pour se citer lui-même, est aussi coauteur d'un livre consacré à cette théorie, cf. [Barthélemy *et al.*, 1992]).

La classification fait un large usage de *partitions*. Le chapitre 3 leur est consacré. Les indices usuels permettant d'évaluer la qualité d'une classe y sont définis : diamètre, rayon, séparation, étoilement, indice de clique. Les considérations relatives aux partitions à proprement parler font l'objet de la partie 3.2 : treillis des partitions, dénombrement des partitions (nombres de Stirling de seconde espèce) et de la partie 3.3 : indices de qualité d'une partition (somme des diamètres, diamètre maximum, somme des indices de clique, somme des étoilements, maximum des rayons), comparaison de partitions. Les parties 3.5 et 3.6 posent la question algorithmique de savoir comment déterminer une classe ou une partition optimales pour un critère donné, question qui conduit, selon le critère, à des problèmes polynomiaux ou au contraire NP-difficiles (on y trouve par exemple, parmi de nombreux autres résultats, la preuve de la NP-difficulté du problème de C. T. Zahn [1964], consistant à approcher une relation symétrique par une relation d'équivalence minimisant la distance de la différence symétrique ; on peut en dériver la NP-difficulté du problème de S. Régner [1964], consistant à agréger un profil de relations d'équivalence en une relation d'équivalence minimisant la distance au profil initial, là encore pour la distance de la différence symétrique). Le chapitre se termine avec quelques méthodes génériques de partitionnement et avec quelques exemples de jeux de données.

L'étude des *systèmes de classes* (des hypergraphes), en particulier ceux issus de données décrites par une dissimilarité (un graphe), fait l'objet du chapitre 4. Après avoir défini ce qu'est un système de classes, les auteurs étudient plusieurs façons de

définir une classe à partir d'une dissimilarité ou d'un graphe, ainsi que les théorèmes de bijection classiques les liant. Ils présentent ensuite une façon alternative de définir les classes, la *binarisation*, et les systèmes associés, dits systèmes binaires. Enfin, la partie 4.5 s'intéresse à la notion de *rigidité*, déjà présente dans les travaux de C. Flament, dès 1962.

La structure de partition n'autorise pas l'emboîtement qui permet de distinguer divers niveaux (par exemple, en biologie, les niveaux de classe, d'ordre, d'embranchement, de genre ou d'espèce). Les *hiérarchies de parties* fournissent au contraire une telle possibilité, tout en conservant l'absence d'empiètement : deux classes sont soit emboîtées soit disjointes. Le chapitre 5 définit le concept de hiérarchie et la représentation graphique d'une hiérarchie, le *dendrogramme*, puis expose les principales propriétés combinatoires des hiérarchies et des indices permettant de les comparer. Il montre aussi que les hiérarchies sont en bijection avec certaines dissimilarités appelées *ultramétriques*. Il continue avec le problème (NP-difficile en général) de l'approximation d'une dissimilarité par une ultramétrique et avec une restriction pour obtenir un problème d'approximation polynomial (recherche de la sous-dominante). Il se termine avec deux heuristiques d'approximation efficaces : la classification ascendante hiérarchique et un algorithme avec garantie de performance.

Le chapitre 6 se concentre sur les *hiérarchies faibles* et sur les *quasi-hiérarchies*. Là aussi, il existe des théorèmes de bijection liant ces modèles à certaines dissimilarités, à savoir les ultramétriques faibles et les quasi-ultramétriques. De nouveaux problèmes NP-complets ou NP-difficiles y sont exposés. Les auteurs s'intéressent ensuite à l'approximation d'une dissimilarité quelconque par une dissimilarité d'un de ces types et analysent des algorithmes polynomiaux permettant d'établir plusieurs propriétés de ces modèles (liées aux *dissimilarités inférieures-maximales*, cf. la partie 6.5).

On revient, dans le chapitre 7, sur le thème de la rigidité introduit au chapitre 4. Mais il s'agit ici d'un cas particulier, celui des systèmes de classes rigides sur une chaîne ou un cycle, structures intéressantes par leur simplicité. Les auteurs y caractérisent les dissimilarités dont les boules, les 2-boules et les cliques maximales sont rigides sur une chaîne ou sur un cycle. Ils montrent que l'ensemble des dissimilarités dont le système de classes est rigide sur une chaîne admet des inférieures-maximales. Ils donnent un algorithme polynomial pour calculer celles-ci. Enfin, ils proposent une heuristique permettant d'approcher une dissimilarité quelconque par une quasi-ultramétrique dont les classes sont rigides sur un cycle.

L'étude de la rigidité se poursuit dans le dernier chapitre, mais maintenant pour des systèmes de classe rigides sur des arbres. Il y est établi que les quasi-ultramétriques rigides sur un arbre admettent des sous-dominantes faibles. Les auteurs donnent un algorithme permettant de calculer celles-ci. Ils s'intéressent ensuite (partie 8.2) aux *arbres phylogénétiques* ou *X-arbres* et décrivent un algorithme calculant un tel arbre (partie 8.3).

On le voit donc, ce livre couvre un large secteur des aspects mathématiques et algorithmiques de la classification. Il s'adresse à mon avis à tous les chercheurs et les doctorants travaillant sur des aspects mathématiques ou algorithmiques de la classification. Les auteurs étaient bien placés pour écrire une telle synthèse :

tous deux ont animé un groupe de recherche sur ce domaine à l'École nationale supérieure des télécommunications de Bretagne. Ils ont de plus contribué activement au développement de la Société francophone de classification (SFC), l'un (Jean-Pierre Barthélemy) en tant que vice-président puis président de la SFC, l'autre (François Brucker) en tant que rédacteur du bulletin de la SFC, de 2002 à 2006. Cette profonde connaissance de la discipline et leur implication directe dans ce domaine, en particulier comme auteurs de nombreux articles de recherche consacrés à cette problématique, leur ont permis d'être à la pointe du sujet et d'insérer des résultats récents sur de nombreux thèmes développés dans ce livre.

Cela contribue évidemment à l'intérêt de ce livre et lui permet de faire autorité en la matière. Ouvrage de référence, il mériterait à ce titre d'être traduit en anglais afin d'accroître son audience. Espérons que cela soit prochainement le cas !

BIBLIOGRAPHIE

BARTHÉLEMY J.-P., COHEN G., LOBSTEIN A., « Complexité algorithmique et problèmes de communications », Paris, Masson, 1992.

RÉGNIER S., "Sur quelques aspects mathématiques des problèmes de classification automatique", *I.C.C. Bulletin* 4, Rome, 1964.

ZAHN C.T., "Approximating symmetric relations by equivalence relations", *SIAM Journal on Applied Mathematics* 12, 1964, p. 840-847.

Olivier Hudry