Clustering: from model-based approaches to heuristic algorithms

Hans Hermann Bock[1]

Institute of Statistics, RWTH Aachen University,
D-52056 Aachen, Allemagne, bock@stochastik.rwth-aachen.de

Résumé. Les méthodes du 'clustering' ont pour but de diviser un ensemble (large) d'objets dans un petit nombre de groupes homogènes (clusters), basé sur des données relevées ou observées qui décrivent les (dis-)similarités qui existent entre les objets – en espérant que ces clusters soient utiles pour l'application concernée. Il existe une multitude d'approches, et cette contribution présente quelques-unes qui sont les plus importantes ou actuelles.

Les approches qui sont basées sur un modèle (model-based clustering) partent d'une vue probabiliste dans laquelle il existe une classification inconnue et les données sont des variables aléatoires dont la distribution dépend de la classe des objets correspondants. Nous présenterons les modèles 'fixed-partition', 'random-partition' et le modèle de mélange. Chacun mène à un critère de classification à optimiser. Nous esquissons des algorithmes, des propriétés mathématiques, et quelques cas spéciaux, mais importants.

Il est facile de définir des critères heuristiques de classification dans des cas où il n'y a pas un modèle probabiliste, et tandis que les méthodes précédentes se concentrent sur des classifications de type 'partition', on peut aussi construire des classifications hiérarchiques ou structurées. - Contrairement aux méthodes qui construisent une classification exhaustive pour l'ensemble total de tous les objets donnés, nous considérerons finalement le cas où on se contente à trouver seulement des (quelques) groupes singuliers et isolés d'objets qui sont bien plus similaires entre eux qu'en moyenne. Ces méthodes sont à la base de beaucoup d'applications en fouille des données (marketing, biotechnology, web logs).

[1] Etudes de mathématiques en 1958-1965 à Karlsruhe, Paris, Freiburg (diplome) ; positions universitaires aux universités de Freiburg, Hannover, et Aachen (Aix-la-Chapelle) ; Professeur en Probabilité et Statistique à Aachen depuis 1978 ; spécialités : analyse des données, clustering et classification, fiabilité; président de la Société Allemande de Classification (GfKl; 1986-1995), président de la International Federation of Classification Societies (IFCS; 1985-1987) ; Editeur de la revue 'Advances in Data Analysis and Classification (ADAC)' et de la série 'Classification, Data Analysis, and Knowledge Organization' (Springer Verlag).