REVUE DE STATISTIQUE APPLIQUÉE

MAURICE ROUX

Deux algorithmes récents en classification automatique

Revue de statistique appliquée, tome 18, nº 4 (1970), p. 35-40

<http://www.numdam.org/item?id=RSA_1970__18_4_35_0>

DEUX ALGORITHMES RÉCENTS EN CLASSIFICATION AUTOMATIQUE

Maurice ROUX

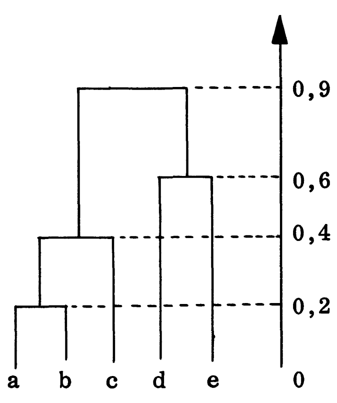
Laboratoire de Statistique Mathématique de l'I.S.U.P.

Faculté des Sciences - Paris

Tout d'abord je tiens à dire ici combien je suis d'accord avec le schéma que vient de nous donner Monsieur TOMASSONE sur l'ensemble des méthodes utilisées pour l'analyse des données.

Nous conformant à ce schéma nous admettrons donc que nous disposons des distances entre les objets à classer. Toutefois les deux procédures que nous allons décrire maintenant s'éloignent quelque peu des méthodes classiques de classification automatique.

I - ULTRAMETRIQUE INFERIEURE MAXIMA



Le but de toute classification est d'obtenir une "hiérarchie indicée", si possible, c'est-à-dire une arborescence où les niveaux d'agrégation des groupes d'objets sont chiffrés par un indice d'autant plus élevé que les groupes agglomérés se ressemblent peu. Ainsi sur la figure, ci-jointe (Fig. 1), les objets a et b ont une forte ressemblance et un indice faible, 0,2, tandis que les groupes [a, b, c] et [d, e] ont une faible ressemblance donc un fort indice, soit 0.9.

Une telle hiérarchie indicée définit une distance d, au sens mathématique du mot, qui a pour valeur la "hauteur" du noeud le plus élevé de l'arbre qui soit encore sur le chemin reliant les deux objets considérés. Sur notre figure la distance entre a et c sera de 0.4. Cette distance a en outre la propriété suivante : si i, j, k sont trois des objets classés par la hiérarchie alors $d(i , k) < \max [d(i , j), d(j , k)]$, ce qui fait que tous les triangles que l'on peut construire avec des objets pour sommets sont isocèles avec la base plus petite que les côtés égaux. Une telle distance est dite ultramétrique.

Réciproquement on montre qu'à toute distance ultramétrique correspond une hiérarchie indicée et une seule. On peut donc considérer l'élaboration d'une classification comme la recherche d'un système de distances sur l'ensemble J des objets à classer qui vérifient l'inégalité ci-dessus. Naturellement, en général, la distance donnée ne satisfait cette inégalité pour aucun des triplets d'objets : les triangles isocèles sont rares dans la nature ! Cependant le système de distances ultramétriques que l'on construira devra s'éloigner aussi peu que possible du système des distances données.

Posons maintenant une définition : la distance δ sur l'ensemble J sera dite inférieure à la distance d sur le même ensemble si, pour toute paire d'objets i et j de J on a :

$$\delta(i, j) \leqslant d(i, j)$$

Soit d la distance donnée sur J. On considère alors l'ensemble D(d) de toutes les distances ultramétriques δ inférieures ou égales à d. Cet ensemble n'est pas vide car il contient au moins la distance δ_0 triviale sur J où toutes les distances sont égales entre elles et égales à la plus petite des valeurs prises par d. Soit alors δ^* l'enveloppe supérieure de la famille des distances de D :

$$\text{pour tous i et j dans J} \quad \delta^*(i, j) = \sup_{\delta \in D(d)} [\delta(i, j)].$$

Il est facile de montrer que δ^* est encore ultramétrique et inférieure à la distance donnée d. Parmi les ultramétriques inférieures à d elle est donc la plus proche de d, au sens de la relation d'ordre entre métriques définie ci-dessus.

Voici un algorithme simple (mais ce n'est pas le seul) pour construire cette distance particulière δ^*, dite ultramétrique inférieure maxima, en raison des propriétés que l'on vient de voir.

Soit T l'ensemble des triplets de points de J. Pour chaque triplet t de T, rendre le triangle correspondant isocèle en réduisant la plus grande des trois distances de ce triangle à être égale à celle qui lui est immédiatement inférieure. Lorsque ceci a été fait, réexaminer l'ensemble T depuis le début de réitérer la modification ci-dessus à chaque fois que celà sera nécessaire. Effectuer ce réexamen de T jusqu'à ce qu'on ne fasse plus aucune modification de la distance qui sera alors égale à l'ultramétrique inférieure maxima.

II - ALGORITHMES D'ECHANGE ET PROGRAMME M.S.H.

Nous examinerons en détail l'application de ce deuxième type d'algorithmes à la construction d'une simple partition de l'ensemble J à étudier, pour ne donner que des indications sur le cas où l'on désire une hiérarchisation complète.

Dans les algorithmes d'échange on se fixe tout d'abord un nombre maximum K de classes possibles à la partition que l'on se propose de construire, et l'on commence par répartir les objets arbitrairement dans ces K classes, soit totalement au hasard, soit en suivant une idée que l'on a à priori sur la manière dont s'organise le domaine étudié. Puis on cherche à changer la classe d'affectation de chacun des objets de façon à maximiser une certaine fonction f que nous préciserons plus tard. Ici plusieurs stratégies sont possibles :

a) la meilleure consiste à essayer pour tous les objets toutes les affectations possibles, et à ne retenir qu'une seule modification d'affectation: celle qui provoque la plus grande augmentation de la fonction f.

b) une procédure beaucoup plus simple et aussi plus rapide consiste à adopter un changement d'affectation dès que celui-ci provoque une augmentation de f.

c) on peut aussi choisir une méthode intermédiaire entre les deux précédentes qui consiste à examiner un objet et à modifier son appartenance de façon à ce que f augmente le plus possible, puis à passer à l'objet suivant et ainsi de suite.

Dans les trois cas on s'arrête lorsqu'il n'est plus possible de modifier l'appartenance des objets sans faire décroitre f. Il va sans dire que la plupart du temps on n'obtient ainsi qu'un maximum relatif de f et non un maximum absolu. Cela tient surtout à ce qu'il faudrait envisager non l'affectation d'un objet à la fois mais de plusieurs pour obtenir ce maximum absolu. La qualité du résultat dépend de la partition initialement choisie , mais en général de la stratégie a) est meilleure que c) elle-même meilleure que b).

Précisons maintenant quelle est cette fonction f. Il s'agit d'une mesure de l'accord entre le résultat et la donnée. Cette donnée étant un système de distances on peut en déduire ce qu'on appelle "l'ordonnance" qui est la relation d'ordre entre les paires d'objets. Une paire est dite plus petite qu'une autre si la distance entre les objets qui constituent cette paire est plus courte que la distance entre les objets de l'autre paire.

De même la partition construite peut se traduire en inégalités : une paire d'objets est plus petite qu'une autre s'il existe une classe de la partition qui contient ces deux objets et s'il n'existe pas de classe qui contienne à la fois les objets de l'autre paire. Une telle relation d'ordre induite par une partition n'est pas un ordre total c'est-à-dire que deux paires d'objets peuvent ne pas être comparables. Soit alors g le nombre d'inégalités contradictoires entre cette dernière relation d'ordre et l'ordonnance donnée, on prend $f = - g$. En d'autres termes g est une mesure de la dissemblance entre le résultat et la donnée, dissemblance qu'il faut minimiser, ou bien maximiser son opposé. Bien entendu on peut inventer d'autres critères que celui-ci mais celui-ci est déjà le fruit de plusieurs essais et semble satisfaisant.

Un mot maintenant du cas où l'on veut une véritable classification au lieu d'une simple partition. Messieurs REGNER et DE LA VEGA de la Maison des Sciences de l'Homme (M.S.H) utilisent à cette fin la procédure décrite en a) mais au lieu de ne retenir qu'une seule agglomération ils effectuent la meilleure moitié d'entre elles, ce qui constitue un niveau de la hiérarchie. Comme dans les procédures classiques il y a une formule pour recalculer la distance entre les nouveaux groupes formés et les anciens et l'on réitére l'opération ci-dessus jusqu'à ce que l'on ne puisse plus augmenter la fonction f. Il se peut que cette procédure d'arrêt laisse certains groupes ou objets non réunis entre eux ce qui signifie que l'arbre obtenu sera non connexe, c'est-à-dire en plusieurs morceaux, ou sous-arbres, non reliés entre eux.

III - EN GUISE DE CONCLUSION

De ces deux méthodes la première a surtout l'avantage d'éclaircir, partiellement, la situation de la classification automatique du point de vue théorique. Pratiquement elle donne le même résultat que celle dite "du saut minimum" ("single linkage cluster analysis" dans les ouvrages de langue anglaise) dont on connaît le désagrément majeur : l'effet de chaîne qui met dans un même groupe deux objets éloignés pourvu qu'existe entre eux une suite de points peu éloignés les uns des autres.

Le deuxième algorithme moins rigoureux théoriquement a, en revanche un intérêt pratique certain.

RESUME

Une première méthode de classification dite de "l'ultramétrique inférieure maxima" est exposée, qui donne les mêmes résultats que l'égrégation par le saut minimum ("single linkage cluster analysis") mais à l'avantage de donner un support théorique à ces résultats. Le seconde méthode décrite, peut, sur le plan pratique, utilement suppléer aux déficiences de la première. Ces deux algorithmes s'écartent notoirement des procédures habituelles.

DISCUSSION

QUESTION (P. DUCIMETIERE) Quelle relation y a-t-il entre l'algorithme "single linkage" et votre premier algorithme ?

Réponse : outre l'éclaircissement théorique déjà signalé nous pensons que cet algorithme risque d'être plus rapide que le single linkage qui nécessite de nombreux tris, opérations "lentes" en ordinateur, mais cela demande a être vérifié sur des exemples réels.

QUESTION (P. DUCIMETIERE) : "L'emploi des stratégies paramétrées de Lance et Williams semble assez rapide...

Commentaire de R. Tomassone à la seconde question de P. Ducimetière - les coefficients α_i, α_j, β, γ, sont les quatre paramètres dont dépendent linéairement les strategies de classification conventionnelles, si d_{ij} est la distance entre deux groupes i et j, et si ces deux groupes fusionnent pour former un groupe k. La distance entre un troisième groupe h et k peut se calculer par :

$$d_{hk} = \alpha_i \, d_{hi} + \alpha_j \, d_{hj} + \beta \, d_{ij} + \gamma \, |d_{hi} - d_{hj}|$$

où les paramètres α_i, α_j, β et γ définissent la stratégie.

Réponse à P. Ducimetière après l'éclaircissement apporté par R. Tomassone "Le paramétrage des stratégies d'agglomaration des objets ne simplifie pas la procédure. Les tris dont nous parlions précédemment arrivent dans la sélection des groupes à fusionner".

QUESTION (R. TOMASSONE) - quand on remplace la nature des distances par la nature des rangs relatifs aux distances, et que l'on forme les groupes en utilisant ces rangs comme mesure de distance, obtient-on les mêmes résultats qu'avec l'algorithme ultramétrique ?

Réponse : la forme de l'arbre reste la même, seules les hauteurs des noeuds sont différentes.

QUESTION (M. TENENHAUS) - on utilise en général des distances différentes en classification automatique et en analyse des correspondances. Comment comparer ces deux méthodes et utiliser leurs complémentarités ?

Réponse : si on veut vraiment comparer ces deux types de méthodes il faut évidemment utiliser la même distance ; mais il nous semble que ces méthodes sont plutôt complémentaires et qu'il vaut mieux utiliser la classification automatique précisément dans le cas où les mesures de distances usuelles sont en défaut, et utiliser alors des indices de dissemblance.

QUESTION (M. MILLIEZ) - la classification ultramétrique est difficilement interprétable à cause des effets de chaine. Si on conçoit uniquement la classification comme une aide à l'interprétation, quelle est l'utilité d'une telle méthode ? La classification ultramétrique peut-elle servir à caractériser la "distribution" des points de départ (aspect multinormal) ?

Réponse : aucune méthode de classification ne nous semble pouvoir caractériser une distribution, même au sens géométrique du terme. La classification par l'ultramétrique inférieure maxima permet toutefois de juger du degré de cohésion des groupes formés à cause du chaînage qui relie immédiatement deux groupes non fortement individualisés.

QUESTION (M. DIDAY) - grands tableaux de données ?

Réponse : Nous n'avons pas encore résolu le problème de grosses données, bien que ce soit un point crucial de la classification automatique mais nous pensons pouvoir y arriver sous peu, par l'intermédiaire du "minimum spanning tree", ou "graphe de longueur minimum", étudié par les chercheurs anglais de Rothamstead.

RÈFÈRENCES

BENZECRI J.P.

1/ description mathématique des classifications

2/ construction ascendante d'une classification hiérarchique

3/ l'algorithme d'échange en classification automatique.

Cours de 3ème Cycle polycopiés, Laboratoire de Statistique Mathématiques, Paris 1969.

DE LA GENIERE J. & DE LA VEGA W.F. - A propos de la classification, rapport interne du centre de calcul de la Maison des Sciences de l'Homme - PARIS 1969.

DE LA VEGA W.F. - Techniques de classification automatique utilisant un indice de ressemblance. Revue Français de Sociologie - PARIS 1967.

GOWER J.C. & ROSS G.J.S. - Minimum spanning trees and single linkage cluster analysis - Applied Statistics, 1969.

LANCE G.N. & WILLIAMS W.T. - A general theory of classificatory sorting strategies 1 - Hierarchical systems. Computer Journal, Vol. 9, 1967.

ROUX M. - un algorithme pour construire une hiérarchie particulière. Thèse de 3ème cycle - Paris, 1968.

SOKAL R.R. & SNEATH P.H.A. - Principles of numerical taxonomy. Freeman and Co. San Francisco and London - 1963.