MATHÉMATIQUES ET SCIENCES HUMAINES

ALAIN GUÉNOCHE

Hiérarchies conceptuelles de données binaires

Mathématiques et sciences humaines, tome 121 (1993), p. 23-34

<http://www.numdam.org/item?id=MSH_1993__121__23_0>

Math. Inf. Sci. hum., (31e année, n° 121, *1993*, p.23-34)

HIÉRARCHIES CONCEPTUELLES DE DONNÉES BINAIRES

Alain GUÉNOCHE[1]

RÉSUMÉ — *En classification conceptuelle d'un ensemble d'objets décrits dans un espace de représentation, on cherche à construire une partition des objets en classes disjointes et simultanément une caractérisation de chaque classe dans les termes de l'espace de représentation. Dans le cas, très courant, où cet espace est engendré par des données binaires nous présentons deux algorithmes, dérivés des méthodes ascendantes et descendantes en classification qui maximisent localement un indice de cohésion des classes. Les caractérisations construites sont des conjonctions de caractères communs qui sont également caractéristiques des classes. De ce fait elles sont monothétiques et et constituent des éléments du treillis Galois.*

SUMMARY — Conceptual clustering of binary atributes.
The conceptual clustering problem is to build not only a partition of a set of objects into separated classes, but also to associate to each class a characterization in the representation space terms. In this text we present two algorithms, derived from classical clustering methods, to realise simultaneously these two functions, in a representation space generated by binary attributes. Characterizations will be expressed like characteristic functions of monothetic classes that correspond to concepts selected from Galois lattice.

La classification conceptuelle [Michalski, Diday & Stepp 1981] d'un ensemble X d'objets, décrits dans un espace de représentation engendré par des descripteurs, est un problème né du rapprochement méthodologique entre l'Apprentissage à partir d'exemples et l'Analyse des Données [Gascuel & Guénoche 1990]. Il s'agit de construire non seulement une partition de X en classes disjointes, ou une hiérarchie de partitions, mais aussi d'associer à chaque classe une *caractérisation* dans cet espace de représentation. En Apprentissage, on appelle *généralisation* [Mitchell 1982] l'opération qui consiste à décrire un concept (une classe d'objets) c'est à dire, lorsqu'il est défini *en extension* par des *instances* (par la liste de ses éléments) à exprimer une ou plusieurs conditions d'appartenance à ce concept (de le définir *en intension*). Une solution héritée de l'Analyse des Données consiste à utiliser une méthode de classification, puis les classes étant définies, à construire leurs caractérisations par des méthodes de généralisation [Nicolas 1988], par exemple à l'aide d'*arbres de décision* [Breiman & al. 1984] ou de méthodes de discrimination conceptuelle [Guénoche 1988].

Nous présentons dans ce texte des méthodes qui réalisent simultanément ces deux opérations, de partition et de caractérisation, et ce dans un espace de représentation très général engendré par des attributs binaires. Les variables considérées peuvent être soit des variables de *présence/absence* de caractères, soit des variables *dichotomiques* dont les deux modalités sont arbitrairement codées 0 et 1. Cette distinction est nécessaire, car si, dans le premier cas on peut ne s'intéresser qu'aux combinaisons d'attributs simultanément présents, dans le second il convient de donner aux deux valeurs 0 et 1 des rôles identiques. Cet espace de représentation renvoie à deux formalismes différents pour le calcul : un formalisme latticiel dans lequel seules les valeurs codées 1 sont significatives et un formalisme booléen dans lequel les deux valeurs sont traitées de façon équivalente. C'est dans un formalisme qui emprunte aux deux que nous

[1] G.R.T.C. - C.N.R.S., 31, chemin Joseph Aiguier 13402 Marseille Cedex 9.

nous plaçons, en déclarant préalablement que pour certains attributs (pas nécessairement tous) les valeurs 0 sont significatives.

Intuitivement, une caractérisation d'une classe Y de X est un ensemble de propriétés que vérifient la plupart des éléments de Y et que ne possèdent pas, sauf exception, les éléments de X\Y. Une fonction caractéristique de Y, au sens ensembliste, c'est à dire qui pour tout élément de X prend la valeur 1 s'il appartient à Y et la valeur 0 sinon, est bien sûr une caractérisation au sens le plus strict. Nous ne produisons que ce type de caractérisations, appelées en Apprentissage *fonctions complètes et consistantes* ; ce sont des fonctions booléennes dont nous utiliserons une expression sous forme normale disjonctive.

Pour caractériser X en entier, on retient les attributs qui prennent sur X une seule des valeurs 0 ou 1. Le plus souvent il n'y a pas ou peu d'attributs de ce type. On cherche donc à partitionner X en classes qui présentent sur plusieurs attributs une seule des deux formes. Plus il y en a, mieux la conjonction de ces attributs caractérise les objets de la classe. C'est l'importance de ces parties communes qui fait la qualité des partitions choisies ; c'est donc un nouveau critère de classification qui est introduit dans ces méthodes dites *conceptuelles.*

Pour fixer les idées, nous utiliserons un exemple où huit espèces animales sont décrites par sept attributs (les réponses "oui" et "non" sont respectivement codées 1 et 0) :

A : L'animal pond-il des œufs ? B : Est-il couvert de plumes ? C : Est-il couvert d'écailles ?
D : Est-il couvert de peau ? E : Présence de dents ? F : L'animal vole-t-il ?
G : L'animal nage-t-il ? H : L'animal respire-t-il dans l'air (1) ou dans l'eau (0) ?

Pour les animaux considérés on obtient le tableau ci-dessous :

	A	B	C	D	E	F	G	H
1 Autruche	1	1	0	0	0	0	0	1
2 Canari	1	1	0	0	0	1	0	1
3 Canard	1	1	0	0	0	1	1	1
4 Requin	1	0	0	1	1	0	1	0
5 Saumon	1	0	1	0	0	0	1	0
6 Grenouille	1	0	0	1	0	0	1	1
7 Crocodile	1	0	0	1	1	0	1	1
8 Barracuda	1	0	1	0	1	0	1	0
9 Cygne	1	1	0	0	0	0	1	1

Les espèces mentionnées n'ont qu'un seul caractère commun : pondre des œufs. Maintenant si l'on regarde ces descriptions avec un peu plus d'attention, on remarquera qu'il y a d'une part les animaux qui ont des plumes et n'ont pas de dent et qui vivent dans l'air et d'autre part ceux qui ne volent pas, qui nagent et qui, s'ils ont des écailles, vivent dans l'eau. Ce faisant nous avons partitionné nos animaux en classes disjointes, chacune ayant comme caractéristique l'ensemble de ses traits communs. Dire que ces animaux appartiennent à l'une des deux classes, chacune ayant une *cohésion* plus forte que l'ensemble, est une caractérisation plus *spécifique* que le seul attribut qu'ils ont en commun. On pourra recommencer cette opération sur l'une ou l'autre classe, c'est à dire les subdiviser de façon qu'elles aient une plus large part de description commune.

Nous allons donner un cadre formel à cette démarche, un sens précis au terme généralisation, et présenter des algorithmes pour construire ces partitions et les conjonctions de propriétés de leurs classes.

1. NOTATIONS ET TERMINOLOGIE

Soit X l'ensemble des n objets décrits par un ensemble **A** = {A, B, C, .. } de m attributs, chacun, par exemple A, pouvant prendre les valeurs a (dite *forme directe*, codée 1) et a' (dite *forme conjuguée*, codée 0). On peut considérer X comme un tableau de valeurs 0 ou 1 avec n lignes correspondant aux éléments de X et m colonnes correspondant aux éléments de **A**. On pose **A*** = {a, a', b, b' .. j, j'.. } l'ensemble des 2.m formes des m variables. On appelle *monôme de longueur* k une conjonction de k formes, c'est à dire une partie à k éléments de **A***.

Un monôme est noté par la concaténation de ses formes. Par exemple μ = ab'd est un monôme de longueur 3. Un monôme est dit *complet* si chaque variable apparaît sous une seule forme, donc s'il est de longueur m = | **A** |. L'ensemble H = $\{0,1\}^m$ tient lieu de d'*espace de représentation* des objets.

On note X(μ) l'ensemble des éléments de X dont les valeurs des variables sont les formes présentes dans μ, ou encore l'*extension* de μ. Par exemple X(ab) est l'ensemble des espèces de X ayant 1 en première et en seconde colonnes de la table (soit {autruche, canari, canard et cygne}). Un monôme est dit *vide* sur X si X(μ) = $\varnothing$. En particulier, si une variable apparaît dans un monôme sous ses deux formes, ce monôme est vide. Parallèlement H(μ) est l'ensemble des m-uples de $\{0,1\}^m$ qui *contiennent* μ. Comme il y a 8 attributs, en fixant les deux premiers on a 2^6 = 64 éléments dans H(ab).

Une disjonction de monômes $\Phi = \mu_1 \vee \mu_2 \vee \ldots \vee \mu_p$ est une *formule normale disjonctive.* Pour tout élément x de H, Φ(x) = "vrai" ssi il existe i tel que x $\in$ H(μ_i), "faux" sinon. On définit :

$$X(\Phi) = \cup_{i=1,..p}\ X(\mu_i).$$

Parallèlement H(Φ) est l'ensemble des éléments de $\{0,1\}^m$ pour lesquels Φ prend la valeur "vrai". Si l'on note $\underline{\mu}_i$ le monôme complet correspondant au i-ième élément de X, la fonction

$$\underline{\Phi} = \vee_{i=1,\,..n}\ \underline{\mu}_i$$

est *équivalente* au tableau X, c'est à dire que l'on a à l'évidence H($\underline{\Phi}$) = X et $\underline{\Phi}$ ne prend la valeur "vrai" que sur les éléments de X.

2. CLASSIFICATION CONCEPTUELLE

Souvent en Apprentissage, d'autant plus s'il s'agit de particularités présentes, seules les formes directes sont à prendre en compte ; mettre en évidence qu'un ensemble d'objets a comme caractéristique de ne pas posséder une certaine propriété, peut n'avoir aucun sens. Nous restreignons **A***, l'ensemble des formes directes et conjuguées des attributs, aux seules formes *pertinentes* à prendre en compte. Cela revient à éliminer les formes conjuguées qui n'apportent aucune information, c'est à dire à ne pas tenir compte des 0 de certaines variables. Ceci est fait en déclarant au préalable quels sont les attributs dont les deux formes sont pertinentes.

Empruntons maintenant un petit peu au formalisme galoisien. La table de données binaires décrit une correspondance entre objets et formes directes ou conjuguées, pourvu qu'elles soient pertinentes. Pour tout objet x $\in$ X, on définit $f(\{x\})$ comme l'ensemble des éléments de **A*** en correspondance avec x et pour toute partie Y $\subseteq$ X, $f(Y) = \cap_{x \in Y} f(\{x\})$, c'est à dire l'ensemble des valeurs de **A*** qui sont constantes sur Y ou encore la conjonction des caractères communs aux éléments de Y. La conjonction des éléments de f(Y), notée π(Y), est appelée *monôme propre* de Y. Le monôme propre de Y est équivalent à l'*intension* de la classe Y.

Il se peut que π(Y) soit une fonction caractéristique de Y, c'est à dire dans nos notations X(π(Y)) = Y. Dans ce cas Y est une classe monothétique et π(Y) est appelé *monôme caractéristique* de Y. Rappelons qu'une classe est dite monothétique si elle peut être caractérisée par une conjonction d'attributs, autrement dit dans ce formalisme, si aucun élément de X\Y ne donne à π(Y) la valeur "vrai". Si la classe Y est monothétique, c'est que Y est un fermé de la correspondance de Galois entre X et **A***.

Considérons une partition de X en q classes $\{X_1, .. X_q\}$. Par définition on a :

$$\cup_{i=1, q} X_i = X \text{ et pour tout } i \neq j \; X_i \cap X_j = \varnothing.$$

On dit qu'il s'agit d'une *partition conceptuelle* si et seulement si chaque classe X_i de la partition est monothétique, donc que son monôme propre est une fonction caractéristique. En fait c'est une partition de X par ses fermés, souvent appelés concepts, d'où la terminologie employée.

Pour une classe Y, on peut considérer la longueur, notée $l(\pi(Y))$, du monôme $\pi(Y)$ comme une mesure de cohésion de Y et donc de la qualité de sa caractérisation ; Y est d'autant mieux caractérisée que $\pi(Y)$ contient plus de variables. Mentionnons sans plus que l'on peut affiner cette notion en donnant aux variables des poids proportionnels à l'importance qu'on leur accorde dans la description.

Nous appelons *généralisation* de X la formule disjonctive induite par une partition conceptuelle

$$\Phi = \pi(X_1) \vee \pi(X_2) \vee \ldots \vee \pi(X_p)$$

disjonction des monômes caractéristiques des classes de X. A chaque partition conceptuelle de X correspond une généralisation unique.

Exemples

La fonction $\Phi = \pi(x_1) \vee \pi(x_2) \vee \ldots \vee \pi(x_n) = \vee_{i=1, .. n} \pi(x_i)$ est une généralisation de X particulière puisqu'elle n'est vraie que sur les éléments de X.

Pour toute variable A pertinente sous ses deux formes, la formule $\Phi_A = \pi(X(a)) \vee \pi(X(a'))$ est une généralisation puisque tout élément de X possède A sous forme directe ou conjuguée donc $\{X(a), X(a')\}$ est une partition conceptuelle puisque les deux formes sont mutuellement exclusive. Par contre si a' n'est pas pertinente, $\pi(X(a'))$ peut ne pas exister, ou n'être pas caractéristique.

Ainsi définies, on voit qu'il existe un grand nombre de généralisations. En particulier, chaque fois que l'on subdivise une classe suivant un attribut, on crée une nouvelle formule Φ plus spécifique, en ce sens qu'elle couvre moins d'éléments de X, et qu'elle tient compte de plus de variables pour caractériser les deux classes introduites. Nous allons quantifier ce phénomène à l'aide d'un indice très simple déjà présenté dans Guénoche [1989].

En caractérisant Y par $\pi(Y)$ on utilise une partie de la description des éléments de Y. On peut dire que l'on a résumé Y à la conjonction de ses caractères communs, en utilisant une partie de l'information. On quantifie cette part à l'aide d'un *indice de cohésion*, égal au produit du nombre d'éléments de Y par la longueur du monôme $\pi(Y)$.

$$E[\pi(Y)] = l(\pi(Y)).| Y |.$$

La classe Y est d'autant mieux caractérisée que $E[\pi(Y)]$ est grand.

Comme il apparait clairement dans la figure 1, cette indice est la surface du rectangle Y x $f(Y)$ dans la table binaire de correspondance entre X et **A***. C'est le produit des cardinaux de ses intension et extension. On peut étendre sans problème l'indice de cohésion aux partitions conceptuelles par une formule additive. Notons E^i la contribution de la classe X_i. Comme ces classes sont disjointes on peut poser :

$$E[\Phi] = \Sigma_{i=1, .. p} \; E[\pi(X_i)] = \Sigma_{i=1, .. p} E^i$$

3. ALGORITHMES

Ainsi présenté le problème de la généralisation d'un ensemble X se ramène à un problème de partition conceptuelle optimale de X avec, éventuellement, un nombre de classes fixé : Construire $\{X_1, X_2, .. X_q\}$ une partition de X en q classes monothétiques, donc une généralisation Φ telle que :

- pour tout i, $\pi(X_i)$ soit une fonction caractéristique de X_i et
- $E[\Phi]$ soit maximum.

Si l'on raisonne dans la terminologie galoisienne, il s'agit de construire q fermés de X qui en réalisent une partition. Il n'est bien sûr pas question de construire tout le treillis de Galois pour en extraire une partition optimale. Les méthodes décrites ci-dessous sont donc aussi des méthodes pour construire une partie des fermés.

On observe que $E[\Phi]$ est d'autant plus grand que les monômes $\pi(X_i)$ sont de longueur maximum et donc les descriptions en intension de X_i semblables. On retrouve donc le paradigme de la classification : construire des classes d'objets similaires. On notera qu'ici la ressemblance entre deux objets est mesurée par la part de description qu'ils ont en commun. Les deux approches méthodes de subdivision et méthodes de classification ascendantes, classiques en analyse de données, peuvent être adaptées pour réaliser à chaque itération un optimum de E. Sans garantir une optimisation globale, elles réalisent des optimum locaux.

Nous rappelons tout d'abord quelques propriétés élémentaires surtout si l'on raisonne en terme de fermetures.

- Pour toutes classes $Y \subset Y'$, on a $\pi(Y') \subseteq \pi(Y)$. En effet, tout attribut constant sur Y' l'est a fortiori sur Y, donc $\pi(Y')$ est inclus dans $\pi(Y)$. Si les deux monômes sont identiques, Y n'est certainement pas monothétique. C'est la formulation du principe de dualité entre extension et intension ; plus l'une est générale, plus l'autre est spécifique.

- L'union de deux classes monothétiques n'est pas nécessairement monothétique ; leur intersection l'est toujours. L'intersection de deux fermés est un fermé ; par contre pour l'union il n'en va pas de même, comme on peut le constater dans la table binaire ci-dessous. Les monômes propres $\pi(\{1,2\})$ = abde', $\pi(\{3,4\})$ = bde sont caractéristiques, mais $\pi(\{1,2,3,4\})$ = bd ne l'est pas puisque $\{5,6\} \subset X(bd)$.

	A	B	C	D	E
1 :	1	1	1	1	0
2 :	1	1	0	1	0
3 :	0	1	1	1	1
4 :	1	1	0	1	1
5 :	0	1	1	1	0
6 :	0	1	0	1	0

3.1 Méthode ascendante

De même que dans les méthodes ascendantes de classification, on considère initialement que chaque élément de X constitue à lui seul une classe. A chaque étape on va réunir deux des classes les plus voisines, si leur union est monothétique. On voit que la seconde remarque impose quelques vérifications préliminaires avant la fusion des classes.

D'abord il se peut que l'union de deux classes X_i et X_j ait un monôme propre qui prenne la valeur "vrai" c'est une autre classe X_k c'est-à-dire $X_k \subset X(\pi(X_i \cup X_j))$. Il faut donc

réunir simultanément ces trois classes si l'on veut que cette réunion soit monothétique. Soit X_{ij} la classe résultant de l'union de X_i et X_j.

Ensuite il se peut que certains éléments d'une autre classe $X_{k'}$ donnent également la valeur "vrai" au monôme propre de X_{ij}, c'est à dire :

$$X_{k'} \not\subset X(\pi(X_{ij})) \text{ mais } X_{k'} \cap X(\pi(X_{ij})) \neq \varnothing.$$

La classe $X_{k'}$ ne peut faire partie de X_{ij} et $\pi(X_{ij})$ ne sera jamais caractéristique. De ce fait cette réunion $X_i \cup X_j$, qui entraine éventuellement d'autres classes dans un regroupement X_{ij}, ne sera jamais faite.

L'algorithme s'arrête lorsqu'il n'y a plus d'union qui crée une classe monothétique. Dans ce cas on réunit tout le monde en une seule classe. Bien qu'il n'y ait plus de classe complémentaire, on admettra qu'elle est monothétique.

On part donc de la partition en n classes pour laquelle la valeur E_n de l'indice de cohésion est maximum, puisque toutes les formes pertinentes sont utilisées dans la généralisation de X. On aboutit à la partition en une classe, pour laquelle sa valeur E_1 est minimum ; elle est même nulle si X n'a aucune forme commune. A chaque itération, le nombre de classes diminue, ainsi que la valeur de E. Pour achever la définition d'une méthode ascendante, il reste à définir la distance $d(X_i, X_j)$ entre les classes X_i et X_j.

Nous avons noté X_{ij} la classe résultant de l'union de X_i et X_j et définissons d par la part de cohésion perdue :

$$d(X_i, X_j) = \sum_{X_k \subset X_{ij}} l(\pi(X_k)).| X_k | \; - \; l(\pi(X_{ij})).| X_{ij} |.$$

La valeur de d correspond à la surface hachurée dans la figure 1 où X_{ij} est simplement l'union des deux classes .

Soient P_q une partition en q classes et $P_{q'}$ la partition en $q' < q$ classes obtenue en regroupant les classes X_i, X_j de P_q et peut-être d'autres pour assurer la contrainte "monothétique". On pose $E_q = E[\Phi_q]$ et $E_{q'} = E[\Phi_{q'}]$, Φ_q et $\Phi_{q'}$ étant les généralisations correspondant à P_q et $P_{q'}$. On a :

$$E_{q'} = E_q - d(X_i, X_j).$$

Pour calculer l'indice de cohésion de $P_{q'}$, on supprime les parts d'indice qui correspondent aux classes réunies et on ajoute celle correspondant à X_{ij}.

En conséquence, une partition à q classes monothétiques étant donnée, celle à q' classes qui maximise E est bien obtenue en réunissant les deux classes les plus voisines pour cette distance. Soulignons que cette réunion n'est effectuée que si la classe résultante est monothétique.

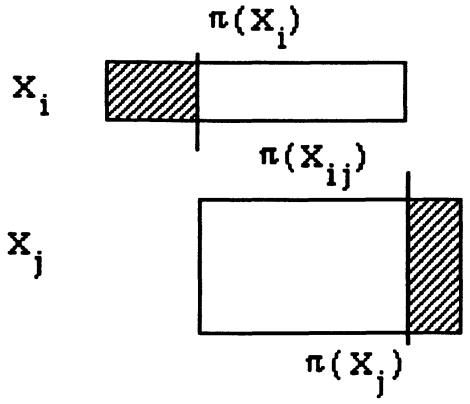


Figure 1. Monôme propre de l'union de 2 classes
Les rectangles faces à X_i et X_j figurent les sous-tableaux binaires de leurs formes communes. Les parties hachurées sont les parts perdues dans X_{ij}.

Cette méthode ascendante s'apparente à la méthode du lien unique [Sokal & Sneath 63], pour une matrice de dissimilarité égale à la distance de la différence symétrique (dans le cas où toutes les formes sont pertinentes), à ceci près :

- A chaque étape il faut calculer les distances de la nouvelle classe aux autres classes inchangées, alors que dans les méthodes hiérarchiques elle est évaluée à l'aide des valeurs initiales, par le maximum, la moyenne ou d'autres formules.

- Nous avons ajouté une contrainte sur la nature des classes, ce qui n'a de sens que si les objets, et les classes restent descriptibles dans un espace de représentation.

Cette méthode dans le cadre booléen est semblable à la *monomial clustering* présentée par Pitt & Reinke [88] comme exemple d'une méthode de classification conceptuelle de complexité polynômiale. Dans leur présentation, les similitudes étaient évaluées une fois pour toutes, au lieu d'être ré-évaluées à chaque apparition d'une nouvelle classe.

Le choix des deux classes à réunir est souvent problématique, car on observe fréquemment, à partir de données binaires, des valeurs identiques de similitude. Afin de maximiser E, il faut réunir à chaque étape le plus grand nombre de classes dont les distances deux à deux sont minimum. Cette procédure, un peu compliquée, s'exprime dans le formalisme des graphes que nous adoptons.

Il est bien connu qu'une dissimilarité peut être considérée comme la valuation des arêtes du graphe complet dont l'ensemble des sommets est X ; la distance d(x,y) est la longueur de l'arête (x,y). Soit δ la valeur minimum de distance. Elle définit un graphe seuil G_δ dont les sommets sont les classes et une arête lie deux classes ssi leur distance est égale à δ. Réunir un nombre maximum de classes revient à déterminer un nombre maximum d'arêtes disjointes. En théorie des graphes un tel ensemble d'arêtes s'appelle un *couplage maximum* dans G_δ. Le problème du couplage de cardinalité maximum, classique en théorie des graphes [Berge 1970], n'est pas détaillé ici ; la méthode utilisée construit un arbre alterné ; il suffit ici de savoir qu'elle est de complexité polynomiale, ce qui n'est pas pénalisant du point de vue du temps de calcul dans l'exécution de l'algorithme.

Exemple
Nous traitons les données de l'exemple initial en considérant que les variables B, C et D ne sont pertinentes que sous forme directe, (puisque ce sont trois modalités mutuellement exclusives sur la nature de l'enveloppe externe de ces espèces), alors que toutes les autres le sont sous les deux formes. Aux 9 espèces décrites correspondent les 9 monômes propres qui sont caractéristiques :

```
 1 : abe'f'g'h      2 : abe'fg'h      3 : abe'fgh       4 : adef'gh'
 5 : ace'f'gh'      6 : ade'f'gh      7 : adef'gh       8 : acef'gh'
 9 : abe'f'gh
```

La valeur initiale de l'indice de cohésion est égale à la somme de leurs longueurs, soit $E_9 = 54$. Les distances calculées suivant la formule ci-dessus donnent :

```
      1  2  3  4  5  6  7  8
2 :   2
3 :   8  2
4 :  28 45 28
5 :  12 24 12  6
6 :   6 15  6  6  6
7 :  12 24 12  2 18  2
8 :  28 45 28  2  2 18  6
9 :   2  8  2 18  6  2  6 18
```

Les valeurs minimum permettent de construire le graphe seuil G_2 ci-dessous dont les arêtes en gras forment un couplage parfait. On obtient :

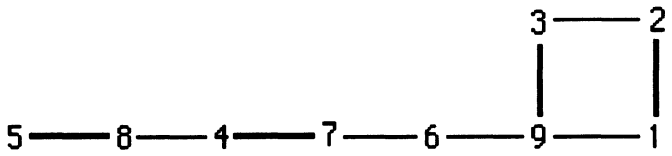


Figure 2. Graphe seuil des distances minimum

```
Classes  1  2  réunies en  10  :  monôme abe'g'h
Classes  4  7  réunies en  11  :  monôme adef'g
Classes  5  8  réunies en  12  :  monôme acf'gh'
Classes  3  9  réunies en  13  :  monôme abe'gh
Valeur de l'indice de cohésion : 46
```

Les distances entre ces nouvelles classes sont recalculées : On obtient le couplage (10, 13) et (6, 11).

```
      10 13 11 12
13 :   4
11 :  41 22
12 :  39 22 14
 6 :  13  4  4 14

Classes  10  13  réunies en  14  :  monôme abe'h
Classes  11  6   réunies en  15  :  monôme adf'g
Valeur de l'indice de cohésion : 38
```

Il ne reste plus que 3 classes. L'union des classes 15 et 12 a comme monôme propre af'g, qui n'est pas caractéristique, puisque le cygne qui est dans la classe 14 possède également ces caractères. Les autres réunions ont pour seul caractère commun a qui est partagé par tous, si bien que les trois classes sont réunies en une seule à la dernière itération.

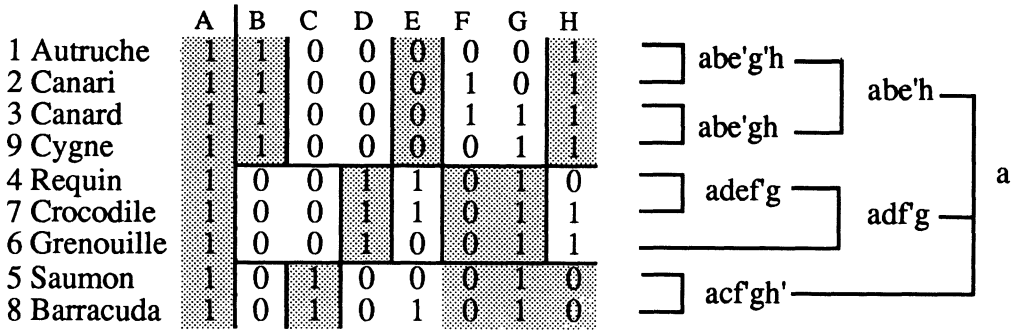
```
      14 15
15 :  29
12 :  33  7

Classes  14  15  12  réunies en  16  :  monôme a
Valeur de l'indice de cohésion  9
```

A la manière des arbres de classification, on peut représenter cette hiérarchie de partitions par le dendrogramme de la figure 3 ; les objets sont ordonnés par un parcours de plus profonde descente de l'arbre de classification.

Une généralisation en 3 classes, séparées par un trait continu, nous donne la fonction disjonctive Φ_3 dans laquelle chaque monôme est bien caractéristique de sa classe :

$$\Phi_3 = abe'h \vee adf'g \vee acf'gh'.$$

$E[\Phi_3] = 4x4 + 3x4 + 2x5 = 38$; la part de description utilisée est grisée.



3.2. Méthode descendante

Pour les méthodes ascendantes, il est bien connu que les partitions en un petit nombre de classes, celles qui sont utilisées dans la pratique, sont les conséquences des choix réalisés en début de procédure. Par contre les méthodes descendantes sont des méthodes de subdivision ; on commence par segmenter X en deux classes, puis chacune d'entre elles, jusqu'à atteindre le nombre de classes cherché, ou que chaque classe soit réduite à un singleton. Dans un espace de représentation booléen, on choisit la variable qui optimise un certain critère, par exemple le Chi2 [Williams & Lambert 1959]. C'est plus ou moins ce que nous allons faire en réalisant encore des subdivisions qui optimisent localement l'indice de cohésion.

Plus généralement, considérons une conjonction de variables, sous forme quelconque, c'est à dire d'un monôme α. A partir de ce monôme on réalise une subdivision de X par l'enchaînement des opérations :

- construire $X(\alpha)$, et son monôme propre $\pi(X(\alpha))$;
- construire le complémentaire de $X(\alpha)$ dans X, noté X' et de même $\pi(X')$.

On a ainsi subdivisé X en deux classes, $X(\alpha)$ et X'. La première est toujours une classe monothétique puisque $\pi(X(\alpha))$ est la fermeture de α ; c'est donc une fonction caractéristique de $X(\alpha)$ dans X. Il n'en est pas toujours de même pour X'. D'abord il se peut qu'il n'y ait aucune forme dans $\pi(X')$; $l(\pi(X')) = 0$ et les éléments de X n'ont aucun caractère commun. Ensuite le monôme propre de X n'est pas nécessairement caractéristique. Dans cette méthode de subdivision, nous nous restreindrons aux bipartitions conceptuelles. Si au moins un attribut est pertinent sous ses deux formes, il en existe toujours.

Il n'est pas envisageable d'étudier les subdivisions correspondant à tous les monômes. On se restreindra :

- aux monômes de longueur 1, c'est à dire que pour toute variable A qui n'est pas dans $\pi(X)$, on étudie la bipartition $\{X(a), X(a')\}$.
- aux monômes de longueur 2. Ces derniers correspondent aux combinaisons des attributs deux à deux. Pour toute paire attributs on calcule les effectifs des cases 00, 01, 10 et 11.

Chaque case "non vide" peut donner lieu à une classe Y. Tous ces tableaux croisés sont calculés initialement et constituent ce que l'on appelle le tableau de Burt. Pour ne pas opposer une classe d'effectif trop faible à sa classe complémentaire qui contiendrait presque tous les objets, on peut se limiter aux cases d'effectif *suffisant* c'est à dire supérieur à un certain seuil ; on peut ainsi fixer un nombre minimum d'éléments par classe, ce qui peut garantir des classes équilibrées. On note $B_X(\alpha) = | X(\alpha) |$.

A l'évidence $\pi(X(\alpha))$ contient α, mais peut être de longueur supérieur à 2 ; il contient alors plusieurs monômes de longueur 2. Ceci nous amène à établir une proposition qui permet de ne pas développer plusieurs fois la même bipartition. De ce fait l'algorithme qui en découle est très efficace.

PROPOSITION

Pour toute classe X et $\{a,b,c\} \subset \mathbf{A}^*$, on a

$$X(ab) = X(ac) \Leftrightarrow c \in \pi(X(ab)) \text{ et } B_X(ab) = B_X(ac)$$

DÉMONSTRATION

$\Rightarrow$ Si X(ab) = X(ac), tous les éléments de X qui possèdent les caractères a et b possèdent aussi le caractère c, donc $B_X(ab) = B_X(ac)$ et, de la même façon, $B_X(ab) = B_X(bc)$ par symétrie. Le monôme propre de X(ab) contient c. De la même façon $b \in \pi(X(ac))$.

$\Leftarrow$ Si $c \in \pi(X(ab))$, alors $X(ab) \subseteq X(ac)$. Mais si de plus $B_X(ab) = B_X(ac)$ alors on a l'égalité.

On examine les monômes dans l'ordre décroissant des effectifs des cases du tableau de Burt que l'on utilise pour marquer les monômes qui conduisent à des bipartitions déjà étudiées. Plus précisément : Si $k \in \pi(X(ab))$ et $B_X(ak) = B_X(ab)$, on posera $B_X(ak) = 0$.

Le principe de cette méthode descendante est de subdiviser à chaque itération une classe. Le choix de cette classe peut dépendre de son indice de cohésion (classe dont la valeur est minimum), du nombre d'éléments (classe de cardinal maximum), de la longueur des monômes caractéristiques. Quel qu'il soit, décrivons la subdivision d'une classe Y ; initialement Y = X.

1. Initialiser une liste **L** de monômes avec les formes directes des variables K qui n'appartiennent pas à $\pi(Y)$, [sinon Y(k) ou Y(k') est vide].

Calculer le tableau de Burt de Y.

Ranger dans **L** les monômes α de longueur 2, tels que $B_Y(\alpha) \geq 1$, dans l'ordre décroissant des effectifs que l'on mémorise également.

2. Pour tout $\alpha \in \mathbf{L}$ faire

Calculer $Y(\alpha)$, $\pi(Y(\alpha))$

Pour tout $i,j \in \pi(Y(\alpha))$, posons $\beta := ij$

Si $\beta \in \mathbf{L}$ et $B_Y(\beta) = B_Y(\alpha)$ supprimer β de **L**

Soit Y' le complémentaire de $Y(\alpha)$ dans Y : Calculer $\pi(Y')$

$E[\Phi_\alpha] := E[\pi(Y(\alpha))] + E[\pi(Y')]$

Pour tout $i,j \in \pi(Y')$, posons $\beta := ij$

Si $\beta \in \mathbf{L}$ et $B_Y(\beta) = | Y' |$ supprimer β de **L**

Si $\pi(Y')$ n'est pas caractéristique de Y' Alors $E[\Phi_\alpha] = - E[\Phi_\alpha]$

3. Choisir parmi les Φ_α la généralisation d'indice E maximum et segmenter Y

Cet algorithme est de complexité polynomiale, puisque l'on examine au plus m+m.(m-1)/2 bipartitions et que pour chacune on calcule le monôme propre de chaque classe par des opérations en O(n.m) et qu'on vérifie avec la même complexité qu'il est caractéristique. L'algorithme est donc dans le cas le pire en $O(n.m^3)$.

Exemple
Traitons les données initiales avec les mêmes formes pertinentes. Les variables B, C, D étant mutuellement exclusives, nous ne considérons pas les croisements de ces variables deux à deux. Le tableau de Burt sur X est :

	AB	AC	AD	AE	AF	AG	AH	BE	BF	BG	BH	CE	CF	CG	CH	DE	DF	DG	DH	EF	EG	EH	FG	FH	GH
00	-1	-1	-1	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	4	2	1	1	3	0
01	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	2	4	5	6	4	2
10	-1	-1	-1	6	7	2	3	4	2	2	0	1	2	0	2	1	3	0	1	3	0	2	1	0	3
11	4	2	3	3	2	7	6	0	2	2	4	1	0	2	0	2	0	3	2	0	3	1	1	2	4

On examine d'abord les partitions de X suivant B, C, .. G et l'on obtient :

X(b) = {1,2,3,9}, π(X(b)) = abe'h, X' = X(b') = {4,5,6,7,8} et π(X') = af'g. Comme nous l'avons déjà vu, cette classe n'est pas monothétique. Les cases ab, be', bh dont l'effectif est 4 sont supprimées de **L** ; aucune autre case n'est supprimée du fait de la classe complémentaire. La correspondance entre une classe et son monôme caractéristique s'écrit souvent comme le produit cartésien de deux ensembles. On notera de façon synthétique ces résultats sous la forme :

B : {1,2,3,9} × abe'h | {4,5,6,7,8} × abe'f ¬ M E = -31 ab, be', bh

Poursuivons l'examen des autres attributs :

C : {5,8} x acf'gh' | {1,2,3,4,6,7,9} x a ¬ M E = -17 ac, cf', cg, ch'

D : {4,6,7} x adf'g | {1,2,3,5,8,9} x a ¬ M E = -18 ad, df', dg

E : {4,7,8} x aef'g | {1,2,3,5,6,9} x ae' E = 24 ae, ef', eg, ae'
F : {2,3} x abe'fh | {1,4,5,6,7,8,9} x af' E = 24 af, bf, e'f, fh, af'
G : {3,4,5,6,7,8,9} x ag | {1,2} x abe'g'h E = 24 ag,ag',bg',e'g',g'h
H : {1,2,3,6,7,9} x ah | {4,5,8} x af'gh' E = 24 ah, ah', f'h', gh'

Il reste à examiner les monômes de longueur 2 non éliminés, et d'abord f'g d'effectif 6 et e'h d'effectif 5.

f'g : {4,5,6,7,8,9} × af'g | {1,2,3} × abe'h ¬ M E = -30

e'h : {1,2,3,6,9} × ae'h | {4,5,7,8} × af'g ¬ M E = -27

Puis les cases d'effectif 4 :

e'f' : {1,5,6,9} × ae'f | {2,3,4,7,8} × a ¬ M E = -17

e'g : {3,5,6,9} × ae'g | {1,2,4,7,8} × a ¬ M E = -17

f'h : {1,6,7,9} × af'h | {2,3,4,5,8} × a ¬ M E = -17

gh : {3,6,7,9} × agh | {1,2,4,5,8} × a ¬ M E = -17

Il reste encore à examiner les cases d'effectif ≤ 2. Nous ne les détaillerons pas. A chaque fois π(Y') = a, donc Y' n'est pas monothétique.

La meilleure des bipartitions étudiées est celle obtenue en coupant suivant B, mais elle n'est pas conceptuelle, non plus que celles suivant f'g ou e'h. On se rabattra donc sur l'une des partitions d'indice 24. Si l'on privilégie les classes d'effectifs équilibrés on choisira la subdivision suivant E, qui isole les espèces à dents, ou encore la subdivision suivant H qui sépare les poissons. Ici on obtient comme généralisation à deux classes l'une des deux formules :

$$\Phi_2 = \{4,7,8\} \times aef'g \vee \{1,2,3,5,6,9\} \times ae'$$

$$\Phi_2 = \{1,2,3,6,7,9\} \times ah \vee \{4,5,8\} \times af'gh'$$

Des poids sur les attributs, ou mieux encore sur les formes, auraient permis de les départager. Ceci peut être fait à l'itération suivante. Si l'on cherche une bipartition conceptuelle de la classe X(ae'), on en trouve 2 d'indice 22, par contre pour la classe X(ah), on en trouve une d'indice 26, qu'on lui préfèrera. On trouve finalement une partition en trois classes, différente de celle de la méthode ascendante mais de même valeur d'indice de cohésion.

$$\Phi_3 = \{1,2,3,9\} \times abe'h \vee \{6,7\} \times adf'gh \vee \{4,5,8\} \times af'gh'$$

RÉFÉRENCES

BERGE, Cl., *Graphes et Hypergraphes*, Dunod, Paris, 1970.

BREIMAN L., FRIEDMAN J.H., OLSHEN R.A., STONE C.J., Classification and Regression Trees, *Wadsworth International Group*, 1984.

GASCUEL O., A. GUÉNOCHE, "Approche symbolique-numérique en Apprentissage", *Actes des journées nationales du PRC-IA*, B. Bouchon Meunier (Ed.), Hermès, Paris, 1990, p. 91-112.

GUÉNOCHE A., "Generalization and Conceptual Classification : Indices and Algorithm", *Data Analysis, Learning symbolic and numeric knowledge*, E. Diday (Ed.), Nova Science Publishers, New York, 1989, p. 503-510.

GUÉNOCHE A., "Classification conceptuelle dans l'Algèbre de Boole", *XXII-ièmes Journées de Statistique*, Tours, 28 Mai-1 Juin 1990.

GUÉNOCHE A., "Méthodes booléennes pour la caractérisation des exemples par les contre-exemples", *Actes des V-ièmes Journées Françaises de l'Apprentissage*, Cassis, 1988, 11p.

MICHALSKI R.S., DIDAY E., STEPP R.E., "A recent advance in data analysis : Clustering Objects into Classes Characterized by Conjunctive Concepts", *Progress in Patern Recognition*, KANAL L.N.& ROSENFELD, A., (Eds.), North-Holland, 1981, p. 33-56.

MICHALSKI R.S., STEPP R.E., "Learning from observation : Conceptual Clustering", *Machine Learning : an Artificial Intelligence Approach*, Tioga, Palo Alto, 1983, p. 331-363.

MITCHELL T., "Generalization as search", *Artificial Intelligence*, 18, 1982, p. 203-226.

NICOLAS J., "Généralisation en logique des prédicats", *Actes des Journées nationales PRC-GRECO I.A.*, Toulouse, Mars 1988, p. 255-274.

PITT L., REINKE. R.E., "Criteria for Polynomial-Time (Conceptual) Clustering", *Machine Learning*, 2, 4, 1988, p. 371-396.

SOKAL R.R., SNEATH P.H.A., *Principles of numerical taxonomy*, Freeman, San Francisco, 1963.

WILLIAMS W.T., LAMBERT J.M., "Multivariate methods in plant ecology", *J. of Ecology*, 47, 1959, p. 83-101.