Cours 8 : Classification automatique de données par la méthode des centres mobiles.

Comme l'algorithme de classification hiérarchique ascendante, l'algorithme des centres mobiles (*K-mean clustering* en anglais) est un outils de "fouille de données"(*data mining*). La fouille de données joue un rôle important dans presque tous les domaines scientifiques, du marketing qui l'a fait naître[1], à la génétique en passant par l'informatique (reconnaissance de forme) ou la linguistique.

1 L'algorithme des centres mobiles :

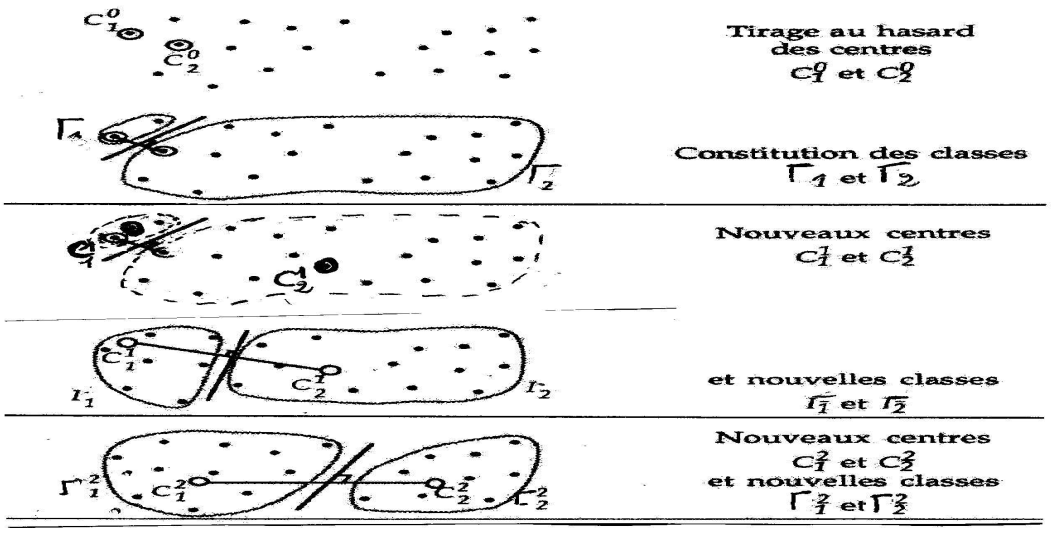
L'objectif de la méthode est de partitionner en différentes classes des individus pour lesquels on dispose de mesures. On représente les individus comme des points de l'espace ayant pour coordonnées ces mesures. On cherche à regrouper autant que possible les individus les plus semblables (du point de vue des mesures que l'on possède) tout en séparant les classes le mieux possible les unes des autres. Ici encore (comme dans le cas de la classification hiérarchique ascendante) on choisit de procéder *de façon automatique*, c'est-à-dire qu'on ne cherche pas à utiliser l'expertise que l'on aurait des individus pour trouver des regroupements avec ce que l'on connait les concernant mais plutot un moyen de *faire apparaître*, uniquement à partir des mesures, des ressemblances et des différences à priori peu visibles. Cette idée, travailler automatiquement, à l'aide de l'ordinateur et *en aveugle*, est appelée *apprentissage non supervisé*.

La méthode des centres mobiles s'applique lorsque l'on sait à l'avance combien de classes on veut obtenir. Appelons k ce nombre de classes. L'algorithme est le suivant :
Etape 0 : Pour initialiser l'algorithme, on tire au hasard k individus appartenant à la population, C_1^0, C_2^0, ..., C_k^0 : ce sont les k *centres* initiaux. On notera que l'indice numérote les différents centres et l'exposant indique qu'il s'agit des k centres initiaux. On choisit aussi une distance entre individus.
On va ensuite répéter un grand nombre de fois les deux étapes suivantes :
Etape 1 : Constitution de classes : On répartit l'ensemble des individus en k classes Γ_1^0, Γ_2^0, ..., Γ_k^0 en regroupant autour de chaque centre C_i^0 pour $i = 1, \ldots, n$ l'ensemble des individus qui sont plus proches du centre C_i^0 que des autres centres C_j^0 pour $j \neq i$ (au sens de la distance choisie).
Etape 2 : Calcul des nouveaux centres : On détermine les centres de gravité G_1, G_2, ... , G_k des k classes ainsi obtenues et on désigne ces points comme les nouveaux centres $C_1^1 = G_1$, $C_2^1 = G_2$, ... , $C_k^1 = G_k$
Répétition des étapes 1 et 2 : on répète ces deux étapes jusqu'à la stabilisation de l'algorithme,c'est-à-dire jusqu'à ce que le découpage en classes obtenu ne soit (presque) plus modifié par une itération suplémentaire.

Le schéma ci-dessous illustre la méthode (à noter qu'en pratique, bien sur, on ne fait pas ces calculs "à la main" mais à l'aide d'un logiciel d'analyse de données). Dans cette figure la distance choisie est la distance euclidienne : en effet, pour répartir les points du nuage en deux groupes, ceux qui sont les plus proches d'un point C_1^0 et ceux qui sont les plus proches d'un autre point C_2^0 au sens de la distance euclidienne, il suffit de tracer la médiatrice du segment $[C_1^0, C_2^0]$.

Mais est-on sûr que cet algorithme conduit bien à une partition *meilleure* que celle dont on est parti, c'est-à-dire celle qui était issue du tirage aléatoire initial de k centres ? Pour répondre à cette question, il faudrait préciser ce que l'on entend par *meilleure*. Nous allons pour cela introduire la notion d'inertie d'un nuage de points.

[1] Selon `http ://fr.wikipedia.oeg/Exploration_de_données`, cet algorithme qui remonte à 1956 serait devenu célèbre en mettant en évidence pour les magasins Wal-Mart un lien entre l'achat de couches pour bébés et l'achat de bières le samedi après midi. L'analyse des résultats d'une classification automatique permit de comprendre qu'il s'agissait de messieurs envoyés par leurs compagnes chercher des couches, jugées trop volumineuses, et qui s'offraient alors des packs de bière. On réorganisa les rayons des magasins en disposant couches et packs de bière à proximité et les ventes de bière s'envolèrent.



2 Inertie inter et intra classes :

On appelle *inertie totale* d'un nuage $\Gamma = \{M_i, i = 1, \ldots, n\}$ la somme pondérée des carrés des distances de ses points au centre de gravité du nuage. Donc, si G désigne le centre de gravité de Γ, l'inertie totale de Γ est, si tous les points du nuage sont de même poids égal à $\frac{1}{n}$, est donnée par la formule :

$$\mathcal{I}(\Gamma) = \frac{1}{n}\left(d_2(M_1, G)^2 + d_2(M_2, G)^2 + \ldots + d_2(M_n, G)^2\right). \quad (1)$$

Notons que le centre de gravité est précisément le point G pour lequel cette somme pondérée est minimale. L'inertie "mesure" la dispersion du nuage, elle sera grande pour un nuage très dispercé et petite lorsque le nuage est constitué de points bien regroupés. Si le nuage Γ est composé de k classes Γ_1, Γ_2, ... , Γ_k (disjointes deux à deux), celles-ci seront d'autant plus homogènes que les inerties de chaque classe, $\mathcal{I}(\Gamma_1)$, $\mathcal{I}(\Gamma_2)$, ... , $\mathcal{I}(\Gamma_k)$, calculées par rapport à leurs centres de gravité G_1, G_2, ... , G_k respectifs, seront faibles. La somme de ces inerties est appelée *inertie intraclasse* :

$$\mathcal{I}_{intra} = \mathcal{I}(\Gamma_1) + \mathcal{I}(\Gamma_2) + \ldots + \mathcal{I}(\Gamma_k)$$

Les inerties des classes $\mathcal{I}(\Gamma_1)$, $\mathcal{I}(\Gamma_2)$, ... sont simplement calculées avec la formule (1) ci-dessus où l'on remplace le centre de gravité G par celui de la classe G_1, G_2, ... et le poids $\frac{1}{n}$ par celui de la classe.

L'inertie totale d'un nuage n'est généralement pas égale à la somme des inerties des classes qui le composent, c'est-à-dire à l'inertie intraclasse (sauf dans le cas où les centres de gravité de toutes les classes sont confondus) car il faut prendre en compte également la dispersion des classes par rapport au centre de gravité du nuage. Il s'agit de l'*inertie interclasse* définie par

$\mathcal{I}_{inter} = \overline{p}_1 d_2(G_1, G)^2 + \overline{p}_2 d_2(G_2, G)^2 + \ldots + \overline{p}_k d_2(G_k, G)^2$, où $\overline{p}_j$ désigne le poids total de la classe Γ_j.

On montre le résultat suivant appelé *décomposition de Huygens :*

Théorème 1 *L'inertie totale d'un nuage de points composé de différentes classes disjointes deux à deux est la somme de son inertie intraclasse et de son inertie interclasse, c'est-à-dire :*

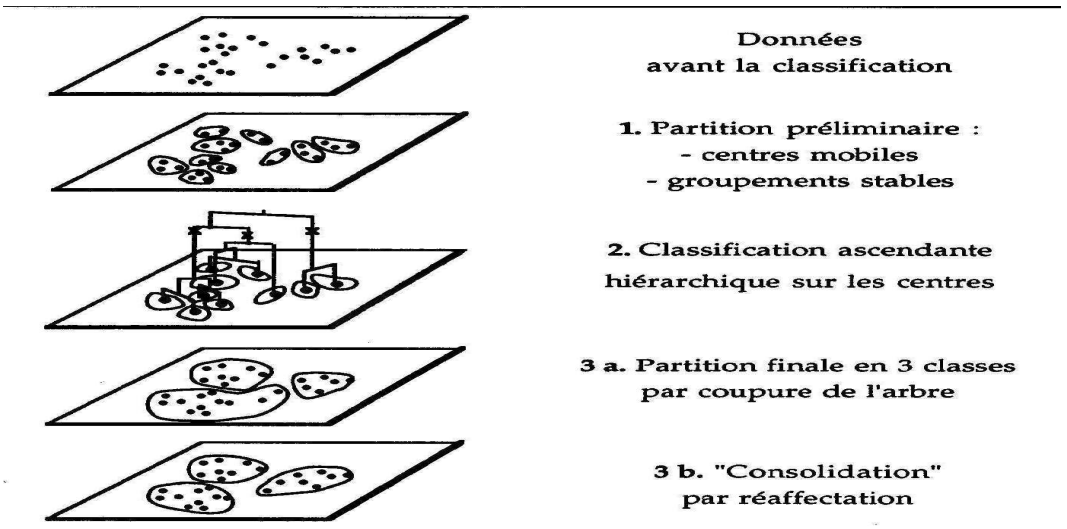
$$\mathcal{I}(\Gamma) = \mathcal{I}(\Gamma_1 \cup \Gamma_2 \cup \ldots \cup \Gamma_k) = \mathcal{I}_{intra} + \mathcal{I}_{inter}.$$

On a vu que lorsqu'un nuage est composé de plusieures classes, si chacune est très bien regroupée autour de son centre de gravité, son inertie intra, qui est la somme des inertie de chaque classe sera petite. La partition d'un nuage en k classes, pour un nombre de classes fixé, sera d'autant *meilleure* que son inertie intra sera petite, ou, puisque son inertie totale reste la même quelque soit la partition, que son inertie extra est grande. Or on peut montrer justement que l'inertie intra classe ne peut que décroître lorsque l'on passe d'un regroupement en classes $\{\Gamma_1^i, \Gamma_2,^i \ldots, \Gamma_k^i\}$ au suivant $\{\Gamma_1^{i+1}, \Gamma_2,^{i+1} \ldots, \Gamma_k^{i+1}\}$ par une itération de l'algorithme des centres mobiles. Si cette décroissance était toujours stricte, le nombre de partitions différentes d'un ensemble fini de points est lui-même fini (même s'il est gigantesque), on serait sur d'atteindre ainsi le minimum. En pratique, la décroissance n'est pas toujours stricte et on n'est donc sûrs de rien. Mais cet algorithme est populaire car il est facile à utiliser et il suffit souvent de peu d'itérations pour avoir deja une partition de qualité.

3 Méthodes mixtes

L'algorithme des centres mobiles a cependant deux défauts principaux :
1) tout d'abord il exige de l'utilisateur de choisir à l'avance le nombre de classes de la partition, ce qui est parfois difficile.
2) Ensuite on s'aperçoit que la partition que l'on obtient peut varier sensiblement en fonction du choix des centres initiaux. Cela vient du fait que, si l'inertie intra décroit effectivement à chaque itération, ce n'est pas forcément vers le minimum recherché mais parfois vers un *minimum local* qui n'est pas du tout optimal. En pratique, comme le déroulement de l'algorithme est généralement rapide, on n'hesite pas à l'exécuter plusieurs fois avec des choix différents des centres initiaux et on compare les partitions obtenues pour ne retenir que celle dont l'inertie intra est minimale, ou, si aucune n'est clairement minimale, la partition qui revient le plus souvent (groupements stables).

Au dela de la classification hierarchique ascendante et de la méthode des centres mobiles, il existes beaucoup d'autres méthodes (par exemple des méthodes stochastiques comme les reseaux de neurones) mais l'utilisateur privilégie souvent, lorsque le nombre d'individus est très grands et qu'il est alors difficile de choisir d'avance le nombre de classes, une classification mixte comme indiquées sur la figure suivante :



Si l'on a des milliers, voir des dizaines de milliers d'individus à classifier, on commence par les répartir en un (trop) grand nombre de classes (par exemple $k = 100$) par la méthode des centres mobiles. Puis, on ne retient que les centres des classes (avec leur poids qui sera proportionnel au nombre d'individus dans chaque classe) $\{(C_1^n, p_1), (C_2^n, p_2), \ldots, (C_{100}^n, p_{100})\}$ et on effectue une classification hiérarchique ascendante sur ces centres.Une partition est alors obtenue par coupure du dendrogramme que l'on choisit aussi judicieusement que possible (par exemple *au plus grand saut*) pour avoir le *bon* nombre de classes. On peut alors calculer leurs centres de gravité et finalement alouer chaque individu au centre le plus proche, ce qui *consolide* la partition.