6. Classification automatique

Etant donné "n" observations, comment puis-je les regrouper en un certain nombre de groupes (disons k) de façon à ce que les groupes obtenus soient constitués d'observations semblables et que les groupes soient le plus différents possible entre eux.

C'est la réponse à cette question que veulent fournir les méthodes de classification automatique. La diversité des méthodes en classification automatique est déconcertante au premier abord. Des volumes entiers ont étés publiés sur le sujet. Nous nous bornerons ici à énoncer les grands principes qui sous-tendent toutes ces méthodes.

Il est important de bien cerner ce qui distingue la classification automatique de l'analyse discriminante. Dans l'AD, les groupes sont connus à priori. Ils servent à définir une équation (ou des équations si le nombre de groupes est supérieur à 2) permettant de classer une nouvelle observation dont le groupe est inconnu. En classification automatique, il n'y a pas de groupes à priori. La méthode cherche dans le nuage de points les zones denses qui formeront des groupes qu'il restera à interpréter par la suite.

6.1 Ce qui distingue les différentes méthodes

Trois éléments permettent de caractériser les différentes méthodes :

i. La classification procède séquentiellement en regroupant les observations les plus `semblables' en premier lieu (méthodes hiérarchiques) ou elle regroupe en k groupes toutes les observations simultanément (méthodes non-hiérarchiques).

ii. Le critère de `ressemblance' entre deux observations.

iii. Le critère de `ressemblance' entre deux groupes ou entre une observation et un groupe.

6.2 Les méthodes hiérarchiques

Historiquement, elles furent les premières développées, principalement en raison de la simplicité des calculs. L'avènement des puissants ordinateurs leur a fait perdre une certaine popularité au profit des méthodes non-hiérarchiques. Toutefois, dans certains domaines (ex. paléontologie), elles demeurent d'utilisation courante en raison de leur capacité d'organiser les ressemblances suivant une hiérarchie, ce qui est le principe de classification habituel lorsqu'on parle d'espèces animales ou végétales.

Définition: On appellera groupe un ensemble de une ou plusieurs observations regroupées par la méthode de classification.

L'algorithme de base est le suivant :

i. a-t-on plus d'un groupe (si non, on termine).

ii. calculer les `ressemblances' entre toutes les paires de groupes.

iii. fusionner les deux groupes montrant la plus grande ressemblance (similarité) ou la plus faible dissemblance (dissimilarité).

Les méthodes hiérarchiques diffèreront entre elles par le choix du critère de ressemblance et par la façon de mesurer les ressemblances entre un nouveau groupe fusionné et les autres inchangés.

6.2.1 La méthode "single linkage" ou "plus proche voisin"

Soit la mesure de ressemblance S. Soit trois groupes R,P et Q. Supposons que R est fusionné avec P. On calculera la ressemblance de (R+P) avec Q de la façon suivante :

$$S(R+P, Q)= \max[S(R,Q) , S(P,Q)]$$

6.2.2 La méthode "complete linkage" ou "voisin le plus éloigné"

Toujours avec la même notation, on calculera plutôt la ressemblance entre les anciens groupes et le nouveau groupe fusion de R et P par:

$$S(R+P , Q)= \min[S(R,Q) , S(P,Q)]$$

6.2.3 Les méthodes intermédiaires

Ces méthodes définissent une ressemblance `moyenne' entre groupes. Il en existe plusieurs variantes.

ex. "average linkage" :

$$S(R+P , Q)= n_p/(n_p+n_r)* S(P,Q) + n_r/(n_p+n_r)* S(R,Q)$$

où n_p : nombre d'observations constituant le groupe P.

D'autres méthodes compromis existent, certaines ne peuvent être utilisées qu'avec certains types de mesures de ressemblance. On peut même, avec le logiciel `Clustan' par exemple, définir nos propres methodes de calcul de ressemblances entre un groupe fusionné et les autres.

La méthode du `single linkage' est peu utilisée de nos jours en raison du problème de chaînage que cette méthode occasionne. Très souvent, en effet, on se retrouve avec un groupe démesurément gros et plusieurs petits groupes satellites. Le `complete linkage ne présente pas ce problème. La méthode du complete linkage tend, au contraire à former des groupes de taille égale.

6.3 Le dendrogramme

Puisque les méthodes hiérarchiques fusionnent les groupes à des degrés décroissants de ressemblance, il est naturel de représenter les résultats de la classification au moyen d'une structure arborescente que l'on appelle dendrogramme

Il peut être intéressant de fournir ici l'algorithme pour la construction d'un dendrogramme qui consiste à ordonner les observations de telle sorte qu'il n'y ait aucun croisement entre les diverses branches du dendrogramme.

i. placer les observations selon un ordre quelconque de gauche à droite
ii. s'il ne reste qu'un seul groupe, on termine.
iii. prendre les groupes compris entre les groupes qui fusionnent et les déplacer rigidement à la droite de la dernière observation du groupe fusionné situé le plus à droite.
iv. retourner à ii.

6.4 Exemples de classification hiérarchique

6.4.1 Exemple de Davis

L'exemple suivant est tiré de Davis. Les observations consistent en des lames minces provenant de grès. Les variables mesurées décrivent la structure de la roche dont: taille et forme des grains, taille et forme des vides et proportion de vides. Davis utilise le cosinus entre les vecteurs d'observations comme mesure de similarité (1: plus ressemblant, -1: moins ressemblant). La similarité après fusion est calculée comme la moyenne des similarités calculées avant fusion (forme de "average linkage"). Il obtient les résultats suivants:

Question: Quelle modification faudrait-il faire à la mesure de similarité si l'on voulait qu'une corrélation négative forte soit indicatrice d'une forte similarité?

6.4.2 Exemple des montérégiennes

L'exemple suivant montre l'application des méthodes liens simples et liens complets aux 68 données provenant des montérégiennes (voir chapitre sur l'ACP). On note les faits suivants:

i. La méthode à liens simples a donné de piètres résultats. On observe un phénomène de chaînage important puisqu'un seul grand groupe est discernable, peu importe le niveau de similarité ou l'on coupe. On ne retrouve pas du tout les groupes pétrologiques connus.

ii. La méthode à liens complets a donné de bons résultats. On retrouve dans l'ensemble les groupes pétrologiques connus. Les groupes sont de taille comparable.

6.5 Les méthodes non-hiérarchiques

Ces méthodes sont plus proches des méthodes factorielles vues jusqu'ici. Leur essor est relié à la possibilité récente d'effectuer à très faible coût des montagnes de calculs. L'idée sous-jacente consiste à rechercher les zones denses du nuage d'observations. L'algorithme de base est le suivant:

A- Initialisation

i. on choisit aléatoirement k (à préciser) groupes d'observations avec un nombre donné d'observations dans chaque groupe (disons n/k), on initialise le critère de dispersion à une valeur infinie.

B- Itération

ii. on calcule le vecteur moyenne (centre) des k groupes.

iii. on calcule pour chaque observation la distance au centre de chaque groupe et on calcule le critère de dispersion présent. Si la dispersion décroît, on continue, sinon on arrête l'algorithme.

iv. on affecte chaque observation au groupe dont elle est le plus près; on obtient ainsi k nouveaux groupes et on retourne à ii.

Cet algorithme porte le nom de "agrégation autour de centres variables". Une version légèrement différente, connue sous le nom de "nuées dynamiques" consiste à représenter chaque groupe non pas par son centre, mais par un ensemble de points (noyau) choisis aléatoirement à l'intérieur de chaque groupe. On calcule alors une distance "moyenne" entre chaque observation et ces noyaux et l'on procède à l'affectation.

Une autre méthode, celle des "k-means", suit le principe général de la méthode d'agrégation autour de centres variables sauf que dès qu'une observation est affectée à un groupe, on recalcule immédiatement la moyenne de ce groupe. Cette méthode converge plus rapidement, toutefois, le résultat final peut dépendre de l'ordre dans lequel les observations sont lues.

6.5.1 Critère guidant les itérations ; convergence

Pour qu' un algorithme de classification non-hiérarchique soit acceptable, il est essentiel de démontrer sa convergence à un optimum local, sinon on pourrait boucler indéfiniment. On parle d'optimum local plutôt que d'optimum global car le critère est minimisé en fonction d'une certaine partition initiale et d'un algorithme donné. Le résultat n'est pas nécessairement le meilleur ensemble que l'on puisse obtenir puisqu'une partition initiale différente aurait pu donner de meilleurs résultats. Autrement dit, le critère d'optimisation représente une fonction, dans l'espace des partitions possibles, qui n'est pas nécessairement convexe.

Un critère pouvant être utilisé est la dispersion dans les groupes, i.e. la somme des carrés des distances des observations au centre du groupe auquel elles sont affectées (c.f. AD).

$$\text{critère} = \text{trace}(D) = \text{trace}\,[(X-C)'(X-C)]$$

où X est la matrice nxp des observations.
C est la matrice nxp des moyennes des groupes répétées autant de fois qu'il y a d'observations dans le groupe considéré
trace: somme des éléments diagonaux

Ce critère permet de contrôler les itérations. En effet la quantité trace (D) ne peut augmenter entre deux itérations successives m et m+1. On applique donc l'algorithme jusqu'à ce que trace(D) diminue d'une valeur considérée négligeable.

Démonstration: Trace(D) ne peut pas croître entre deux itérations m et m+1 :

i. Lorsqu'on calcule $Trace(D)_m$ à étape ii de l'itération m, on a la partition de l'itération m-1 et les centres de groupes de l'itération m.

ii. Si, à l'itération m, après l'affectation on calculait la dispersion en utilisant les centres de groupes de cette même itération, on aurait nécessairement une décroissance de la dispersion puisqu'on a déplacé les observations d'un groupe à l'autre en utilisant le critère de distance minimale. De même, à l'étape iii de l'itération suivante (m+1), on calcule la dispersion avec les mêmes groupes que précédemment mais cette fois par rapport à leur véritable centre et non par rapport au centre de l'itération précédente. Le centre d'un groupe est le point par rapport auquel la dispersion est minimale, donc celle-ci ne peut que décroître (ou demeurée inchangée) à nouveau.

6.5.2 Groupements stables ou formes fortes

La partition finale obtenue correspond à ce qu'il convient d'appeler un optimum local. Le critère de dispersion ne peut plus décroître et on arrête donc l'algorithme. Les résultats obtenus dépendent toutefois de la partition initiale (choisie aléatoirement). Des groupes initiaux différents auraient pu donner une partition finale différente.

Afin de contrer cette ennuyeuse situation et de déterminer quels sont les groupes vraiment homogènes, il suffit d'appliquer l'algorithme plusieurs fois en modifiant les groupes initiaux et de noter quelles sont les observations qui se regroupent toujours ensemble. C'est ce que l'on appelle des formes fortes.

Il convient de noter qu'il y a un maximum de k^r formes fortes possibles (où r est le nombre de partitions initiales différentes). Ainsi avec k=4 et r=5, on aurait 1024 formes fortes possibles (à condition bien sûr qu'il y ait plus de 1024 observations). En pratique toutefois le nombre de formes fortes ayant un effectif non-négligeable est beaucoup moindre que cette valeur (avec n=1024, r=4, on aurait possiblement seulement 7 ou 8 formes fortes ayant plus de 10 observations).

Exemple: Supposons que l'on ait 20 observations et que l'on applique un algorithme de classification non-hiérarchique à partir de 4 partitions aléatoires initiales. Si, à chaque application de l'algorithme, on demande de former 3 groupes, on pourra former le tableau suivant résumant les résultats de la classification.

Observation #	1er essai	2e essai	3e essai	4e essai
15	1	3	3	2
17	1	3	3	2
10	1	3	3	2
12	1	3	3	2
11	1	3	3	2
1	1	3	3	2
4	1	3	3	2
5	1	3	3	2
7	2	1	2	1
2	2	1	2	1
3	2	1	2	1
8	2	1	2	1
16	2	1	2	1
18	3	2	1	3
9	3	2	1	3
20	3	2	1	3
6	3	2	1	3
13	3	2	1	3
14	3	2	2	3
19	3	2	2	3

Dans le tableau ci-haut, on distingue 4 formes fortes. Une première constituée des observations [15,17,10,12,11,1,4,5], une 2e formée des observations [7,2,3,8,16], une 3e regroupant [18,9,20,6,13] et la 4e formée de [14,19]. Cette dernière forme forte est très semblable à la 3e, puisque sur 4 essais, un seul a regroupé ces observations séparément. Le choix de trois groupes semble donc ici judicieux. Si l'on avait observé plus de formes fortes, il aurait alors été souhaitable de répéter la classification en spécifiant plus (ou moins) de groupes à former.

On discerne mieux maintenant une différence fondamentale entre les méthodes hiérarchiques et les méthodes non-hiérarchiques. Avec les premières, on obtient une seule classification représentée par le dendrogramme. L'utilisateur détermine à postériori combien de groupes lui semblent significativement différents. Avec les méthodes non-hiérarchiques, le nombre de groupes doit être fixé à priori. De plus dépendant du nombre de

répétitions de la classification, on obtiendra possiblement un nombre beaucoup plus grand de formes fortes et l'utilisateur devra interpréter ces formes fortes pour possiblement les fusionner, les éliminer de l'étude ou reprendre l'étude avec un nombre de groupe initial différent. Malgré ce désavantage, les méthodes non-hiérarchiques sont supérieures aux méthodes hiérarchiques pour ce qui est de la description des zones denses du nuage des observations. De plus le problème de la mesure de la ressemblance entre deux groupes fusionnés ne se pose évidemment pas. Notons qu'il est possible d'utiliser conjointement les deux méthodes en formant d'abord des groupes (formes fortes) par méthode non-hiérarchique, puis en utilisant une mesure de ressemblance entre les groupes obtenus et en effectuant une classification hiérarchique sur ces groupes. Comme on le voit, il y a large place laissée à notre imagination.

Question: Si on demande de former plus de groupes qu'il n'en existe réellement, que se passera-t-il lorsqu'on appliquera l'algorithme de classification plusieurs fois?

6.5.3 Exemple: classification non-hiérarchique sur les données des montérégiennes.

Par un algorithme de type "k-means", on a obtenu les groupes suivants que l'on représente sur le plan factoriel des axes 1 et 2 de l'ACP. On notera une bonne concordance entre les groupes obtenus par classification automatique et les groupes pétrologiques.

6.6 Critères de ressemblance entre observations.

Il existe une multitude de critères de ressemblance entre observations. Dans un livre récent sur le sujet (Legendre L. et Legendre P.,1984), on en répertorie au-delà de 70 différents. Toutefois, tous appartiennent à deux grands types : les mesures de distance et les mesures de similarité.

Les mesures de similarité obéissent aux deux propriétés suivantes :

i. symétrie: $s(x,y) = s(y,x)$
ii. $s(x,y)$ est maximum quand $x=y$

Les mesures de distance obéissent aux propriétés suivantes :

i. symétrie: $d(x,y) = d(y,x) \geq 0$
ii. $d(x,y) = 0 <==> x=y$
iii. inégalité triangulaire: $d(x,y) \leq d(x,z) + d(y,z)$

Chaque mesure possède ses avantages et ses inconvénients, le choix d'une plutôt que de l'autre s'effectue selon la nature des données et la nature des relations entre observations que l'on veut mettre particulièrement en valeur. La méthode utilisée (hiérarchique ou non-hiérarchique) influence aussi le choix. En effet, on favorisera plutôt les mesures de distance avec les méthodes non-hiérarchiques.

6.7 Aides à l'interprétation

Une fois les groupements obtenus, il faut pouvoir les interpréter. Pour ce faire, toutes les méthodes d'analyse des données vues jusqu'ici peuvent être mises à profit. L'analyse discriminante en particulier permet de déterminer selon quel(s) facteur(s) s'est effectué le regroupement. On obtient parfois d'excellentes visualisations à l'aide de l' ACP ou de l' AC. On peut dessiner des enveloppes représentant les groupes sur les plans factoriels, examiner les relations entre les groupes, etc. Les statistiques élémentaires pour chaque groupe obtenu peuvent également donner d'excellentes pistes. On ne doit évidemment pas négliger toute information extérieure susceptible d'éclairer l'interprétation.

Précisons également que dans certaines applications, les regroupements sont effectués non pas sur les valeurs originales des p variables, mais sur des valeurs issues de celles-ci comme par exemple les coordonnées obtenues de l'ACP sur un nombre restreint ($q < p$) de vecteurs propres, ou encore les coordonnées obtenues de l'AC.

Question: Les groupes obtenus avec les p coordonnées d'une ACP (matrice des covariances) seront-ils les mêmes que ceux obtenus avec les p variables initiales lorsque le critère de dispersion utilise la distance euclidienne?

6.8 Segmentation

La segmentation d'une image consiste à identifier des zones homogènes dans l'image. On identifie d'abord des pixels sur l'image qui servent d'amorce aux différents groupes. L'algorithme type, dit de croissance des régions, procède en cherchant à amalgamer aux différents pixels les pixels voisins. Un critère de distance (dans l'espace des variables) est utilisé.

Algorithme :

i. Déterminer un seuil d_{max} correspondant à la plus grande différence acceptable pour qu'un pixel puisse être joint à un germe donné.

Les germes sont considérés à tour de rôle en commençant par le premier.

ii. Tous les pixels x_i tel que $d(x_i,g_j) < d_{max}$ et tel que partant de x_i il est possible de rejoindre g_j en empruntant uniquement (par le côté) des pixels respectant ce même critère sont associés au germe g_j et sont étiquetés définitivement. Parmi ces pixels, il peut se retrouver d'autres germes non encore traités (la catégorie est alors absorbée et ce germe est retiré de la liste des germes à traiter).

iii. Considérer le germe suivant dans la liste des germes à traiter et aller à ii. jusqu'à épuisement de la liste.

À terminaison, il et possible que certains pixels n'aient été affectés à aucun des germes initiaux. Dans ce cas, on peut soit augmenter la valeur de d_{max}, soit créer de nouveaux germes dans ces zones et repartir l'algorithme, soit laisser ces pixels comme non classés.