Élagage et aide à l'interprétation symbolique et graphique d'une pyramide

Kutluhan Kemal Pak, Mohamed Cherif Rahal, Edwin Diday

CEREMADE – Université Paris Dauphine
Place du Maréchal de Lattre de Tassigny
75775 Paris cedex 16
{Pak, Rahal, Diday}@ceremade.dauphine.fr
www.ceremade.dauphine.fr

Résumé : Le but de ce travail est de faciliter l'interprétation d'une classification pyramidale construite sur un tableau de données symboliques. Alors que dans une hiérarchie binaire le nombre de paliers est égal à ***n-1***, si ***n*** est le nombre d'individus à classer, dans le cas d'une pyramide ce dernier peut atteindre ***n(n-1)/2.*** Afin de réduire ce nombre, on élague la pyramide et on utilise un critère de sélection de paliers basé sur la hauteur. De plus on décrit tous les paliers retenus par des variables que l'on sélectionne également en utilisant "**le degré de généralité**" ainsi que des mesures de dissimilarités de type symbolique-numérique. L'aide à l'interprétation se sert d'outils graphiques et interactifs grâce à la bibliothèque OpenGL. Enfin une simulation montre comment évoluent ces sélections quand le nombre de classes et de variables croit.

Mots clés. Classification pyramidale. Classification hiérarchique. Données symboliques. Élagage d'une pyramide. Sélection de variables. Sélection de classes et description. Interprétation d'une classification.

1. Introduction

La classification automatique a pour but la recherche de groupes homogènes, selon un critère bien déterminé, la proximité entre les objets à classer par exemple. Les méthodes de classification automatique sont généralement applicables sur des ensembles de données ou d'objets décrits par des attributs, les habitants d'une ville, les patients d'un service médical… etc. Chaque méthode de classification a ses propres objectifs et sa propre représentation : Arbre, Graphe, Groupement sous forme d'ensembles (Voir (Jain et Dubes (1988))).

Dans le cas de la classification ascendante pyramidale (CAP) qui a été proposée par (Diday 1984), puis développée par (Bertrand (1986)), (Brito (1991)), (Mfoumoune (1998)), (Rodriguez (2000)), (Pak (2004)), et (Rahal (2004)) généralisant la classification ascendante hiérarchique (CAH) (Benzécri (1973)). Il en résulte qu'une représentation en groupes "non disjoints″ et emboîtés d'une pyramide est plus fidèle et riche en information par rapport aux données initiales qu'une représentation de type hiérarchique . Rappelons qu'une pyramide ***P*** construite sur un ensemble *E = {1,2, …,n}* est un ensemble fini de sous-ensembles non vides *{A, B, ….}, (A,B… Í E)* tel que : 1) *E*∈ *P* (le plus grand palier de la pyramide contient tous les individus), 2)Tous les singletons *{1},{2},…{n}* appartiennent à *P 3) " A,B* deux classes de la pyramide ***P*** on a soit *A ÇB = Æ ou A Ç B* Ì *P.* 4) ∃ un ordre *q* compatible avec *P.* Si on définit un index *f(A)* ³ *0* pour chaque classe *A* de *P* tel que *f* est isotonique sur *P : f(A)* £ *f (B)*

si A Ì B, alors la paire *(P, f)* est appelée une ***pyramide indicée.***

Le principe de la CAP consiste, en commençant avec *n* ensembles à un individu (les singletons), à agréger successivement deux classes jusqu'à ce qu'une classe coïncidant avec l'ensemble *E* soit formée, chaque classe pouvant être agrégée deux fois (au lieu d'une seule dans le cas de la CAH), à chaque classe formée est attribuée une hauteur correspondant à la fonction d'indiçage précédemment citée. C'est ainsi qu'une pyramide indicée est construite

L'introduction de l'analyse de données symboliques (Diday (1987)), (Bock, Diday (2000)), a conduit à étendre les méthodes d'analyse de données à variables numériques ou qualitatives à des données mieux adaptées à la description de concepts dont l'extension est formée d'individus décrits de façon standard à l'aide de variables dites "symboliques" car à valeur intervalle, ensemble de valeurs parfois pondérées et munies de règles et de taxonomies. Pour la classification pyramidale symbolique, où chaque palier est un objet symbolique complet[1], (Brito (1991)) propose, en utilisant le **degré de généralités** (utilisé comme mesure de généralisation et comme fonction d'indiçage dans le cas des pyramides indicées symboliques), une généralisation de CAP en gardant les mêmes principes d'agrégation. Nous pouvons également citer les algorithmes CAPS (Classification Ascendante Pyramidale Symbolique) et CAPSO (Classification Ascendante Pyramidale Symbolique avec Ordre donné) présentés par (Rodriguez (2000)).

	y_1	y_2	y_3	y_4
$\omega 1$	[1, 2]	[2, 3]	[1, 2]	[0, 1]
$\omega 2$	[7, 9]	[7, 10]	[0, 6]	[11, 20]
$\omega 3$	[4, 5]	[0, 7]	[-6, 3]	[-30, 30]
$\omega 4$	[-1, 0]	[-7, 2]	[-3, 3]	[10, 100]

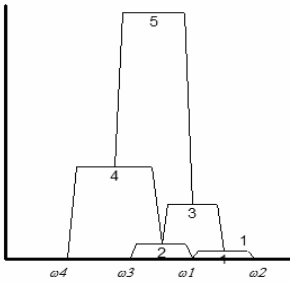

FIG. 1 : *Tableau de données symboliques et la pyramide indicée associée (SODAS2.5[2]-HIPYR)*

Etant donné que la classification pyramidale induit un grand nombre de classes, pouvant atteindre *n(n-1)/2* où n est le nombre d'objet à classer, le but de ce travail est de faciliter l'interprétation d'une classification pyramidale symbolique. Nous présentons dans les deux sections qui suivent deux algorithmes. Le premier consiste à élaguer la pyramide et donc permet une réduction du nombre de paliers à visualiser en éliminant ceux qui sont trop proches les uns des autres en terme de hauteurs et de sauts cet algorithme est proche du point de vue algorithmique de la phase d'affinage effectuée à la fin de l'étape d'agrégation dans la CAH et la 2-3 CAH (Bertrand (2002)) et qui améliore l'algorithme proposé par (Mfoumoune (1998)). Quant au deuxième, qui est un algorithme de sélection de paliers, qui met en évidence les paliers les plus intéressants, en terme de sauts également, selon deux critères de sélection, un critère externe (saut par rapport au prédécesseurs) et un critère interne (saut des

[1] Un objet symbolique est dit **complet** lorsqu'il décrit toutes les propriétés de son extension. (Un objet complet est la partie intentionnelle d'un concept).

[2] Le logiciel SODAS est téléchargeable sur le site :
http://www.ceremade.dauphine.fr/~Etouati/sodas-pagegarde.htm

descendants les uns par rapport aux autres). On peut citer aussi les travaux de (Gordon (1994), (1998)) et de (Jain et Dubes 1988) dans le cadre de la sélection et de l'identification de classes pertinentes dans une classification sur plusieurs méthodes. La quatrième section est consacrée à la sélection et la visualisation des variables explicatives (pour les paliers). Enfin la cinquième présente des tests sur des données réelles et sur des jeux de données simulées.

2. Élagage ou squelettisation des pyramides

Rappelons la notion de successeur et de prédécesseur d'un palier : Soient P_1 et P_2 deux paliers de la pyramide **P**, on dit qu'un palier P_1 est **successeur** de P_2 et P_2 est **prédécesseur** de P_1 si : 1) $P_1 \grave{I} P_2$ (Strictement), 2) Il n'existe pas de palier $P_i \in$ **P** tel que $P_1 \grave{I} P_i \grave{I} P_2$ (strictement). Par construction, un palier p (non réduit à un singleton) possède exactement 2 successeurs, $fils_g(p)$ et $fils_d(p)$ et au plus 2 prédécesseurs $pere_d(p)$ et $pere_g(p)$. Par exemple, dans la figure 1 le palier *5* est le père (prédécesseur) des paliers 4 et 3.

Le but est d'éliminer les paliers dont "l'existence" n'est pas "primordiale" selon un critère défini par un seuil s et de ressortir les paliers (classes) les plus intéressants en fonction de s sans perte d'information. La squelettisation démarre avec une pyramide déjà construite. Le seuil s varie entre 0 et la hauteur maximale $haut_{Max}$ de la pyramide qui est celle de la classe la plus générale donc la plus haute c_{Max}, ou Max représente le nombre total des paliers de la pyramide en question. On autorise la suppression d'un palier si la différence de hauteur avec ses prédécesseurs est inférieure au seuil s, c'est à dire qu'on supprime le palier le moins haut, pour respecter la généralité, qui est l'idée de base de l'élagage. Lorsque ce palier est supprimé, les mises à jour se propagent dans toute la pyramide car la suppression d'un palier provoque la modification de ceux qui deviennent "orphelins" et qui permet de respecter la relation père-fils. Dans le même cadre, (Diday (1984)) a proposé un algorithme d'élagage dit algorithme de suppression d'arètes inutiles dans une pyramide, où l'auteur propose de supprimer les paliers de même hauteur et de les remplacer par un seul palier pouvant être père de «plus de deux», un autre algorithme a été proposé par (Mfoumoune (1998)) ce dernier est limité par certaines conditions, dont le nombre de paliers parents vaut obligatoirement 1. Dans l'algorithme suivant, tous les paliers, quel que soit le nombre des parents du palier examiné, peuvent être supprimés si le seuil s est atteint. L'algorithme d'élagage que nous proposons se déroule en quatre étapes comme suit :

Phase d'Initialisation :

On pose $L = \{c_1, c_2, \ldots, c_{Max}\}$, la liste de tous les paliers (classes) de ***P*** triés selon leurs hauteurs. Etant donné que c_{Max} (le plus haut palier de la pyramide) ne contient aucun prédécesseur, l'algorithme se déroulera sur $Max - 1$ étapes ou itérations. En effet, l'ordre n'a pas vraiment d'importance, le parcours peut se réaliser de manière aléatoire, mais on part du principe que les paliers qui seront supprimés ne doivent pas donner naissance à d'autres, on doit donc commencer par ceux qui sont les plus bas pour empêcher, au maximum, la formation de paliers "inutiles".

Soit c_i le palier de hauteur $haut_i$ examine au passage i. c_i contient au plus deux prédécesseurs c_k de hauteur $haut_k$ et c_l de hauteur $haut_l$, ou $k \neq l$, $k > i$ et $l > i$ donc $haut_k > haut_i$ et $haut_l > haut_i$.

Phase de sélection :

Le critère $cr = min(haut_k - haut_i, haut_l - haut_i)$. Si $cr < s$ alors c_i est à éliminer et on passe

à la phase de mises à jour. Sinon, on passe au palier suivant de la liste *L*.

Phase de mises à jour :

Soit c_k le parent avec lequel le seuil est atteint. On suppose que c_i est le fils gauche (resp. droit) de c_k.

Si c_i n'est pas un palier initial, alors c_g et c_d sont ces deux fils avec $g \neq d$, $g < i$ et $d < i$.

Sinon les deux fils, c_g et c_d, de c_i sont égaux à lui-même avec $i = g = d$.

Le nouveau fils gauche (resp. droit) de c_k devient c_g (resp. c_d) remplaçant c_i. D'une manière symétrique, c_g (resp. c_d) aura un nouveau père c_k. Après avoir défini les nouveaux liens du nouveau palier (mise à jour de la relation père-fils).

Phase de Suppression:

On supprime tous les paliers qui ont une relation, directe ou indirecte, avec le palier c_i autres que c_k et c_g (resp. c_d). On commence par c_i qui sera définitivement enlevé de la liste *L*. Il en résulte que c_d (resp. c_g) n'aura donc plus que le palier c_i comme père et c_l n'aura plus de fils donc c_i qui devra également être supprimé pour éviter les croisements. On détaille cette phase en fonction des paliers c_d et c_l.

- c_d (resp. c_g) et les suppressions : si c_d (resp. c_g) n'a pas d'autre prédécesseur (autre que c_i), c_d (resp. c_g) et tous ses descendants $desc_d = \{c_m, ..., c_p\}$ (resp. $desc_g$) ou tous les $c_j \in desc_d$, $j < d$, devront être enlevés de la pyramide ***P*** et de la liste *L* car c_d (resp. c_g) sera un palier interne ainsi que tous ces descendants droits (resp. gauches) par définition et construction des pyramides. Si la suppression du palier c_j provoque la formation de paliers orphelins (internes) alors c_j est la cible des mêmes modifications que son ancêtre c_d, par conséquent la mise à jour de ***P*** et de *L* en fonction de tous ses descendants.

- c_l et les suppressions : la suppression du deuxième parent de c_i, c_l, provoque la suppression de tous ses prédécesseurs gauches (resp. droit) $c_j \in pred_g(c_l)$. (resp. $pred_d(c_l)$). Pour tout $c_j \in pred_g(c_l)$, son prédécesseur aura comme nouveau fils gauche (resp. droit) le fils gauche(resp. droit) de c_j. En fonction des paliers supprimés et modifiés on met à jour P et la liste *L*.

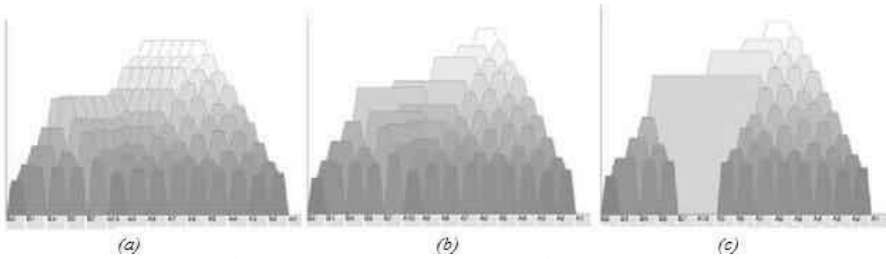

FIG. 2 – *Exemple d'élagage : (a) une pyramide sans élagage, (b) résultat d'un élagage faible, (c) résultat d'un élagage moyen*

Exemple : On examine la pyramide de la figure 2.a: il existe plusieurs paliers de même hauteur et faisant partie de la même descendance. C'est pourquoi, en appliquant l'algorithme précédent avec un seuil très faiblement supérieur à 0, on obtient la figure 2.b. Il n'apparaît plus donc de classes de même hauteur, qu'on peut qualifier d'inutiles. Lorsqu'on augmente la discrimination, donc le seuil, on voit apparaître les classes dans leurs généralités. Par exemple avec un seuil assez important on obtient la figure 2.c : On voit apparaître deux classes C*l*1 et C*l*2 dont les extensions formées par les éléments *{B*1*, B*2*, B*4*, B*5*}* et *{A*1*, A*2*,*

*A*3, *A*4, *A*5, *A*6, *A*7, *A*8, *A*9*}*. L'élément *A*10 jouant les intermédiaires entre les deux classes, s'est vu disparaître lors de la construction de la pyramide élaguée.

Remarques :

- Les suppressions se poursuivent jusqu'à atteindre un palier initial qu'on ne supprime pas de la pyramide pour décrire l'ordre final induit. On qualifiera ces individus d'*inactifs*.
- Le résultat est une **pyramide** car il n'existe aucun palier interne qui resulte des agrégations de paliers *inactifs.*
- Ce n'est pas une hiérarchisation car on supprime des paliers (même initiaux). On obtient une pyramide qui ne tient pas compte de certains individus (en fonction du seuil).
- L'élagage permet d'éliminer l'indiçage large de la pyramide.
- Le nombre de paliers obtenus dépend du seuil, une barre défilante permet de choisir interactivement ce dernier ayant une valeur entre 0 et $haut_{Max}$. On peut ainsi développer des élagages indépendants jusqu'à obtenir la pyramide souhaitée.

3. La sélection de paliers

Dans cette section, nous présentons un algorithme de sélection de paliers, qui est une amélioration de celui de (Rahal et Diday (2004)), où les auteurs s'intéressent uniquement à la sélection de paliers qui ont de grands sauts par rapport à leurs prédécesseur. Dans l'algorithme que nous proposons, contrairement au précédent, nous nous intéressons aux sauts externes (i.e. par rapport aux plus bas pères) mais aussi aux sauts internes (ceux des paliers fils). Deux critères de sélection sont utilisés dans cet algorithme : *Scext*, *Scint.* Le premier pour la sélection des paliers qui ont de grands sauts externes (par rapport à leurs prédécesseurs) et le second pour les sauts internes (par rapport à leurs successeurs).

Dans l'implémentation de la méthode, nous avons utilisé $Scint = m - s$ et $Scext = m + s$, où m est la moyenne des sauts de la pyramide et s l'écart type, l'algorithme de sélection de paliers se déroule comme suit :

Étape d'initialisation :

On parcourt tous les paliers de la pyramide -y compris les singletons- et on calcule les deux sauts pour chaque palier *Pi :* soient *Sautgi* pour le saut vers le palier père gauche et *sautdi* pour le saut vers le palier père droit.

Étape de sélection :

On distingue 2 cas :

Cas 1 : Le palier *Pi* est un palier singleton,

On n'examine pas ses successeurs (car il n'en a pas)

Si *min(sautdi, sautgi) ³ Scext* ***Alors***

Le palier Pi est un palier intéressant

Fin de si

Passer au palier *Pi+1*

Cas 2 : *le palier Pi est formé de l'union de deux paliers fils Pi = Pg È Pd*

Si *min(sautdi, sautgi) ³ Scext* ***alors***

" Pj Ì Pi (3 cas sont possibles)

1 – ***Si*** *P j est un palier extrême gauche* ***alors*** *sautdj = Scint*

2 – ***Si*** *Pj est un palier extrême droit* ***alors*** *sautgj = Scint*

3 – ***Si*** *Pj n'est pas un palier extrême* ***alors*** *Max(sautdi, sautgi) = Scint*

Si *ces trois conditions sont vérifiées pour tous les paliers Pj* ***Alors***

Le palier *Pi* est un palier intéressant

Fin de si

Passer au palier *Pi+1*

Dans la figure 3, nous montrons un exemple de sélection de paliers intéressants à partir d'une pyramide construite à partir de 8 individus. Nous remarquons que les paliers sélectionnés (en rouge sur la figure) ont tous un grand saut par rapport à leurs prédécesseurs et que les paliers internes ont de petits sauts (selon les critères de sélections utilisés).

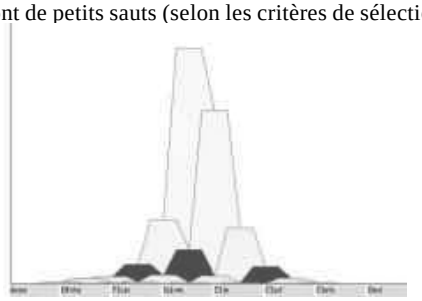

FIG.3 – *Exemple : Sélection de paliers (Scint = m - s* et *Scext = m + s)*.

4. Sélection et représentation de variables

On présente une méthode de description de variables dans le cadre des pyramides symboliques, dont les paliers sont représentés par des objets symboliques complets. Etant donné que le nombre de variables, pouvant expliquer un objet, un individu à classer ou une classe (palier) peut être très grand, le second objectif de ce travail est de réduire l'ensemble des variables descriptives pour chaque classe. Autrement dit, il s'agit de trouver les variables qui "expliquent" le mieux possible un palier par rapport à un autre, dit palier référence. La méthode consiste à sélectionner, parmi l'ensemble des variables qui caractérisent une classe, un sous-ensemble restreint expliquant au mieux les variations locales et globales et de les décrire grâce à une représentation, graphique et conique dont la base est un polygone. On choisit un point *S* qui correspond au sommet du cône, on relie chaque sommet de la base au point *S* pour obtenir ainsi un cône formé d'autant de triangles isocèles que le nombre de variables à représenter, voir par exemple la figure 4 (droite). Chaque coté du triangle formé avec le sommet *S* est gradué de 0 à 1. Cette graduation correspond au rapport de la dissimilarité (ou du degré de généralité) calculé entre la variable représentée par ce triangle et la même variable de la classe référence qu'on définira par la suite.

On privilégie la différence, donc plus la variable contiend d'informations différentes par rapport à la variable référence, plus le triangle est rempli d'une couleur. La méthode qu'on propose peut facilement s'appliquer pour l'explication de n'importe quel objet symbolique si ce dernier peut être comparé à un autre.

Soit : $a_k = \wedge a_{ki}$ et $a_l = \wedge a_{li}$ deux objets symboliques décrits par p variables. Pour visualiser les variables significatives de a_k par rapport à a_l, on peut soit utiliser une mesure de dissimilarité d entre les objets symboliques "élémentaires" relatifs à toutes les variables a_{ki} et a_{li} ou bien utiliser le degré de généralité pour décrire le rapport des généralités entre a_{ki} et a_{li}.

On définit deux sortes de comparaisons suivant la variation de deux objets :

- ***Variation locale :***

Le saut est calculé entre un palier et un de ses prédécesseurs directs, celui avec lequel sa

différence de hauteur est la plus faible. On prend la pyramide de la figure 4 (gauche). On suppose qu'on cherche à décrire les variables du palier sélectionné a_k de couleur rouge. Le parent a_l de plus faible hauteur est celui de couleur verte. Grâce à cette comparaison, on définit le saut d'un palier et on définit l'apport des variables lors de ce saut.

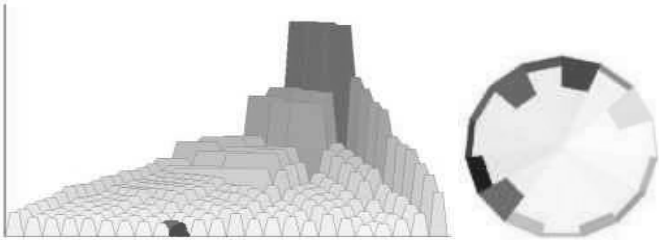

FIG. 4 –*(gauche) Exemple de sélection de palier sur la pyramide, (droite) Cône représentant les variables d'une classe(Vue de haut).*

- ***Variation générale :***

Dans une pyramide symbolique (complète) dans le sens du degré de généralité, le palier le plus haut est une classe qui est composée de l'union de tous les individus initiaux et qui généralise toutes les classes. Dans l'exemple de la figure 4 (gauche), a_k est toujours le palier rouge, tandis que a_l est le palier de plus grande hauteur. Cette version permet de comparer tous les paliers à une référence *unique* et d'avoir une vision sur l'évolution de la généralité des paliers en fonction de leurs variables.

4.1 Mesures utilisées pour la sélection et la description d'une variable

Pour la description et la représentation des variables on a deux possibilités suivant la mesure utilisée : le degré de généralité (Brito 1991) et la mesure de dissimilarité de Minkowski généralisé au cas symbolique (De Carvalho 1996) .

4.1.1 Degré de généralité

Soit p_i un palier de la pyramide ***P*** et soit $S(p_j) = \Lambda a_j$ *(j=1..p)* l'objet symbolique assertion associé, où a_j est un événement élémentaire noté : $a_i = [y_i \in D_i]$. (D_i est la valeur prise par y_j dans l'espace d'observation borné O_j). Le degré de généralité de $S(p_i)$ est définie par :

$$DG(S(p_i)) = \prod_{j=1}^{p} DG(a_j) = \prod_{i=1}^{p} \frac{C(D_j)}{C(O_j)}.$$

Où $C(D_j)$ est le cardinal de D_j si O_j est un ensemble à valeurs discrètes, et la longueur si O_j est continue (intervalle). S'il s'agit d'une variable à valeur histogramme C(D_j) est la moyenne des modalités.

Pour déterminer la hauteur de remplissage *hrj* du j^{eme} triangle correspondant à la j^{eme} variable, on calcule dans un premier temps :

$$ri = \frac{g(a_{kj})}{g(a_{ij})} = \frac{c(D_{kj})}{c(D_{ij})},$$

Le but est de représenter les variations par rapports aux variables références. Etant donné que la hauteur du triangle vaut *1*, et que le rapport des degré de généralité est faible si les variables sont ressemblantes, la hauteur hr_i du taux de remplissage de chaque triangle représentant la variable *j* vaut *:* $hr_j = 1 - r_j$.

4.1.2 La mesure de dissimilarité de Minkowsky généralisée au cas symbolique

On utilise une mesure basée sur la somme pondérée des dissimilarité entre événements élémentaires, la mesure de Minkowski généralisée par De Carvalho (De Carvalho 1996):

$$d(a_k, a_l) = \sqrt[q]{\left[\sum_{i=1}^{p} d(a_{ki}, a_{li})^q\right]}.$$

On recherche à représenter les variations élémentaires et l'apport de chaque variable dans la construction d'un palier, on pose :

$$d(a_k, a_l)^q = \left[\sum_{i=1}^{p} d(a_{ki}, a_{li})^q\right].$$

Pour empêcher que les écarts soient trop faibles on pose :

$$ri = \frac{d(a_{ki}, a_{li})}{d(a_k, a_l)}.$$

La hauteur du taux de remplissage sera définie par la variable *i* vaut *ri* car une dissimilarité décrit en soit la différence, donc plus *ri* est grande plus les variables sont dissociées :

$$hri = ri = \frac{d(a_{ki}, a_{li})}{d(a_k, a_l)}.$$

4.1.3 Réduction du nombre de variables

Il existe le problème du nombre de variables à représenter, deux méthodes sont possibles la première est de représenter uniquement les variables qui ont un *ri* non nul (*ri* ? 0) autrement dit on ne représente que les variables qui ont changé (de valeurs entre la palier en question et le palier référence), sinon nous pouvons utiliser un seuil *s* qui sera par exemple l'écart type (ou la moyenne) des variations des degrés de généralités ou dissimilarités "élémentaires" calculée pour un palier par rapport au palier référence, et représenter les variables significatives par rapport à ce seuil.

Dans le cas élémentaire de la moyenne, la méthode exposée précédemment devient :

On pose : $s = \frac{1}{p}\sum_{i=1}^{p} hri$,

$\forall\, i \le p$, ***si*** $hri \geq s$ ***alors*** le cône contiendra le triangle *i* correspondant à la i^{eme} variable.

On pourait également penser à adapter la notion de ***valeur-test*** au données symboliques, qui va permettre de classer, par ordre d'importance, les variables qui caractérisent au mieux un groupe afin d'identifier une typologie d'individus. Pour obtenir les variables les plus significatives de chaque classe, il faut calculer l'écart entre la moyenne générale de chaque variable et la moyenne de chaque classe. Cette valeur est calculée comme suit :

$$Valeur-test(x_i) = \frac{(\bar{x}_i - \bar{x})}{s_i}$$

Où $\bar{x}$ est la moyenne empirique de la variable, $\bar{x}_i$ est la moyenne empirique de la classe i et s_i est l'écart type de la classe *i*.

Plus celle-ci est grande en valeur absolue, mieux la variable caractérise la classe. Une valeur positive significative correspond à une classe dans laquelle la variable prend des valeurs au-dessus de la moyenne. Au contraire, une valeur négative significative identifie une classe pour laquelle la variable a des valeurs en dessous de la moyenne générale.

4.1.4 Quelques propriétés

– *La rotation :* On peut effectuer des rotations suivant les trois axes de l'espace comme par exemple la fonctionnalité de la figure 5.a(droite) qui est une caption après une transformation suivant l'axe des *x* puis l'axe des *z* du cône de la figure 5.a (gauche).
– *Le zoom :* On peut également effectuer un zoom sur le cône comme le montre l'exemple de la figure 5.b. La deuxième figure n'est qu'une caption du zoom.

Grâce à œs deux propriétés nous pouvons accéder plus facilement aux variables, afin d'étudier leur évolution, lorsque leur nombre est élevé.

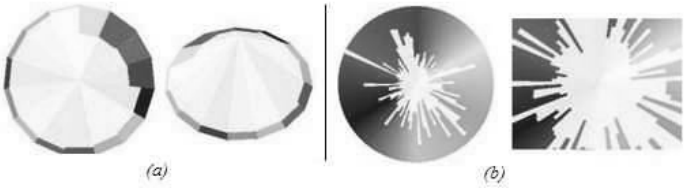

FIG. 5 – *(a) Tri des variables et Rotation, (b) Exemple de Zoom*

5. Tests et résultats

Nous nous proposons dans cette partie d'appliquer l'implémentation des différentes méthodes précédemment présentées. Nous avons effectuées différents tests aussi bien sur des données réelles que sur des données simulées.

5.1 Données réelles

Dans cette partie nous montrons les résultats des méthodes appliquées sur des données météorologiques contenant 27 individus représentant des villes décrites par 24 variables dont 12 variables de type intervalle représentant les températures en degrés centigrades pour chaque mois et 12 autres (continues) représentant la précipitation moyenne par mois en cm^3.

	temp_moy janvier	precip_moy janvier	temp_moy fevrier	..	precip_moy Dec
Berlin	(-2.7,2)	43	(-2.1,3.6)	..	34.1
Bruxelles	(0.7,5.6)	71.2	(0.6,6.4)	..	53
..	..	..	..	..	..
Paris	(0.9,6)	54.3	(1.3,7.6)	..	46.1

FIG. 6 – *Une partie du tableau de données symboliques (météo).*

Les résultats obtenus après élagage et sélection conduisent aux observations suivantes :
La pyramide élaguée de la figure7.2 contient moins de classes, donc moins d'information

mais elle est plus facile à visualiser. L'élagage à 70% donne 9 classes : *C1 = {Las Vegas, Amman, Téhéran}, C2={Lisbonne, Rome, Athènes, Téhéran}, C3 = {Caire, Riadh, Doha, Khartoum, Bamako, Dakar, Havane, Tel-Aviv, Casa Blanca, Lisbonne}, C4={Moscou, Oslo, Helsinki, Copenhague}, C5 ={Zagreb, Genève, Berlin, Amsterdam, Bruxelles, Paris, Copenhague}, C6 = C1 È C2, C7 = C5 È C4, C8 = C3 È C6* et enfin *C9 = C8 È C7.*

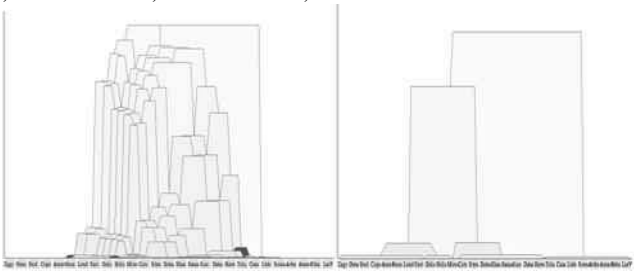

FIG.7.1 –*Sélection de paliers : Pyramide construite sur un tableau de données météorologiques sur 27 villes*

FIG.7.2 – *Résultat d'un élagage à 70% sur la pyramide de gauche.*

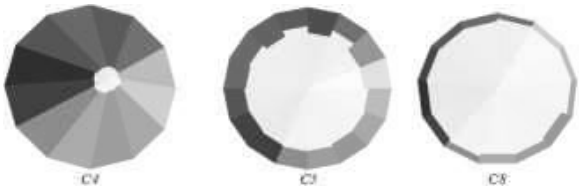

FIG.8 – *Sélection et représentation des variables significatives pour les classes C4, C5, C8.*

Remarquons dans la figure7.2 qui n'est pas une hiérarchie car les intersections entre les classes obtenues (qui est l'un but de la classification pyramidale) ne sont pas vides, par exemple *Copenhague* appartient à deux classes C4 (représenatat les villes qui ont un hiver froid et pluvieux et un été assez doux) et C5 (les villes froides) car entre le mois de Mai et de Novembre elle est plus proche des villes de la classe C4.

Pour la méthode de sélection de paliers nous remarquons 3 classes intéressantes mentionnées en rouge sur la figure7.1, : la classe des villes chaudes ou il pleut moyennement (ou presque pas) qui apparaît également dans les résultats de l'élagage *C8 = {Las Vegas, Amman, Téhéran, Lisbonne, Rome, Athènes, Caire, Riadh, Doha, Khartoum, Bamako, Dakar, Havane, Tel-Aviv, Casa Blanca}*, la deuxième classe obtenue est la classe *C5*, qui apparaît également dans les résultats de l'élagage, des villes moyennement froide ou il pleut beaucoup, la dernière classe est celle des villes *froide C4={Moscou, Oslo, Helsinki, Copenhague}.*

En ce qui concerne les variables, nous nous intéressons à la représentation des variables les plus significatives pour les trois classes *C4*, *C5* et *C8,* nous remarquons dans la figure 8 que le nombre de variables significatives pour ces classes est réduit : 12 variables significatives pour la classe *C4,* 18 pour la classe *C5* et 13 pour la classe *C8,* dans cette représentation, nous n'avons pris que celles qui ont changé (pendant l'étape d'agrégation).

5.2 Données artificielles (simulations)

Tout d'abord nous avons analysé le nombre de classes retenues par la méthode d'élagage en utilisant un seuil à 75% (par rapport au plus haut palier de la pyramide) et la méthode de sélection de paliers. Pour cela nous avons simulé des jeux de données, à l'intérieur d'un carré le nombre d'objets symboliques à classer *n* variant entre 5 et 110, des lois normales assez séparées en utilisant la méthode de Böx Müller en faisant varier la moyenne et l'écart type, nous avons obtenue entre 24.8% et 32.2% de classes retenues avec la méthode d'élagage et entre 10,02 et 13.8 % de classes retenue avec l'algorithme de sélection de paliers (figure 9).

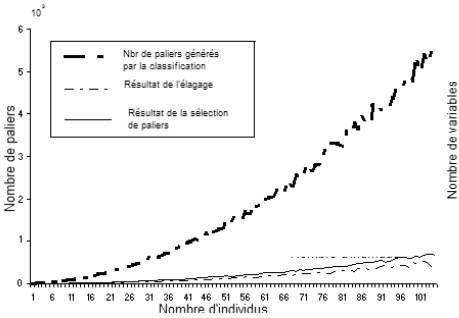

FIG. 9 – *Nbre de palers obtenus en fonction du nombres d'individus à classer selon les 2 méthodes de sélection (élagage et sélection de paliers)*

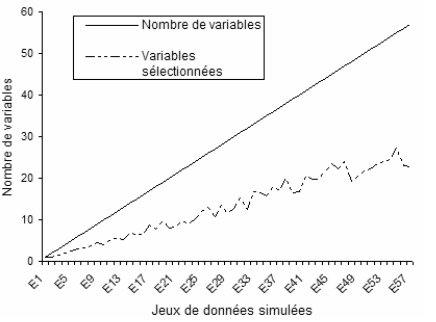

FIG. 10– *Résultats des simulations – variables sélectionnées par rapport au je test (jeux de données n'est pas un*

D'autre part, pour la sélection des variables, nous avons utilisé un autre jeu test en faisant varier le nombre de variables explicatives (dans chacun) . Dans la figure 10 les abscisses représentent les sous-ensembles du jeu test et les ordonnées, le nombre de variables. Nous remarquons que le nombre de variables sélectionnées est nettement plus petit que le nombre total des variables (de l'ordre 15 à 20 %).On remarque enfin que les pentes des nombre de paliers et de variables sont plus grande que celle du nombre de paliers sélectionné, du résultat de l'élagage et du nombre de variables retenues.

6. Conclusion

Dans cet article, nous avons présenté des méthodes pour l'aide à l'interprétation et à la visualisation des classes des pyramides qui s'appliquent également aux hiérarchies. Les deux premières consistent à réduire le nombre de classes en choisissant celles qui s'écartent le plus par rapport au reste de la population et qui sont dense au sens des sauts, et en éliminant celle qui représentent la mesure d'une variation due à un bruit sur les dissimilarités. La troisième est une méthode de sélection et de visualisation de variables qui permet de trouver un sous-ensemble de variables qui a fait qu'une classe s'écarte par rapport aux autres. Il s'est avéré, dans la pratique sur des données réelles et simulées, que les méthodes présentées sont utiles, permettant une interprétation plus aisée, d'une part en ce qui concerne le nombre de paliers à visualiser et à analyser, et d'autre part par rapport au nombre de variables explicatives.

Références

Benzecri J.P. (1973), L'analyse de données. Dunod, Paris, 1973

Bertrand P. (1986), Etude de la représentation pyramidale. Thèse de doctorat, Université Paris IX- Dauphine, 1986.

Bertrand P. (2002), Set Systems for Which Each Set Properly Intersects at Most One Other Set - Application to Pyramidal Clustering. Cahier du CEREMADE numéro 0202. Université Paris Dauphine (IX).

Bock H.H., Diday E.(2000), Exploratory methods for extracting statistical information from complex data. Springer-Verlag.Berlin Heidelberg. 2000. ISBN 3-540-66619-2.

Brito P. (1991), Analyse de données symboliques: Pyramides d'héritage. Thèse de Doctorat. Université Paris IX Dauphine, 1991.

De Carvalho F.A.T. (1996), Histogrammes et Indices de Proximité en Analyse des Données Symboliques, in: Actes de l'Ecole d'Ete sur Analyse des Données Symbolique, Lise - Ceremade, Université Paris - IX Dauphine, Paris.

Diday E. (1984), Une représentation visuelle des classes empiétantes. Rapport INRIA n- 291. Rocquencourt 78150, France, 1984.

Gordon A.D. (1994), Identifying genuine clusters in a classification. Coimputational Statistics and Data Analisys. 18, 561-581

Gordon A.D. (1998), Cluster validation. In Suties in Classification, Data Analysis and Knowledge Organization : Data Science, Classification and Related Methods.Hayashi C, Ohsumi N., Bock H.H. (eds), 22-39, Springer-Verlag, 1998 Tokyo.

Jain AK., Dubes R. (1998), Algorithms for clustering data. Prentice-Hall, Englewood Cliffs.

Mfoumoune M. (1998), Aspects algorithmique de la classification ascendante pyramidale et incrémentale. Thèse de Doctorat. Université Paris Dauphine. 1998.

Pak K.K. (2005), La classification pyramidale symbolique : planaires et spatiales. Thèse de doctorat, CEREMADE, Université Paris Dauphine, à paraître en Décembre 2004.

Rahal M.C., Diday E. (2004), La classification pyramidale symbolique : Sélection de paliers et de varibales. Actes des 11e Rencontres de la SFC Bordeaux, 8 – 10 Septembre 2004, Pages 146-149.

Rodriguez O., Diday.E, (2000). Pyramidal Clustering Algorithms in ISO–3D Project. CEREMADE Paris Dauphine 2000.

Summary

Our aim is to facilitate the interpretation of a pyramid built on a symbolic data table.Whereas in a binary hierarchy the number of clusters is equal to ***n-1***, if ***n*** is the number of individuals to be classified, in the case of a pyramid the number of clusters can reach ***n(n-1)/2***. In order to reduce it we use an algorithm for purring the pyramid and we use criteria of selection of clusters based on the height. Then we describe the selected clusters by variables witch we select also by using "the generality degree" as well as some symbolic-numerical dissimilarity measures, which "explain" them the best. For the representation of the results we use graphical and interactive tools using the OpenGL library. Finally, a simulation shows how these selections envolve when the number of classes and variables grow.

Key words. Pyramidal classification. Hierarchical classification. Symbolic Data. Purring a pyramid. Selection of variables. Selection of classes. Interpretation of a classification.