La revue de Modulad

CLASSIFICATION HIERARCHIQUE SUR TABLEAUX DE DISTANCES

André Carlier
Université Paul Sabatier
31062 Toulouse CEDEX

Résumé:

Ce court papier est une présentation du programme de classification hiérarchique sur tableaux de distance. Il précise son intérêt tant du point de vue traitement de données que du point de vue informatique. Enfin il donne quelques éléments utiles pour en optimiser les performances.

Mots clés:

Classification hiérarchique, tableaux de distances

Classification Hiérarchique sur tableaux de distances

André Carlier
Université Paul Sabatier
31062 Toulouse CEDEX

I GENERALITES

Les données d'un programme de classification sont en général un tableau individus x variables et une métrique définie en général par une option du programme : métrique usuelle, donnant la moyenne quadratique des écarts, métrique réduite utilisant les écarts réduits et enfin métrique du KHI2 pour les tableaux de contingence.

Le programme calcule alors les distances interindividuelles en utilisant le tableau de données et la formule de distance.

II LES LIMITATIONS DES LOGICIELS LES PLUS COURANTS

Lorsqu'on effectue une classification multidimensionnelle, le choix de la formule de distance, ou choix du "mélange multidimensionnel" est particulièrement important. Le choix le plus fréquemment offert (métrique usuelle, métrique réduite, métrique du KHI2) apparaît vite insuffisant. En effet il peut être important de pondérer les variables (métriques diagonales) ou de tenir compte de différences entre variables, ce qui impose l'utilisation de métriques "non diagonales".

Par exemple, si j est associé au temps, on peut être amené à calculer la distance entre deux processus temporels i et i' prenant en compte la dérivée du processus, soit les différences entre les valeurs de la variable entre deux instants consécutifs j et j+1.

Certes, cette limitation dans le choix des métriques peut être levée par d'autres moyens que l'utilisation d'un programme sur tableaux de distances. Un moyen courant consiste alors à effectuer une transformation linéaire sur les données, qui permet alors d'utiliser la métrique usuelle sur le tableau transformé.

Mais il est souvent agréable de travailler sur les données de départ et la démarche consistant à associer un objet informatique (ici la matrice définissant la métrique) à chaque notion statistique (ici la notion de distance) est très naturelle et modulaire. Ainsi un autre moyen réside dans la possibilité de définir en entrée, outre le tableau individus x variables, et l'éventuel tableau des poids, une métrique quelconque (mais rares sont les logiciels qui offrent une telle possibilité). Ces limitations peuvent donc se résoudre sans utiliser les programmes sur tableaux de distance.

Mais il reste un cas où ce programme est nécessaire : c'est celui où les données se présentent directement sous forme d'un tableau de distances ou de dissimilarités, les objets étant équipondérés ou non. Un tel tableau s'obtient par exemple en prenant comme distances les durées (supposées symétriques) de transport entre des villes,...

Les programmes de classification sur tableaux de distances utilisent le fait que les classifications ne dépendent que des distances interindividuelles et d'un éventuel système de poids. Le calcul de ces distances est alors effectué préalablement et c'est directement sur ce tableau qu'opère la classification.

L'intérêt d'une telle démarche n'est pas universel. Elle nécessite cette étape préalable de calcul de la matrice de distance. Donc elle ajoute une étape dans une classification et impose le stockage de la matrice de distances, ce qui n'est pas le cas de la démarche classique.

III L'ASPECT INFORMATIQUE

L'algorithme utilisé est l'algorithme accéléré, décrit dans (Jambu 1978), qui n'est pas spécifique aux tableaux de distances. Il permet de ne consulter à chaque phase d'aggrégation qu'un sous-ensemble du tableau de distances et utilise pour cela la propriété des "voisinages réductibles". Dans un premier temps, on sélectionne dans le tableau de distances celles qui sont les plus petites que l'on range dans un sous tableau de longueur NP. La propriété des voisinages réductibles autorise pour les critères d'aggrégation les plus courants, à ne consulter que cette sous-table jusqu'à épuisement de celle-ci. Il est à remarquer que lorsque le tableau de distance n'est pas euclidien, les résultats des algorithmes peuvent dépendre des paramètres NP et COEF (décrit ci-après), et l'interprétation en terme d'inertie n'est plus qu'analogique.

Le premier problème réside alors dans le choix de la longueur de la sous-table NP : si elle est trop courte, il faudra la régénérer trop souvent, donc balayer tout le tableau de distances restant ; si elle est trop longue, c'est la recherche du plus petit élément qui sera trop coûteuse.

Le second réside dans le choix d'un seuil RO tel que le nombre de distances inférieures à ce seuil soit inférieur à NP, mais pas très inférieur à NP. Le choix de RO se fait alors par l'intermédiaire d'un paramètre COEF, qui est tel que:

$$RO = COEF * (NP/NN) * MOY$$

où MOY est la moyenne des distances du tableau, et où NN est égal aux nombres d'éléments du tableau de distances (N*(N-1)/2 , si N est le nombre d'objets à classer). Les bonnes valeurs de COEF semblent être voisines de 1 (mais cela dépend de l'allure de la distribution des distances), celles de NP sont comprises entre 50 et 100 (ce choix dépend en particulier de l'utilisation ou non des accès disques, dans le cas d'utilisation de mémoire virtuelle).

IV LES CRITERES D'AGGREGATION PROPOSES

Ce sont ceux initialement proposés par Jambu et Lebeaux (1978), dans le programme aggregd, qui a servi de base à cette réécriture, soit :

1 le critère du saut minimum (simple linkage)
2 le critère du diamètre de la réunion (complete linkage)
3 le critère de la distance entre barycentres
4 le critère de Ward (ou inertie inter-classe maximum)
5 le critère de la distance moyenne,

Le critère de Ward est de loin le plus utilisé, et les interprétations en termes d'inertie qu'il permet sont agréables et le font se rapprocher des techniques factorielles. Celui du saut minimum est intéressant dans certains contextes (reconnaissance de formes par exemple) Les autres semblent peu utilisés dans la pratique. Remarquons que le critère 3 ne vérifie pas les propriétés de réductibilité et peut produire un histogramme des indices de niveau d'allure non monotone: les résultats dépendront alors de la valeur de NP et COEF.

V LES ENTREES ET SORTIES DU PROGRAMME

Conçu comme une simple étape d'une chaîne de traitement, le programme utilise le programme de lecture de Thauront (de son programme NUDIS : nuées dynamiques sur tableaux de distances), et ne donne comme sortie que l'histogramme des indices de niveau de la hiérarchie, et une description de celle-ci, selon le format du logiciel ADDAD (cf.Jambu 1978). Un tel choix permet alors une compatibilité avec ce logiciel pour les programmes de description de hiérarchie et de coupure d'arbre.

Enfin ce programme permet de porter les éléments de la matrice de distance à une puissance quelconque. Pour l'intérêt d'un tel choix, on pourra consulter les travaux de Joly et Le Calvé (1986).

BIBLIOGRAPHIE:

JAMBU, M. et LEBEAUX, M.O. Classification automatique pour l'analyse des données , tomes 1 et 2 DUNOD 1978.

JOLY S., LE CALVE G. Etude des puissances d'une distance, Statistique et Analyse des Données 1986, vol. 11 n° 3, pp 30-50.