Introduction à la classification en Sciences Humaines

Georges Le Calvé
Université de Rennes2
6 Avenue Gaston Berger
35043 Rennes Cedex France

1. Introduction

Les étudiants de Sciences Humaines sont des étudiants particuliers, en ce sens que, s'ils n'ont guère, voire aucune connaissance mathématique, ils sont doués d'une facilité d'abstraction, à partir d'analyses. Ceci est typique dans le domaine des Sciences du Comportement. Or cette possibilité de découvrir ce qui fait que tel cas particulier relève d'un cas plus général déjà connu, ou que deux comportements appartiennent à la même pathologie est un exemple de la démarche classificatoire. Ces étudiants pratiquent de manière continuelle cette démarche, et il faudra peu de choses pour les amener à la formaliser. Ils se montrent, de plus, très intéressés si on prend la peine de se débarrasser de ce qui est purement technique pour se consacrer aux idées.

C'est dans cette optique que nous avons conçu ce cours à leur destination. Dans un premier temps il s'agit de leur faire découvrir ce qu'est une classe ; dans un deuxième de définir un outil mathématique qui servira tout au long de l'année, et dans un troisième de préciser quelques problématiques et les réponses que l'on peut y apporter.

2 La notion de classe

L'idée de classe est introduite à partir de notions et d'expressions empruntées à la vie courante. « Qui se ressemble s'assemble » indique très clairement que les classes sont constituées de groupes homogènes, ayant les mêmes propriétés. « Dis-moi qui tu fréquentes, je te dirai qui tu es » montre que l'on attribuera à un individu d'une classe les propriétés des autres.

Une classe sera donc composée d'individus ayant les mêmes caractéristiques, et interchangeables. Il faudra donc une cohésion forte. « Une porte doit être ouverte ou fermée », «c'est blanc ou c'est noir » indiquent des classes non empiétantes et bien séparées.

3 L'outil mathématique

Il est évidemment nécessaire d'utiliser un, ou plusieurs, outils mathématiques. La difficulté est que ces étudiants, comme beaucoup d'autres, mais peut-être encore plus que d'autres, font un rejet des mathématiques. Il est impossible d'utiliser des techniques aussi simples que la sommation sous la forme Σ , la dérivation ou l'intégration, les théorèmes de mécanique classique et leurs conséquences sur les décompositions de la variance, le produit scalaire. Toutes ces notions font appel à des outils sophistiqués dont ils ne maîtrisent pas le maniement. Il est vrai que certains d'entre eux sont non-intuitifs et difficiles.

On ne peut donc utiliser que des notions simples. Nous avons décidé de recourir à la notion de distance. Celle-ci est en effet très intuitive, d'un maniement facile, demande peu d'axiomes et ceux-ci sont aisés à justifier.

3-1 Distances entre points

A priori les étudiants font référence, au début tout au moins à des notions de distance qui sont en fait des distances géométriques, voire géographiques ; il s'agit le plus souvent de la distance euclidienne. Il est aisé de leur faire comprendre qu'il en existe d'autres, par exemple la distance « city-block » pour se déplacer dans une ville ou dans les couloirs et escaliers de l'Université. On

peut également leur introduire facilement la distance géodésique (sur la sphère), utile pour se déplacer à la surface du globe. Il est même ensuite très simple de dévier sur la géométrie de Lobatchevski. Celle-ci mérite un intérêt particulier, dans le sens où, pure curiosité intellectuelle pendant longtemps, elle trouve enfin une application pratique.

On en déduit facilement les axiomes : la nullité de la distance d'un point à lui même, la symétrie, l'inégalité triangulaire introduite comme « la ligne droite est le plus court chemin d'un point à un autre ».

Pour simples qu'ils soient, ces axiomes ne sont cependant pas vérifiés par certaines évaluations (des dissimilarités) qui ne sont donc pas des distances, mais qui mesurent cependant des notions d'éloignement ou d'écartement. distance évaluée en temps de parcours, qui est une notion non symétrique, plus grand saut à effectuer pour passer d'une rive d'une rivière à une autre en empruntant une suite de rochers, qui est une notion ne vérifiant pas l'inégalité triangulaire, « distance affective » entre deux individus, très utilisée en sociométrie, et qui n'est ni symétrique ni transitive.

On passe ensuite à des calculs de distance sur des variables statistiques, supposées, dans un premier temps, prendre des valeurs numériques. Les étudiants découvrent qu'une bonne solution est de combiner les différences ou les écarts individuels. Ils proposent alors d'en faire la somme, débouchant sur la distance d_1 , d'en conserver la plus grande, ce qui génère la distance d_∞, de compter le nombre de fois où cette différence n'est pas nulle, ce qui donne la distance d_0 ; mais il faut leur expliquer que l'on peut prendre la racine carrée de la somme des carrés des écarts de ces distances, qui est pourtant la distance euclidienne d_2.

Toutes ces distances sont des cas particuliers d'une seule, où on combine de manière plus ou moins additive des puissances de ces écarts individuels : la distance de Minkovsky. Ces distances différant les unes des autres par la puissance, il est clair qu'elles feront jouer des rôles plus ou moins importants aux grands écarts individuels.

Il est alors fondamental d'insister sur le fait qu'aucune de ces distances n'est meilleure qu'une autre, pas plus que la distance kilométrique n'est meilleure que la distance en temps de parcours. Elles sont plus ou moins bien adaptées à des situations concrètes, et le choix offert n'est qu'apparent, une problématique entraînant presque automatiquement le choix de l'une d'entre elles. Ceci a comme conséquence que, sur les mêmes données, on peut ou on doit choisir des distances différentes suivant le problème à résoudre

Ainsi la distance Euclidienne d_2 et la distance du sup d_∞ favoriseront les grands écarts, et donc insisteront sur la séparation des classes, tandis que les distances d_0 et d_1 feront jouer des rôles plus ou moins importants aux petits écarts, insistant ainsi sur l'homogénéité des classes

3-2 Distances entre ensembles

Désirant construire des classes homogènes et séparées, il est clair que l'on cherchera des ensembles tels que les distances entre les points d'une même classe sont faibles, tandis que les distances entre deux classes sont grandes. Il est donc nécessaire d'introduire la notion de distance entre classes.

L'analogie géographique est intéressante. Si on veut estimer la distance entre deux pays, on pourra prendre la distance minimale entre leurs frontières, ou au contraire la distance maximale entre une ville de l'un et une ville de l'autre, ou encore la distance entre leurs capitales ou plutôt leurs centres géographiques. Si cette dernière est bien une distance, ce n'est pas le cas des autres, puisque la plus petite ne vérifie pas l'inégalité triangulaire et que la plus grande ne vérifie pas $d(A,A) = 0$

4 Quelques algorithmes de classification

4-1 Les nuées dynamiques

Une première technique très simple à introduire, est celle connue sous le nom de « k-means », et son extension les nuées dynamiques. Dans un cas comme dans l'autre on procède en deux étapes, généralement réitérées : la première désigne le leader, noyau, ou centre de gravité d'une classe ; la seconde demande à chaque individu dans quelle classe il veut aller (règle d'affectation à la classe la plus proche).

Ces deux étapes supposent que l'on sache le nombre de classes que l'on désire, comment désigner le leader et comment calculer la distance d'un point à une classe. Nous avons vu qu'il y a plusieurs choix possibles : par exemple pour le leader on peut choisir le « point central » d'une classe comme étant celui qui est à distance minimum de tous les autres. Comme il y a plusieurs distances, il y a plusieurs réponses. La médiane correspond à d_1, le centre de l'étendue à d_∞ , le mode à d_0 .Le centre de gravité se découvre en prenant la distance d_2 et en minimisant la somme des carrés. On retrouve toujours la difficulté de calcul de la distance euclidienne qui, comme l'illustre Pythagore, additionne des carrés et non des nombres comme dans le city-block. Il est clair que le choix de la distance, et donc du « leader » conditionne le résultat. Quant à la distance d'un point à une classe il s'agit de la notion évoquée au paragraphe précédent.

Cette démarche est bien connue des étudiants, puisque c'est à peu de choses près celle qui est utilisée par des enfants pour constituer des équipes ou même des bandes, voire hélas des « gang ». Si on prend cette analogie, qui fait référence à la notion de « distance affective », il est aisé d'interpréter le choix de chacune des distances. Si on prend la plus petite distance pour mesurer l'attirance d'une classe, on affecte un individu à la classe de son meilleur ami. Si on prend la distance du sup, on l'affecte à la classe où il n'a pas d'ennemi, la distance entre les points centraux étant un compromis entre les deux. Ces situations sont exactement celles rencontrées lorsque l'on cherche un groupe pour passer une soirée, ou pour passer une semaine en vacances, ou pour entretenir des relations de bon voisinage.

4-2 Les arbres hiérarchiques

Cette technique procède d'une autre approche que la précédente. On se refuse à faire des classes en nombre fixé, en raison de deux attitudes opposées : l'une affirme que deux individus ne sont jamais semblables et qu'il faut autant de classes que d'individus, l'autre affirme que, vu de loin, tous les individus sont interchangeables.

Une analogie est alors très pratique : celle de l'astronaute qui voit les distances apparentes des villes diminuer au fur et à mesure qu'il s'élève dans l'espace. Si, de nuit, il regarde les villes éclairées à la surface de la terre, il les voit toutes séparées. Mais, rendu sur la Lune ou sur Mars, il les voit toutes confondues. Ce qui est intéressant est ce qui se passe au cours de l'ascension. En effet, si les distances apparentes diminuent régulièrement, dès que l'une d'entre elles est inférieure au pouvoir séparateur de l'œil elles sont confondues. Naturellement ceci se produit d'abord pour les villes les plus proches. On a là le principe de la construction d'un arbre hiérarchique ascendant. Si l'on choisit comme distance l'une des trois définies précédemment entre groupes, on obtient les techniques connues de lien minimum, lien complet, ou lien moyen..

Il est très important d'insister sur le fait que, comme pour les techniques de partitionement, ces choix correspondent à des problématiques différentes.

Les méthodes de Ward sont plus difficiles à introduire. Elles nécessitent l'introduction d'une notion supplémentaire que nous nommerons *l'épaisseur* d'une classe. Il s'agit d'une combinaison, généralement additive, des distances entre les points de la classe, et dépend donc du type de la distance choisie. Ainsi, pour d_0, nous prendrons le nombre de points dans la classe, pour d_∞ la plus grande de ces distances,(c'est à dire le diamètre) ; pour la dis tance Euclidienne, toujours plus compliquée en raison du théorème de Pythagore, la racine carrée de la somme des carrés des distances, ce qui, à un coefficient près, est l'écart-type. Cette notion d'épaisseur, dont on n'utilise en général que la version Euclidienne, trouve beaucoup d'autres applications que celle-ci. C'est une notion naturelle de dispersion. Elle permet également d'introduire une nouvelle distance entre deux classes, basée sur la variation de l'épaisseur : la distance $\delta(A,B)$ entre deux classes sera l'épaisseur

de A, plus celle de B, moins l'épaisseur de la classe obtenue en les réunissant. (Ce qui, dans certains cas , revient à l'épaisseur de la différence symétrique AΔB). L'algorithme de Ward est alors un algorithme classique au sens de cette distance en utilisant l'épaisseur Euclidienne.

Il est fondamental d'insister sur ce dernier aspect : les classifications changent en fonction de la distance choisie ; on doit en montrer des exemples. La conséquence est que *« il n'existe pas de classification naturelle sur des observations »* de même que *« il n'existe pas de classification ou de technique meilleure qu'une autre."* La classification dépendant de la distance choisie et celle ci de la question posée, il est clair que le résultat est la conséquence de la manière dont on a regardé les données, et que a priori, aucun regard n'est meilleur qu'un autre. La seule chose qui puisse être dite est que, pour tel problème précis, telle classification semble plus adaptée que telle autre.

4.3 Arbres additifs

La notion d'arbre additif s'introduit très facilement à partir de la ntion d'arbre généalogique, d'arbre d'évolution, ou de *stemme* en critique des textes. Par contre, si la notion est naturelle, les algorithmes de construction sont compliqués. Les étudiants admettent très bien que d'autres qu'eux, par exemple des machines, se livrent à des calculs dont ils connaissent la raison mais non le déroulement, qui, d'ailleurs' ne les intéressent pas. Le travail consiste alors à leur apprendre à interpréter de tels arbres .

5 Conclusion

Bien que d'un niveau très faible en mathématique, ou peut-être à cause de cela, les étudiants refusent d'exécuter des opérations – calculatoires ou logiques- dont ils ne comprennent pas la philosophie. Or la Classification a cet avantage énorme sur d'autres techniques que les idées qui la sous-tendent sont très simples, pourvu que l'on utilise un langage et une formulation à leur portée.

SUMMARY

The notion of cluster is a very intuitive notion very often used in the current life. Looking for clusters in a population is a more sophisticated problem. Every student finds quickly that we must define "well separated and homogeneous" subsets, the main difficulty being, precisely the definition of the notion of distances between, and inside, subsets.

The course begins with the study of the main distances between points and continues with the notion of distances between a point and a subset. K-means and related methods are then studied as applications of these notions.

The whole course is conducted by the way of examples.