Chapitre 3

Classification, segmentation

3.1 Introduction

La classification [1] ou segmentation recouvre l'ensemble des méthodes permettant de regrouper des individus selon leurs similarités en un nombre fini de classes. On constitue ainsi une partition des individus. De façon un peu schématique, on peut dire qu'il existe deux grandes familles de méthodes de classification :

- les méthodes de classification globale reposant, implicitement, sur la construction d'un modèle probabiliste et l'optimisation de critères globaux comme des rapports de variance par exemple,
- les méthodes itératives reposant sur des regroupements locaux d'individus voisins reposant uniquement sur des notions de distances entre individus (ou groupes d'individus).

Dans la première famille, la méthode la plus connue est la méthode des nuées dynamiques ou centres mobiles tandis que dans la seconde famille la méthode la plus répandue est la méthode de classification hiérarchique ascendante.

3.2 Méthode des centres mobiles

La méthode des centres mobiles ou k-moyennes [2] est un algorithme de réallocation dynamique qui repose sur la maximisation d'un critère global construit comme étant le rapport de l'inertie intraclasse sur l'inertie interclasse. Ce critère sous entend que l'on cherche une partition telle que les individus d'une même classe soient le plus semblables possible (variance intra classe faible) et que les classes différent le plus possible entre elles (variance interclasse élevée).

Soit $\mathbf{x} = \{x_{ij}\}_{i=1,\cdots,n, j=1\cdots,p}$ une matrice d'observations. On choisit a priori le nombre de classes K. On note g_k le centre de gravité de la classe k.

Algorithme des kmeans

Initialisation Choisir le nombre de classes K puis choisir K points (individus) au hasard parmi les observations

Itérer jusqu'à ce que le critère de variance interclasse ne croisse plus de manière significative.
Pour $i = 1, ..., n$,

1. en anglais : clustering
2. en anglais : kmeans

- Allouer l'individu i à la classe k telle que dist$(x_i, g_k) \leq$ dist(x_i, g_l) pour tout $l \neq k$.
- Calculer les centres de gravités g_k des K classes.

La partition obtenue par l'algorithme des k-moyennes dépend des représentants initialement choisis (essayez de vous en convaincre sur un exemple simple). De façon à s'affranchir en partie de cette dépendance, on exécute l'algorithme des k-moyennes (K et dist étant fixés) avec des initialisations différentes, et on retient la meilleure partition.

La qualité d'une partition est mesurée par la quantité

$$\sum_{k=1}^{K} \sum_{i \in C_k} \text{dist}(x_i, g_k)$$

qui mesure la cohésion des classes obtenues.

On remarque que la distance dist peut-être définie en fonction du type de variables observées. Cependant, dans la version la plus usuelle de l'algorithme des k-moyennes la distance considérée est la distance euclidienne. Dans le cas où les variables ne sont pas toutes quantitatives, on travaille généralement directement avec un tableau de distances. Dans ce cas, on ne calculera plus le centre de gravité de la classe mais il sera remplacé par le mode de la distribution conditionnellement à la classe.

3.3 Classification hiérarhique ascendante

La classification hiérarchique ascendante est une méthode itérative qui consiste, à chaque étape, à regrouper les classes les plus proches. A la première étape chaque individu constitue une classe. L'algorithme démarre donc de la partition triviale des n singletons.

Algorithme de la classification hiérarchique ascendante

Initialisation Les classes intiales sont les singletons. Calculer la matrice des distances 2 à 2.

Itérer les deux étapes suivantes jusqu'à l'aggrégation en une seule classe.

- regrouper les deux classes les plus proches au sens de la distance entre groupe choisie,
- mettre à jour la matrice des distances 2 à 2.

On peut tracer un graphique représentant la décroissance du rapport de la variance intra classe sur la variance totale (R^2 partiel) en fonction du nombre de classes. La présence d'un rupture importante dans cette décroissance aide au choix du nombre de classes. D'autre part, on trace généralement aussi le *dendogramme*. C'est une représentation graphique des aggrégations successives sous forme d'arbre.

Une CAH est souvent utilisée pour initialiser une méthode des centres mobiles (nombre de classes, centre des classes). Si le nombre d'observations est grand, il est d'usage de réaliser la CAH sur un échantillon tiré au hasard dans la base de données.

3.4 Exemple : composition du lait chez différents mammifères

Nous considérons de nouveau le jeu de données dans lequel on a la composition du lait pour 25 mammifères.

Choisissons tout d'abord le nombre de classes. La figure 3.3 représente la décroissance du R^2 partiel en fonction du nombre de classes. On en déduit qu'il semble raisonnable de considérer 3 ou 4 classes.

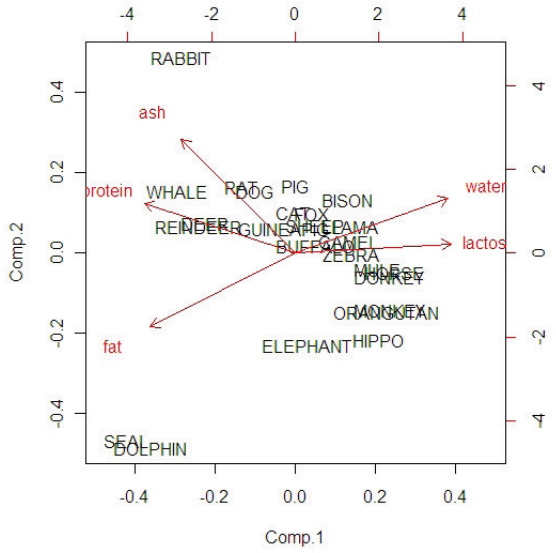

FIGURE 3.1 – Analyse en composante principale, biplot.

3.5 Combinaison de différentes méthodes de classification

Il est courant de combiner les méthodes de classification introduites précédemment. En effet, la méthode de classification hiérarchique n'est raisonnablement applicable que si le nombre d'observations est relativement faible. Son résultat constitue néanmoins souvent une intialisation intéressante pour une méthode des k-moyennes. Il fournit en effet à la fois des critères pour sélectionner le nombre de classes et une initialisation des centres de classe.

Pour les grand ensembles de données, comme on en rencontre fréquemment en data mining, on peut mettre en place la stratégie suivante :

1. Réaliser une classification par nuées dynamique sur un sous échantillon tiré au hasard et de taille environ 10% de n. On choisit un nombre de classes grand.
2. Exécuter une classification hiérarchique ascendante sur les barycentres des classes obtenus puis déterminer un nombre de classe optimal K.
3. Réaliser une classification par k-moyennes pour K classes et en choisissant comme valeurs initiales des centres de classe les barycentres des classes de l'étape précédente. On pourra pondérer ces centres par le nombre d'individu dans les classes.

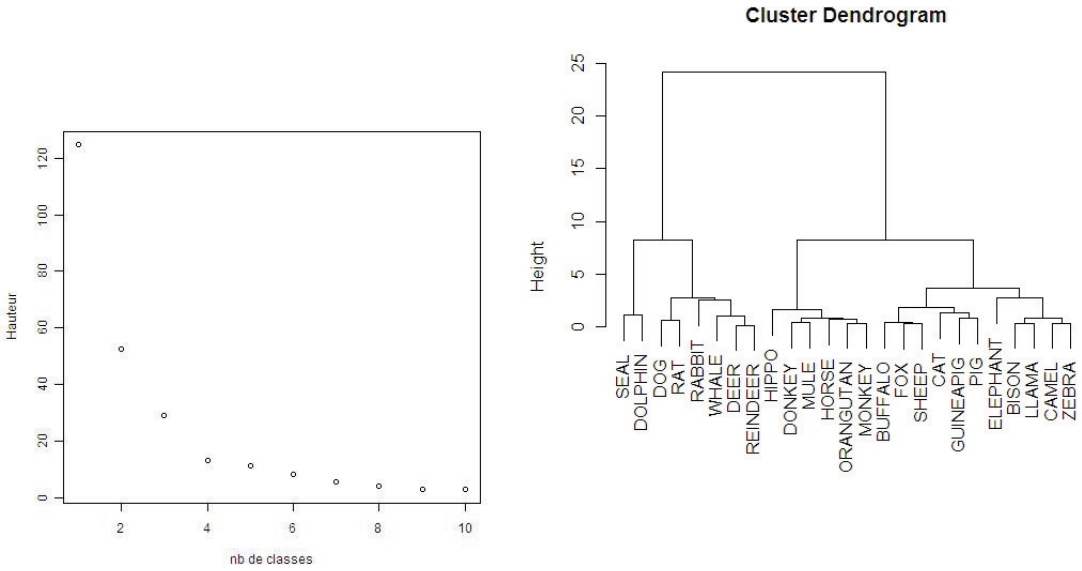

FIGURE 3.2 – Décroissance du R^2 partiel en fonction du nombre de classes (à gauche) et dendogramme (à droite)

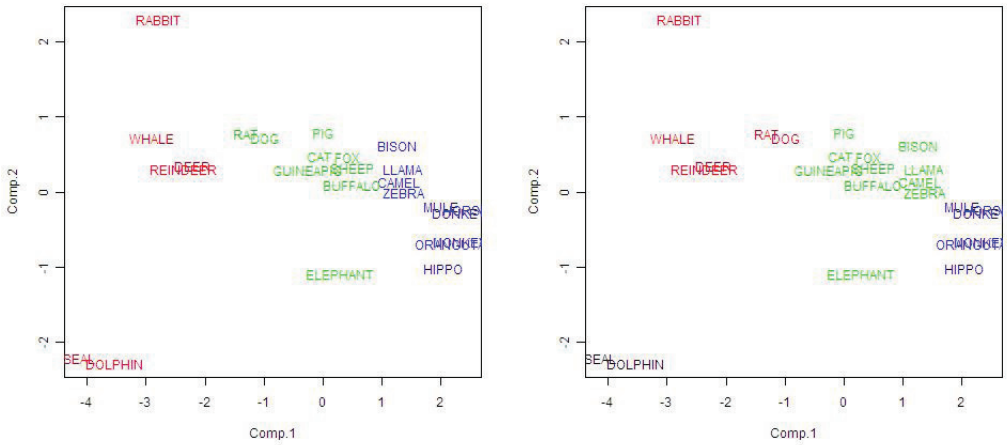

FIGURE 3.3 – Projection des individus sur le premier plan factoriel de l'ACP avec matérialisation de 3 classes (à gauche) et 4 classes (à droite) par un code couleur.

Dans un second temps, on enchaîne généralement d'autres analyses telles que

- Une analyse en composantes principales qui permet de représenter les classes dans un sous espace factoriel et de se faire une idée de la pertinence de la classification obtenue.
- Une analyse discriminante qui permet d'aider à l'interprétation des classes.