Statistique et analyse des données

SOCIÉTÉ FRANÇAISE DE CLASSIFICATION

Résumés - Journées de Statistique, Nice 22-26 mai 1978

Statistique et analyse des données, tome 3, n° 2 (1978), p. 31-44

<http://www.numdam.org/item?id=SAD_1978__3_2_31_0>

SOCIETE FRANÇAISE
DE CLASSIFICATION

(Résumés - Journées de Statistique, Nice 22-26 MAI 1978)

SUR LES TYPES DE PARTITIONS

J.P. BARTHELEMY

E.N.S.C.M.B.
25030 - BESANÇON Cédex

L'étude des types de partitions (suite ordonnée des cardinaux des classes), c'est à dire des partages d'un entier peut être envisagée des points de vue suivants :

- Etude mathématique d'un "problème fondamental lié à la démarche du taxinomiste"
- Comparaison de classifications (menées concurremment) selon la taille et le nombre des classes indépendamment des éléments qui constituent ces classes.

C'est dans cette optique que nous nous proposons de présenter quelques résultats récents (essentiellement des propriétés métriques) sur les partages d'un entier.

L'ensemble P_n des partages de n est muni d'une relation d'ordre (non latticielle) pour laquelle il est modulaire. Cette remarque conduit à la construction de métriques qui s'interprètent en termes de graphes, certaines d'entre elles étant liées à la taxinomie et la théorie de l'information.

LA BI-CLASSIFICATION

G. BROSSIER

U.E.R. DES SCIENCES ET TECHNIQUES
UNIVERSITE DE HAUTE BRETAGNE - RENNES II

Les concepts de bi-classe et de bi-classification sont nés d'une observation lors de dépouillements d'analyses des données. Ces analyses factorielles mettent en évidence les oppositions entre objets placés aux deux extrémités d'un axe. Alors que les classifications mettent en évidence les ensembles d'objets qui se ressemblent.

Il est souhaitable d'avoir simulatanément en résultat d'une analyse de données les classes d'objets qui se ressemblent et les oppositions entre les classes. Le concept de bi-classe relève de cette analyse.

Une bi-classe est formée de deux classes qui s'opposent. C'est à dire que deux objets (individus, variables, questions...etc) appartenant à la même classe d'une bi-classe se ressemblent. Ceux qui appartiennent aux deux classes différentes d'une bi-classe s'opposent.

Le principe de la méthode est de rechercher directement dans un tableau de données les bi-classes. On traite ainsi simultanément les informations du type ressemblance et du type opposition.

Nous avons développé plusieurs types d'algorithmes permettant de rechercher une telle structure :

- une adaptation des algorithmes de recherche de partition à la recherche d'une bi-partition (type nuées dynamiques et transfert)
- une bi-classification hiérarchique qui agrège des bi-classes pour former des bi-classes
- une méthode qui utilise la notion de bi-classes floues pour fournir des familles de k bi-classes. Cette dernière méthode montrant l'équivalence entre un axe factoriel et une bi-classe.

CONSTRUCTION D'UNE ULTRAMETRIQUE LA PLUS PROCHE AU SENS DES MOINDRES CARRES APPROXIMATION VS OPTIMISATION

J.L. CHANDON

UNIVERSITE D'AIX MARSEILLE 3

I.A.E.

29, av. Robert Schumann

13670 - AIX en PROVENCE

Soit un ensemble X d'objets à classer et soit un indice de proximité d défini pour l'ensemble IP des paires d'objets de X. d est une application symétrique de IP dans R_+, indicatrice du degré de dissimilarité existant entre deux objets.

L'objectif de la classification est de simplifier l'information contenue dans le tableau des proximités en remplaçant les proximités entre paires d'objets par des proximités entre classes d'objets.

Nous nous intéressons à la construction d'une classification ascendante hiérachique (CAH) sur l'ensemble X. JOHNSON (1967) et BENZECRI (1973) ont démontré que pour obtenir une hiérarchie totale indicée sur X, il suffit de transformer l'indice de dissimilarité d en dissimilarité ultramétrique S.

L'écart entre deux dissimilarités d et S étant mesuré par la norme euclidienne $\| d - S \|$, le problème de la construction d'une ultramétrique optimale S^x minimisant $\| d - S \|$ a été résolu par CHANDON, LEMAIRE, POUGET (1978). Toutefois, l'algorithme de séparation ét évaluation progressive proposé est relativement lent. Il devient impraticable dès que le nombre d'objets est supérieur à 15.

Pour classer un grand nombre d'objets il est nécessaire de disposer d'un algorithme rapide permettant de construire ou d'approximer S^x. Deux algorithmes basés sur le concept de préordonnance sont proposés. Le premier, très rapide, ne garantit pas l'obtention de l'optimum S^x. Néammoins, il conduit toujour à une bonne approximation et il améliore toujours l'ultramétrique obtenue par l'algorithme de la moyenne de LANCE, WILLIAMS (1967), lorsque celle ci n'est pas optimale. Le second, moins rapide, améliore encore cette estimation.

Les deux algorithmes sont appliqués aux données classiques de RAO (1952), FITCH, MARGOLIASH (1967), KAMEN (1971) et MILLER, NICELY, afin d'illustrer l'amélioration du critère des moindres carrés obtenue par rapport à plusieurs autres méthodes de classification hiérarchique.

RECONNAISSANCE DES FORMES PAR UNE METHODE D'APPROXIMATION POLYNOMIALE A NOEUDS VARIABLES

C. CHARLES

UNIVERSITE PARIS IX

LASADE

Place de Lattre de Tassigny - 75016 PARIS

Y. LECHEVALLIER

IRIA-LABORIA

Domaine de Voluceau - 78150 - LE CHESNAY

Lorsque nous désirons mémoriser de façon condensée l'information fournie par une fonction g(x) échantillonnée en différents points d'un intervalle $[a,b]$, nous cherchons une fonction dépendant d'un petit nombre de paramètres qui approxime au mieux g(x), au sens des normes L_2 et L_∞. En pratique, les fonctions les plus utilisées sont les polynomes ; toutefois, ceux-ci ne peuvent avoir de pentes infinies suivies de pentes nulles sur de courts intervalles si leurs degrés sont faibles, ce qui est nécessaire pour éviter les instabilités numériques. Nous avons donc choisi une approximation par des fonctions polynomiales par morceaux. Ces fonctions sont définies par division de l'intervalle $[a,b]$ en sous intervalles dont les limites sont variables. Bien que la fonction approximante soit non linéaire et qu'on ne possède de solution générale, ce problème est de grande importance puisque la clé du succès d'une approximation locale réside dans l'emplacement de ces points limites.

Les algorithmes que nous présentons emploient de façon simultanée la classification automatique (algorithme de transfert, algorithme de Fisher) et les techniques de régression (norme L_2) ou de résolution d'un système linéaire surdimensionné au sens de Chebyshev (norme L_∞).

Ces algorithmes présentent l'avantage de pouvoir être aisément généralisés aux cas de contours et permettent d'introduire des contraintes de continuité, de dérivabilité (fonctions splines) aux points limites, donnant ainsi ces propriétés à la fonction approximante globale sur $[a,b]$.

UNE METHODOLOGIE DE CLASSIFICATION : L'ANALYSE PAR CRITERES ORDONNES SUR UN ENSEMBLE MUNI D'UNE RELATION DE PROXIMITE.

C. COCHET

IRIA, CEPIA
Domaine de Voluceau
BP 105
78150 LE CHESNAY

Nous proposons, dans le cadre de cette communication, une méthodologie d'approche, "l'analyse par critères ordonnés sur un ensemble muni d'une relation de proximité" (APCO), qui vise des problèmes se posant en des termes complexes à propos du morcellement de certains systèmes. Le déroulement d'une méthode qui permet de discerner des sous parties à l'intérieur d'un système peut prendre deux formes essentielles.

Dans un premier cas il est possible d'identifier un procédé de morcellement dont les résultats seront conformes aux spécifications auxquelles doivent répondre les sous-systèmes discernés. La méthode qu'il faut mettre en oeuvre se réduit alors au procédé qui est ainsi mis en évidence.

Dans tous les autres cas, il n'est pas possible de découvrir une procédure unique qui satisfasse, en fonction des spécificités du domaine étudié, les besoins de l'analyste. L'investigation scientifique auquel est soumis le système sera composé de plus d'un procédé de morcellement.

Nous avons choisi d'établir la synthèse des résultats de morcellement en exploitant les partitions obtenues pour chaque procédé utilisé. Ce moyen autorise l'usage d'une structure mathématique riche : le treillis géométrique des partitions.

Un ensemble de procédures méthodologiques doivent être utilisées pour mener à bonne fin l'analyse des problèmes qui sont soulevés ici, c'est à dire ceux justifiables de différentes analyses de morcellement.

Dans un premier temps, chaque analyse de classification doit être appliquée au champs d'observation du système qui la concerne. Pour développer cette opération, il faut disposer d'une théorie des systèles dont les concepts soient suffisamment évolués afin de rendre possible cette procédure.

Ensuite, nous avons établi une méthodologie qui permet d'envisager une synthèse des différents morcellements utilisés. Pour atteindre cet objectif, nous avons exprimé les résultats des procédés de morcellement sous forme de "proximités" affectant les points faisant l'objet de cette étude.

Tous les résultats obtenus, sous forme de partition des parties morcellées sont alors synthétisés dans une analyse unique qui est le réflet des différents points de vue présents dans l'élaboration de la stratégie de classification. On débouche alors, par l'exploitation des résultats centraux de cette analyse et des produits intermédiaires, sur des quantifications qui permettent de mesurer l'intervention des éléments de la stratégie et même, à la limite, de mettre en évidence certaines composantes qui sont le résumé d'autres, plus complexe.

ANALYSE ET CLASSIFICATION DE SPIRALES OBTENUES A PARTIR DE DOCUMENTS ARCHEOLOGIQUES

J.F. DESNOS

CENTRE DE CALCUL INTERUNIVERSITAIRE DE LYON

R. FAGES

UER Mathématiques - UNIVERSITE DE LYON

Il s'agit de la recherche d'une classification chronologique et géographique de stèles thessaliennes (cf [1]), dont un des ornements principaux consiste en deux spirales symétriques. Le mode de construction du tracé de ces spirales définit un important critère de classement des monuments, permettant de caractériser les ateliers et de suivre les déplacements géographiques des artisans, en relation avec des situations historiques connues.

Les questions qui nous ont été posées par l'équipe de recherche en Archéologie sont les suivantes :

. Peut-on distinguer plusieurs techniques de construction de spirales à partir des documents photographiques dont on dispose ?

. Si oui, peut-on définir une classification conforme de ces spirales et représenter chaque classe par une construction type ?

Dans une première phase (cf [2] et [3]) nous avons cherché à affecter l'ensemble des spirales à trois types prédéfinies : "développante de cercle", "à centres, sommets d'un polygone régulier" et "exponentiel".

Une analyse plus élaborée nous a conduit à étudier la variation du rayon de courbure le long de la spirale et à chercher une meilleure approximation des différentes modalités de cette dernière, en faisant appel à des fonctions de la forme $R = f(\theta)$, où (R, θ) sont les coordonnées polaires. Trois familles de fonctions ont été retenues :

$R = a\theta + b$, R en escalier, $R = b \exp(a\theta)$.

Nous cherchons maintenant à améliorer la finesse de l'approximation et à augmenter le nombre de types de fonctions d'ajustement. Ceci devrait nous permettre, d'après nos dernières constatations d'accroître sensiblement le nombre de spirales reconnues.

ANALYSE CANONIQUE OU POINT DE VUE DE LA CLASSIFICATION AUTOMATIQUE

E. DIDAY

UNIVERSITE PARIS IX-DAUPHINE
Place de Lattre de Tassigny
75775 - PARIS Cédex 16

Quand le tableau des données est de grande taille, il est légitime de chercher des combinaisons linéaires dépendant des tendances locales qui peuvent apparaître dans la population. Il s'agit de détecter ces tendances et simultanément les composantes canoniques qui leur sont le mieux associées.

Suivant que les données sont centrées ou non on propose plusieurs algorithmes qui tendent à minimiser le critère. Dans le cas où toutes les variables sont qualitatives, le problème se pose en terme d'analyse factorielle des correspondances et revient à minimiser le critère. Dans le cas où toutes les variables sont qualitatives, le problème se pose en terme d'analyse factorielle des correspondances et revient à chercher les classes d'objets qui induisent les plus grands X^2 de contingences entre les variables. Si l'un des deux paquets de variables est formé de variables d'incidence on aboutit à des méthodes intéressantes d'analyse discriminante locale.

LA NOTION DE DISPERSION EN CLASSIFICATION AUTOMATIQUE

R. FAGES

UER de Mathématiques
UNIVERSITE DE LYON I

Alors que généralement les méthodes de classification automatiques sont fondées sur la notion de dissimilarité entre deux individus de la population à traiter, il est proposé ici une approche ensembliste.

A toute partie A de la population est associée une mesure de dispersion D_A caractérisée par des axiomes dont le plus essentiel est la surraddi-tivité : $D_{A\cup B} \geq D_A + D_B$ si $A \cap B = \emptyset$.

Différents cas sont passés en revue montrant comment construire des dispersions avec les types de variables habituellement rencontrés. Cette dispersion permet de construire des critères de classification fixés ou adaptatifs (notamment la pondération automatique des variables). Il est alors possible de généraliser des algorithmes connus en les plaçant dans un cadre commun.

Enfin est présentée une méthode de classification originale Méthode Non Hiérarchique Descendante(+) particulièrement adaptée à la notion de dispersion. Cette méthode recherche les meilleures partitions en 2,3,4... etc classes sans imposer de structure arborescente. Si celle-ci est découverte, elle résulte alors d'une structure propre aux données et non d'une distorsion apportée par la méthode.

(+) les programmes de calcul automatique TAXI et TACO utilisés avec succès depuis 1974, par différents laboratoires de l'Université de LYON résulte de cette méthode.

R.FAGES 41D rue Phélypeaux 69100 Villeurbanee. Tél.(78)52.04.37

SELECTION ET DISCRETISATION OPTIMALES DE VARIABLES CONTINUES EN VUE D'UN PROBLEME DE RECONNAISSANCE DE FORMES

R. FAGES

UER de Mathématiques
UNIVERSITE DE LYON I

Lorsque les variables sont en partie ou totalité des variables continues, on se ramène à des variables discrètes (ou de classification) par le choix arbitraire de seuils afin d'utiliser les avantages de processus interrogatifs, comme par exemple les pseudo-questionnaires.

Il est abordé ici le problème du choix optimal pour chacune des variables continues, du nombre de classes et des seuils correspondants, afin de minimiser la probabilité d'erreur de l'identification par la règle de décision de BAYES.

La sélection des variables les plus discriminantes s'en déduit naturellement par le rejet des variables discrétisées par une seule classe. (Cette sélection est étendue aux variables initialement qualitatives).

La technique proposée utilise la majoration la plus fine de la probabilité d'erreur par une mesure d'entropie dérivant du coefficient de BHATTACHARRYA (+).

Un exemple concret est présenté, montrant l'efficacité des sélections et discrétisations obtenues, même dans le cas où les hypothèses nécessaires à la justification du critère d'optimisation ne sont pas vérifiées.

(+) M. TERRENOIRE, D. TOUNISSOUX. "Inequalities using BHATTACHARRYA distance and application to decision process".

3 rd International Joint Conference on Pattern Recognition, Coronado November 1976.

UN NOUVEAU TEST D'UNIFORMITE CONTRE UNE HYPOTHESE ALTERNATIVE UNIMODALE

L. FARINAS del CERRO

et

W. FERNANDEZ de la VEGA

Centre National de la Recherche Scientifique
LABORATOIRE D'INFORMATIQUE
POUR LES SCIENCES DE L'HOMME
31,chemin Joseph Aiguier 13274 Marseille Cedex 2
Tél:(91) 75.90.42

Si F(x) est la fonction de répartition d'une variable aléatoire unimodale dont le support est inclus dans $[0,1]$ il est facile de voir que les points u et v auxquels la différence F(x)-x atteint ses valeurs minimale et maximale respectivement se suivent dans cet ordre. Cette remarque suggère l'utilisation, pour tester l'uniformité contre une hypothèse alternative unimodale, de la statistique

$$S = \mathrm{Max}_{0<a<b<1} \left[a - F^{x}(a) + F^{x}(b) - b \right]$$

où $F^{x}(.)$ désigne la fonction de répartition empirique observée.

On a déterminé, en utilisant une méthode de Monte-Carlo avec 30 000 échantillons simulés indépendants, des seuils approchés pour S, pour des effectifs égaux à 10, 20 et 40. Des calculs de puissance, effectués pour des alternatives du type

$$F(z) = \frac{z^{k}}{a^{k-1}} \quad , \quad z<a$$

$$F(z) = 1 - \frac{(1-z)^{k}}{(1-a)^{k-1}} \quad , \quad z>a$$

indiquent que S se compare favorablement aux statistiques non-paramétriques classiques

DONNEES QUANTITATIVES INCOMPLETES ET CLASSIFICATION

P.P. FEVRE

IRIA - LABORIA
Domaine de Voluceau
78150 - LE CHESNAY

Lorsque des variables sont observées sur une population, il est fréquent que, pour certains individus, des variables ne soient pas relevées.

Pour traiter de telles données, la plupart des auteurs cherchent à "reconstituer" tout d'abord les données non disponibles, afin de pouvoir appliquer, sur ces données complétées, les méthodes usuelles d'analyse.

Le point de vue adopté ici est différent : nous cherchons à mettre en oeuvre directement les méthodes classiques de traitement des données, en ne tenant comote que des observations connues, et sans chercher à reconstituer les observations manquantes.

Pour ce faire, nous calculons, à partir des données disponibles, des approximations des quantités nécessaires au traitement habituel et nous travaillons en nous servant de ces approximations.

En application, afin de montrer ce que permet cette méthodologie, nous montrons comment une méthode de classification, la méthode de Nuées Dynamiques, peut être mise en oeuvre sur des données quantitatives incomplètes.

AMELIORATION D'UN ALGORITHME DE CLASSIFICATION HIERARCHIQUE RAPIDE

M. GRAF-JACCOTTET

UNIVERSITE DE NEUCHATEL
Av. du 1er Mars 26
2000 - NEUCHATEL - SUISSE

HARTIGAN (1975) propose un algorithme de classification hiérarchique rapide "Quick Tree Leader", n'exigeant qu'une lecture séquentielle du tableau des données, qui n'est pas supposé être en mémoire. Il s'agit d'une méthode de type "leader", construisant de proche en proche et en un seul passage à travers les données un arbre hiérarchique non binaire dont les niveaux sont fixés à priori. Le nombre de terminaux et de noeuds de l'arbre n'est donc pas connu à l'avance.

Considérant un arbre non binaire comme un arbre binaire dont certains niveaux sont confondus, nous avons pu améliorer les performances de l'algorithme en introduisant la structure "ainé-benjamin". Ceci permet d'éviter de définir le nombre N de noeuds de l'arbre et de remplacer deux tableaux de dimension N par deux tableaux dimensionnés au nombre maximum de terminaux.

Réf. HARTIGAN J.A. Clustering Algorithms. Chap. 9 (1975) Wiley.

ANALYSE CLASSIFICATOIRE D'UN TEST SCOLAIRE

R. GRAS

Département de Mathématiques et I.R.E.M. de RENNES
Campus de Beaulieu
35042 RENNES CEDEX

Cette communication présente quelques résultats didactiques d'une analyse en classification hiérarchique d'un test mathématique proposé à près de 1 100 élèves de 13 à 15 ans (fin de 3ème de C.E.S.). La classification C1, obtenue à l'aide de l'algorithme de la vraisemblance du lien de I.C. Lerman, est comparée à 2 autres classifications :

- classification C2 par rapport au contenu et à la nature de la tâche de l'item
- classification C3 par rapport à une taxinomie d'objectifs cognitifs de R. Gras.

L'hypothèse d'indépendance entre C1 et C2, puis C2 et C3 est rejetée par un test du χ^2, au seuil de $1^{o}/_{oo}$.

Les 5 classes de C1 conduisent à des interprétations confirmées par une analyse factorielle des correspondances :

- classe α de découverte de règle de production logique
- classe β de fonctions numériques et géométriques
- classe γ de nature numérique, très didactique
- classe δ d'observations de propriétés affines de l'espace
- classe ε d'observations de propriétés métriques de l'espace.

Les niveaux cognitifs croissent puis décroissent de α à ε, en passant par un maximum dans γ. Les classes $\{\delta, \varepsilon\}$ et $\{\alpha, \beta, \gamma\}$ sont relatives aux deux derniers stades de développement cognitif selon Piaget : "opérations concrètes" et "opérations logico-formelles".

PRESENTATION D'UNE METHODE SIMPLE DE DISCRETISATION DE VARIABLES CONTINUES

J.Y. LAFAYE

I.U.T. de VANNES
Rue Montaigne BP 1104
KERCADO 56008 VANNES

Le codage de variables continues préalablement à une analyse de données pose d'une façon générale le problème de la segmentation de l'intervalle de variation de telles variables en sous-intervalles logiques sur lesquels la densité soit unimodale.

Dans le cadre des procédés graphiques visant la décomposition de mélanges de lois de probabilités, on présente une méthode simple basée sur la définition d'une fonction de densité discrète permettant sous des hypothèses très larges la séparation en composantes homogènes.

On rend finalement compte des résultats obtenus dans le domaine médical ainsi que sur des mélanges de lois de probabilité obtenus par simulation.

CLASSIFICATION VISUALISEE DE GRANDS ENSEMBLES SOUS DOUBLE CONTRAINTE

L. LEBART

C.N.R.S.
CREDOC - 140, rue du Chevaleret
75013 - PARIS

C. ROCHE

S.P.A.E.
Direction des Télécommunications

La procédure présentée répond aux préoccupations suivantes :

Construire une partition d'objet respectant une contrainte de contiguité géographique (un zonage) et dont les effectifs des classes soient bornés par une quantité imposée ; faciliter au maximum la compréhension du programme en multipliant les aides à l'interprétation, en introduisant de nouvelles procédures de visualisation, de façon à permettre une utilisation de routine aisée de la procédure ; assurer un encombrement mémoire réduit et une exécution rapide.

L'algorithme de base de construction d'une classification ascendante hiérarchique adaptée au sous-ensemble réactualisé des couples d'objets contigus n'est pas original (cf par exemple la thèse de A. THAURONT, PARIS, 1975).

La matrice de contiguité ne figure ici que sous la forme de tableau de codage réduit (pour chaque sommet du graphe : adresse des sommets adjacents). C'est sous cette forme qu'elle est actualisée après chaque agrégation, et après chaque intervention du seuil de taille maximale des classes. C'est également à partir de cette forme condensée que cette matrice est soumise à une analyse des correspondances impliquant une technique de diagonalisation particulière, de façon à faire apparaître sur l'imprimante une reconstitution de la carte géographique, sur laquelle seront positionnées les différents noeuds de l'arbre intermédiaire, puis les classes finales. Le principe de cette dernière opération permet d'analyser en quelques secondes des matrices binaires clairsemées d'ordre 1000x1000. Une analyse des correspondances classique effectuées cette fois sur le tableau de données de départ permet également de suivre les évolutions des noeuds et la position des classes non plus dans l'espace géographique, mais dans l'espace des variables. On a ainsi tous les éléments pour suivre et comprendre les mécanismes de formation des classes et le caractère plus ou moins prégnant des contraintes.

SELECTION D'UN JEU DE CARACTERES PAR UNE METHODE DE CLASSIFICATION AUTOMATIQUE

Y. LECHEVALLIER

IRIA-LABORIA
Domaine de Voluceau - 78150 LE CHESNAY

C.Y. SUEN

Dépt. of Computer Science
CONCORDIA UNIVERSITY
1455 de Maisonneuve blvd West - MONTREAL QUEBEC H3G 1M8

L'ensemble des lettres à analyser comprend les 26 lettres de l'alphabet et les 10 chiffres. Pour chaque lettre (ou chiffre) nous avons recueilli plusieurs caractères typographiques.

Le but de notre analyse est d'essayer de trouver, pour chaque lettre, un caractère de notre ensemble, représentant au mieux celle-ci mais assez éloigné des caractères des autres lettres.

La première étape de notre analyse est une étape de description. Elle consiste à associer à chaque caractère un vecteur de R^n. Cependant le problème est insoluble si la description des individus (ici les caractères) ne comporte pas les traits pertinents permettant de les différencier d'où la nécessité préalable du codage. L'image de chaque caractère étant digitalisée en 29 colonnes et 39 lignes, le codage que nous avons adopté consiste à quadriller l'image du caractère en intervalles ou en fenêtres plus ou moins grands.

La deuxième étape de notre analyse consiste à trouver une règle simple d'affectation décidant, d'après sa description (dans R^n) de l'appartenance ou non d'un nouvel individu à l'une des classes. Cette procédure nécessite la définition de deux fonctions, l'une est une fonction d'agrégation, l'autre une fonction d'écartement.

ARBRES VALUES ET ULTRAMETRIQUES

B. LECLERC

C. M. S.
54, bd Raspail
75270 - PARIS Cédex 06

Il est maintenant bien connu que tout arbre (graphe connexe et sans cycle) valué, défini sur un ensemble X de cardinal n, induit naturellement une ultramétrique r sur X, donc une classification hiérarchique sur X. Quelques travaux ont commencé à paraître, cherchant à étudier l'ensemble des ultramétriques ainsi définies à partir de tous les arbres valués donnés par un indice de distance d sur X. Inversement, on a aussi posé le problème suivant : soit r une ultramétrique sur X : représenter r par un arbre (Benzécri et Jambu, 1976).

Après avoir rappelé les résultats antérieurs, nous précisons le lien entre arbres valués et ultramétriques. On établit d'abord que celles-ci se caractérisent, parmi les indices de distance, par des propriétés où interviennent uniquement leurs arbres minimaux. Mais, s'il est vrai qu'une ultramétrique r est parfaitement définie par l'un quelconque de ses arbres minimaux (valué par la restriction de r), le nombre N(r) de ceux-ci est compris, dans le cas général, entre $4^{n-1}/n^2$ et (n-1)!. Ceci pose le problème du choix d'un arbre particulier pour représenter r, qui n'a pas de réponse évidente que lorsque r a été obtenue à partir d'un arbre valué lisible directement dans les données (c'est le cas dans certaines méthodes classificatoires : lien simple et lien complet).

On s'intéresse ensuite aux propriétés du nombre N(r) qui paraît être un descripteur intéressant de la classification hiérarchique associée à l'ultramétrique r.

CLASSIFICATION DE GRANDS ENSEMBLES DE DONNEES PAR LA METHODE "SINGLE-LINK"

Ph. LEHERT

FACULTE UNIVERSITAIRE CATHOLIQUE DE MONS (BELGIQUE)
FACULTE DE DROIT ET DE SCIENCES ECONOMIQUES DE LILLE

P. HANSEN

FACULTE UNIVERSITAIRE CATHOLIQUE DE MONS (BELGIQUE)
INSTITUT D'ECONOMIE SCIENTIFIQUE ET DE GESTION DE LILLE

La méthode "single-link" ou de "l'ultramétrique inférieure maximum" est très utilisée en classification numérique. Lorsque les ensembles de données à classer sont grands, une méthode de calcul, basée sur la recherche de l'arbre minimum d'un graphe, exige de O (n) emplacements en mémoire et O (n^2) opérations où n désigne le nombre d'objets à classer. Si les dissimilarités entre objets à classer sont des distances de Minkowski, l'arbre minimum peut être calculé sans que toutes les dissimilarités ne le soient. On présente un algorithme très rapide pour ce problème, ainsi que les résultats d'une classification d'un ensemble de 10 000 étoiles.

LA METHODE DES POLES D'AGREGATION

H. LEREDDE

Département de Mathématiques
Centre Scientifique et Polytechnique
UNIVERSITE PARIS-NORD (XIII)
Av. J.B. Clément
93430 - VILLETANEUSE

Le principe général de cette méthode de classification non hiérarchique est de constituer pas à pas des agrégats parmi les éléments à classer.

Pour former un agrégat nous choisissons un élément "pertinent", appelé pôle, auquel nous agrégeons les éléments qui lui sont le plus proche, fonction d'un certain seuil. Cet agrégat constitué, nous recommençons la même opération avec les éléments non encore agrégés, et ceci tant qu'il reste des éléments à classer. A la fin nous obtenons une partition de l'ensemble des éléments, chaque classe de la partition étant l'un des agrégats (pour l'idée de base de cet algorithme, on pourra consulter I.C. LERMAN, reconnaissance et classification de structures finies en analyse des données, Université de RENNES I, 1977, rapport IRISA n° 70).

Nous avons mis en oeuvre cette méthode sous forme d'un algorithme appelé MPAGD (Méthode des Pôles d'Agrégation sur les Distances) où l'on étudie la distribution des distances entre éléments à classer.

La principale difficulté de ce type d'algorithmes est de parvenir à un système permettant d'arrêter la formation d'un agrégat que l'on puisse appliquer indépendamment de la distribution des distances et de la nature des données. Cet algorithme est une tentative dans cette voie.

LES PRESENTATIONS FACTORIELLES DE LA CLASSIFICATION

I.C. LERMAN

Laboratoire de Statistique - I.R.I.S.A.
UNIVERSITE DE RENNES I - B.P. 25 A
35031 - RENNES Cédex

Différentes tentatives de rapprochement entre l'Analyse Factorielle et la Classification ont été proposées : N. HOWARD (1969), J.P. BENZECRI (1971), M. GONDRAN (1975), M. JAMBU suivant J.P. BENZECRI (1976). Toutes ces tentatives correspondent en fait, comme nous le verrons ci-dessous à des présentations "factorielles" du problème de la Classification. La tentation de telles présentations est en effet grande compte tenu du caractère plus établi de l'histoire de l'Analyse Factorielle en Composantes qui fournit la solution optimale pour le critère de l'inertie expliquée.

Notre but dans cet article est d'analyser chacune de ces tentatives pour préciser sa véritable nature, la généraliser et l'adapter à des situations nouvelles. Ce faisant, nous contribuerons par des résultats nouveaux à chacune des approches en les situant, dans un effort de synthèse, les unes par rapport aux autres. Ce qui nous permettra de nous rendre compte de l'intérêt relatif de ces différentes tentatives.

Les différents types de rapprochement entre l'Analyse Factorielle et la Classification se distinguent en ce que certains traitent du problème de la recherche d'une classification et que d'autres traitent de celui de la recherche d'un arbre binaire des Classifications. Ils se distinguent également de par la nature du critère optimisé : deux critères seront considérés ici ; le premier est basé sur l'inertie expliquée (i.e. variance) et le second sur la notion de proximité entre parties disjointes au sens topologique du terme (i.e. saut minimum). Ces différentes approches se distinguent encore par la forme de l'équation factorielle retenue : s'agit-il de rechercher un système d'axes factoriels dans l'espace de représentation du nuage ou bien, une suite de facteurs dont chacun se trouve défini comme une fonction sur l'ensemble des sommets auquel il est relatif. Ces approches se distinguent enfin de par la nature de la structure projective de l'espace engendré par la solution de l'équation factorielle retenue.

CONTRIBUTION DE LA THEORIE DES GRAPHES A LA TAXINOMIE MATHEMATIQUE

B. MONJARDET
UNIVERSITE DE PARIS V
Centre de Mathématique Sociale (E.H.E.S.S.)

Il est bien connu que deux méthodes classiques en classification hiérarchique, celle du "lien simple" (ou de l'ultramétrique sous-dominante) et celle du "lien complet" s'interprètent aisément en termes de théories des graphes. Les classes des partitions de l'arbre hiérarchique sont en effet, dans le premier cas, des classes connexes, dans le second cas, des cliques, des graphes "seuils" associés à la dissimilarité considérée. La considération de ces graphes seuils ajoutée à une condition de cohérence, ramène le problème de la classification hiérarchique à celui de la classification des sommets d'un graphe. A cet égard, les classements formés par les classes connexes d'une part, des cliques d'autre part, apparaissent comme deux solutions extrêmes maximisant respectivement un critère de séparation entre classes et un critère d'homogénéité à l'intérieur des classes. Mais en fait, la théorie des graphes permet d'envisager bien d'autres possibilités intermédiaires entre les deux solutions extrêmes. Ces possibilités sont d'ailleurs apparues dans des contextes variés. Par exemple, l'analyse de réseaux sociométriques et la recherche de leur décomposition en groupements homogènes amène à définir des concepts de "cliques généralisées" d'un graphe (Luce, 1950). Inversement, des études théoriques sur le nombre chromatique ou la connectivité d'un graphe conduisent à des notions pouvant être utiles en classification.

Le but de l'exposé est de présenter ces apports de la théorie des graphes à la taxinomie mathématique, qu'il s'agisse ou non de classification hiérarchique. La littérature sur le sujet étant déjà fort vaste (plus de 150 références), on ne tentera pas d'être exhaustif, mais on essaiera d'indiquer les principales contributions, en distinguant les apports conceptuels de ceux plus techniques.

UNE NOUVELLE METHODE DE CLASSIFICATION BASEE SUR UN BALAYAGE DE PEANO

J. QUINQUETON
IRIA-LABORIA
Domaine de Voluceau
78150 - LE CHESNAY

Un problème de reconnaissance des formes consiste à trouver une application d'un ensemble U (l'Univers) dans un ensemble F (les Formes). Généralement c'est un problème n-dimensionnel.

En analyse d'image, par exemple, un adressage des points de l'image qui respecterait leur voisinage permettrait de transformer un problème bidimensionnel en un problème unidimensionnel.

En classification automatique non hiérarchique, le problème consiste à partitionner un ensemble E, de points appartenant à R^n, en fonction de leurs interdistances ou, plus généralement, de leur voisinage. Dans ce cas, si nous trouvons une application T, de R^n dans R^q, avec $q < n$, qui respecte le voisinage des points, nous transformons ce problème n-dimensionnel en problème q-dimensionnel.

Une application réduisant la dimension d'un problème est donc très intéressante en classification. L'idée qu'une courbe du type de celle de Péano (courbe remplissant un carré) peut définir une telle application fut suggérée par le Professeur ALEXANDROV, de LENINGRAD.

Le principe est le suivant : divisons $[0,1] \subset R$ en 2^p intervalles égaux, qui nous donnerons une division de $[0,1]^n \subset R^n$ en 2^{np} hypercubles H. Un adressage de ces hypercubes suivant une courbe de Péano permet de les explorer séquentiellement, en respectant presque partout leur voisinage. Deux adressages de ce type, P et $\tilde{P}$, l'un translaté de l'autre d'un vecteur diagonal de R^n, dont les coordonnées sont égales à $\frac{1}{2^p}$ (la diagonale d'un hypercube H), permettent de respecter le 2n-voisinage. Cette propriété est utilisée pour partitionner un ensemble de points en classes, sans avoir à calculer de distances.

Des exemples sont donnés sur les données de Ruspini (n=2) et sur les données IRIS (n=4).

CLASSIFICATION DE MAXIMUM DE VRAISEMBLANCE DE DONNEES BINAIRES

P. ROUSSEAU

Centre de Recherches Mathématiques
UNIVERSITE DE MONTREAL
Casier Postal 6128
MONTREAL (CANADA)

Une famille d'ensembles de données binaires est par hypothèse représentée par K modèles logistique-linéaires où chaque modèle représente une sous famille de cette famille.

On définit un algorithme qui trouve simultanément la partition de maximum de vraisemblance en K sous familles et les estimateurs de maximum de vraisemblance des paramètres de chaque modèle.

Cet algorithme est une extension des méthodes de partition itérative selon l'approche de Diday.

ETUDE DE L'AGRICULTURE REGIONALE FRANÇAISE PAR UNE METHODE DE CLASSIFICATION AUTOMATIQUE

B. TALLUR

Laboratoire de Statistiques
U.E.R. de Maths et Informatique
UNIVERSITE DE RENNES I

On cherche à classifier l'ensemble des départements français de façon à dégager les principaux types d'agriculture, en utilisant l'Algorithme de la Vraisemblance du lieu dO à I.C. LERMAN. L'économiste a retenu trois caractères descriptifs ; chaque département est en effet caractérisé par :

1 - surfaces par principales cultures (6 modalités),

2 - importance du cheptel (3 modalités),

3 - structure d'exploitation (8 modalités).

On a défini un indice de proximité entre les lignes (resp. les colonnes) d'un tableau de contingence, conformément à la classe des indices de I.C. LERMAN et respectant la métrique du χ^2. Cet indice est ensuite généralisé au cas où les données sont une juxtaposition de plusieurs tableaux de contingence ayant, tous, le même ensemble de lignes (ou de colonnes). Ceci a rendu possible l'application de l'A.V.L. pour obtenir la classification de l'ensemble des départements français selon un seul caractère, dans un premier temps, et selon les deux ou trois caractères réunis, dans un second.

Nous avons aussi appliqué, à ces données, le programme de M. JAMBU de la Classification Ascendante Hiérarchique utilisant la distance du χ^2 et l'agrégation par la maximisation, à chaque pas, du moment centré d'ordre 2 de la nouvelle classe formée par réunion de deux classes. Les résultats issus des deux algorithmes sont comparés dans les deux cas :

a) classification basée sur un seul caractère (structure d'exploitati

et

b) classification basée sur tous les trois caractères.