Un nouvel algorithme de classification spatiale

Mohamed Chérif Rahal

CEREMADE - Université Paris Dauphine, Place du maréchal de lattre de tassigny, 75775 Paris cedex 16
rahal@ceremade.dauphine.fr

Résumé La classification spatiale généralise la classification hiérarchique et pyramidale au cas où les objets à classer sont projetés sur une grille au lieu d'une simple droite. Nous présentons dans cet article un nouvel algorithme qui permet d'obtenir une structure classifiante spatiale en utilisant de nouvelles agrégations.
Keywords : Classification pyramidale, Classification spatiale...

1 Introduction

La classification spatiale de données (Voir [7] et [6]) qui nous intéresse ici, s'applique à tout type de données où n individus sont caractérisés par p variables descriptives (qu'elles soient classiques ou symboliques). Elle a pour objectif d'associer ces individus à chaque nœud d'une grille et d'en extraire simultanément une structure classifiante « compatible » avec cette grille (sans croisement). La classification ascendante spatiale est donc une extension de la classification ascendante pyramidale et hiérarchique où les individus de l'ensemble à classer sont placés sur un maillage planaire (appelé également grille ou réseau) au lieu d'une droite. Le but d'une telle classification est d'offrir en sortie non seulement une structure classifiante (système de classes sous forme d'ensemble) mais aussi un mapping des éléments à classer sur le maillage, qu'il soit régulier ou non . Ceci rend l'interprétation du voisinage entre objet et entre classes plus aisée.

Les pyramides spatiales sont construites de la même manière que les hiérarchies [1] et les pyramides [5] « classiques », sauf qu'à chaque individu de la population à classifier est associé un nœud d'un réseau qui peut être de dimension quelconque au lieu d'être linéaire.

La construction de telles structures s'inspire des méthodes ascendantes de classification - hiérarchique (CAH) (voir [1]), 2/3-hiérarchique (2/3-CAH) (voir [4]) et pyramidale (CAP) (voir [5] et [2]) - et de leurs algorithmes sous-jacents. S'appuyant sur un algorithme ascendant (de la base au sommet), on agrège successivement les objets les plus proches, sous certaines conditions, pour obtenir une structure de graphe spatiale (en trois dimensions). Ce graphe représente aussi bien la structure des classes obtenues que l'indiçage ou la hauteur de ces classes, exprimant ainsi la proximité entre les objets ou les groupes d'objets ayant été agrégés. Les conditions d'agrégation différent selon la structure classificatrice désirée : dans le cas des hiérarchies (planaires ou spatiales) chaque objet ne peut être agrégé qu'une seule fois ; dans le cas des pyramides chaque objet est agrégeable deux fois dans le cas standard et quatre fois dans le cas spatial.

2 Quelques défintions

2.1 Définition d'un m/k-maillage

C'est un graphe dont chaque sommet est le point de rencontre d'un maximum de m arêtes formant m angles consécutifs égaux strictement positifs et dont les plus petits cycles (i.e. ceux qui contiennent le minimum de sommets) contiennent k arêtes de même longueur et forment des « cellules » de surface non nulle qui partitionnent et couvrent l'espace dans lequel il est projeté.

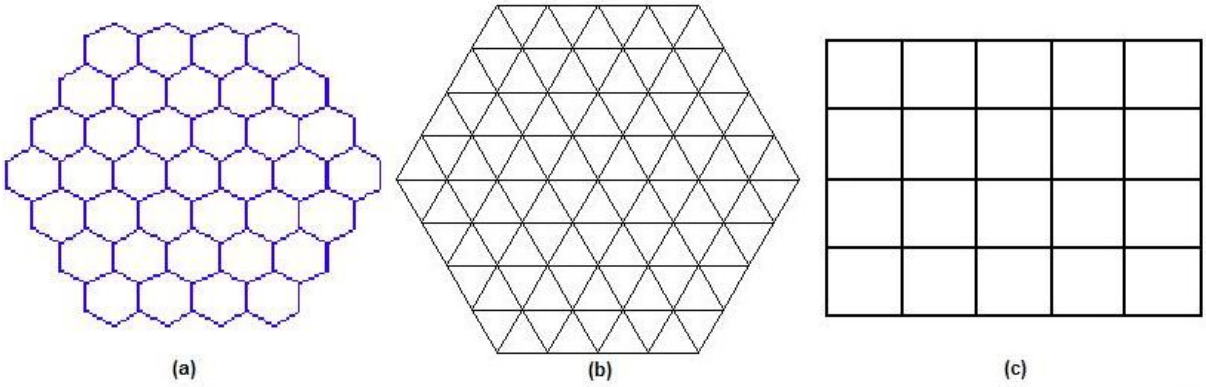

FIG. 1 – Types de grille (maillages) régulières planaires ($2D$) : (a) grille hexagonale, (b) grille triangulaire, (c) grille rectangulaire avec des cellules carrées
.

2.2 Définition d'une pyramide spatiale

Une pyramide spatiale est un ensemble noté $\mathcal{P}$ de parties non vides de E appelées paliers ou classes, satisfaisant les propriétés suivantes :

- $E \in P$;
- pour tout ω dans E, $\{\omega\} \in \mathcal{P}$;
- pour tous p_i et p_j dans $\mathcal{P}$, $p_i \cap p_j = \emptyset$ ou $p_i \cap p_j \in \mathcal{P}$;
- il existe un maillage $\mathcal{G}$ tel que chaque sommet g_{ij} de $\mathcal{G}$ est associé à un élément de E et sur lequel les éléments de $\mathcal{P}$ forment des convexes.

Définition 2.1 ***Composante convexe***
Soit $C \in \mathcal{P}$ associée à un ou plusieurs paliers connectés occupant une sous-grille g de $\mathcal{G}$. g est dite composante convexe ssi :
pour tout couple de nœuds de g le plus court chemin les reliant est dans g.

La figure 2 illustre une pyramide en construction sur une grille régulière de taille 4×6, cette pyramide est constituée de trois paliers maximaux convexes suivants : $A = \{\omega_{20}, \omega_{21}, \omega_{18}, \omega_{19}\}$, $B = \{\omega_{12}, w\omega_{16}, \omega_{17}, \omega_{13}, \omega_{14}, \omega_{15}\}$ et $D = \{\omega_{11}, \omega_{10}, \omega_9, \omega_6, \omega_7, \omega_8, \omega_3, \omega_4, \omega_5\}$ et d'un palier segment $C = \{\omega_1, \omega_2\}$, chacun de ces paliers est représenté par une partie sur la grille. Remarquons que les paliers $A, B, B1, B2, D, D1, D2, D3, D4$ sont des palier auxquels sont respectivement associés les convexes suivants sur la grille :
$Conv(A), Conv(B), Conv(B1), Conv(B2), Conv(D), Conv(D1), Conv(D2)$
$Conv(D3), Conv(D4)$.

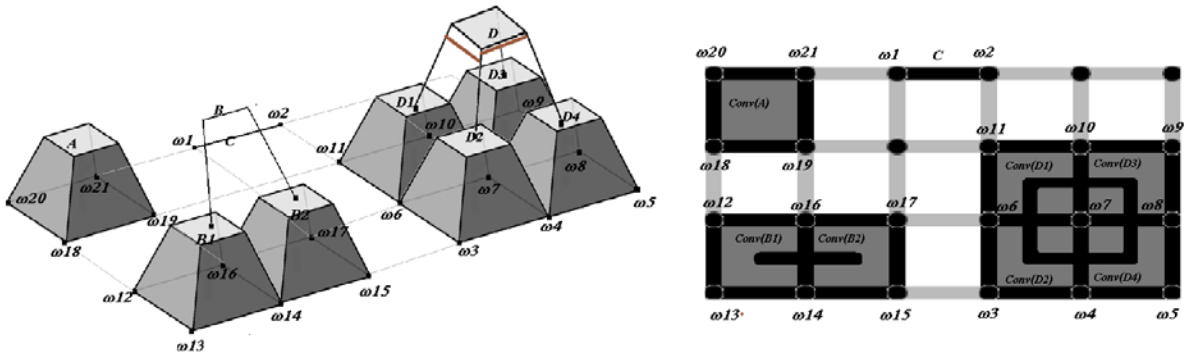

FIG. 2 – Exemple d'une pyramide spatiale en construction, la grille et les différents convexes associés aux paliers

3 Les agrégations spatiales

Nous détaillons et étudions dans ce qui suit les différentes agrégations possibles prises en compte par l'algorithme de classification spatiale que nous présentons dans les prochaines sections. Ces agrégations dépendent d'une part du type de paliers à agréger et de leurs formes d'autres part, par formes nous faisons allusion aux différentes structures « géométriques » qui leur sont associées sur la grille (convexes, segments connexes, ou singletons (un point sur la grille)).

Conditions d'agrégation

On présente dans ce qui suit les conditions d'agrégation de deux paliers. On distingue essentiellement deux cas, le premier est celui où les deux paliers se trouvent dans la même composante convexe, quant au second il est celui où les deux paliers sont dans deux composantes différentes. Dans le premier cas, il faut que les paliers soient connectés ou voisins pour que l'agrégation puisse se faire, dans le deuxième cas les deux paliers doivent être soit maximaux ou à la frontière des composantes convexes qui les contiennent.

Soient h et h' deux paliers de la pyramide en construction $\mathcal{P}$.
Selon les deux cas suivants on dira que h et h' sont agrégeables si les conditions suivantes sont vérifiées :

Cas 1 : h et h' appartiennent à la même composante convexe

- h et h' sont voisins ou connectés.
- h et h' n'ont pas été agrégés quatre fois.

Ou

- h est au nord (resp. au sud) de h' et il n'existe pas de palier h'' tel que h est au nord (resp. au sud) de h'' et h'' est au nord (resp. au sud) de h'.

Ou

- h est à l'est de (resp. à l'ouest) h' et il n'existe pas de palier h'' tel que h est à l'est (resp. à l'ouest) de h'' et h'' est à l'est (resp. à l'ouest) de h'.
- il n'existe pas de palier tel que h ou h' soient à l'intérieur de ce palier.

Cas2 : h et h' n'appartiennent pas à la même composante convexe

Soient C et C' les deux composantes convexes auxquelles appartiennent respectivement les paliers h et h' alors :

- h et h' doivent être à la frontière de C et de C'.
- les composantes convexes relatives à h et h' soient $Conv(h)$ et $Conv(h')$ doivent avoir au moins un coté de taille égale (une bordure de même taille). Cette condition est établie afin de garantir la convexité du palier résultant sur la grille.

4 Cas et types d'agrégations

Il existe dans notre algorithme plusieurs cas d'agrégations, ces cas dépendent d'une part des types de paliers à agréger et des formes ; convexes ou autres ; qui leurs sont associées sur la grille d'autre part. Les cas d'agrégations envisagés dans notre algorithme sont énumérés ci-après :

Construction des paliers segments et de chaînes connectées

Agrégation de singletons

Puisqu'il s'agit d'une classification pyramidale spatiale où chaque palier (i.e. classe) peut être agrégé au plus quatre fois, et afin d'aboutir aux premiers convexes de la pyramide, nous autorisons la construction de paliers connexes. En effet, un palier connexe est dû à l'agrégation de deux singletons en satisfaisant certaines conditions d'agrégation. On agrège deux singletons $\{\omega_i\}, \{\omega_j\}$ s'ils n'ont pas été agrégés quatre fois. Il existe deux cas possibles. Le premier cas est le plus simple, il s'agit d'agréger deux singletons n'appartenant à aucun autre palier déjà construit. Dans le deuxième cas, les singletons ont déjà été agrégés (appartiennent soit à un palier convexe soit à un palier connexe). Nous détaillons dans ce qui suit ces cas de figure :

- **cas 1** : *les deux singletons sont libres (i.e. n'ayant pas encore été agrégés).* C'est le cas d'agrégation le plus simple, dans ce cas on agrège les deux singletons $\{\omega_i\}$ et $\{\omega_j\}$ en un palier connexe $h = \{\omega_i\} \cup \{\omega_j\}$ auquel sera associé un segment connexe $S_{con}(h) = \langle \omega_i, \omega_j \rangle$.
- **cas 2** : *un singleton est libre et le deuxième est déjà agrégé.* Dans ce cas, le singleton ayant déjà été agrégé (moins de 4 fois) peut appartenir à un ou plusieurs paliers connexes ou convexes. Dans le premier cas, il n'y a aucun problème d'agrégation qui se pose, on agrège les deux singletons.

Construction de paliers convexes

- **Agrégation de paliers segments**
  Ce type d'agrégation est purement hiérarchique puisqu'elle vise à former les premiers paliers convexes de la pyramide spatiale à partir de deux paliers segments, on opère exactement de la même manière que dans l'algorithme PyrSpat de Pak [7] pour l'agrégation de paliers « temporaries » afin de former des paliers dits « définitifs ».
- **Agrégation de paliers convexes**
  On agrège deux paliers h et h' constituant respectivement les convexes $g = conv(h)$ et $g' = conv(h')$ sur la grille s'ils ont un coté de même taille autrement dit si l'une des conditions suivantes est vérifiée :

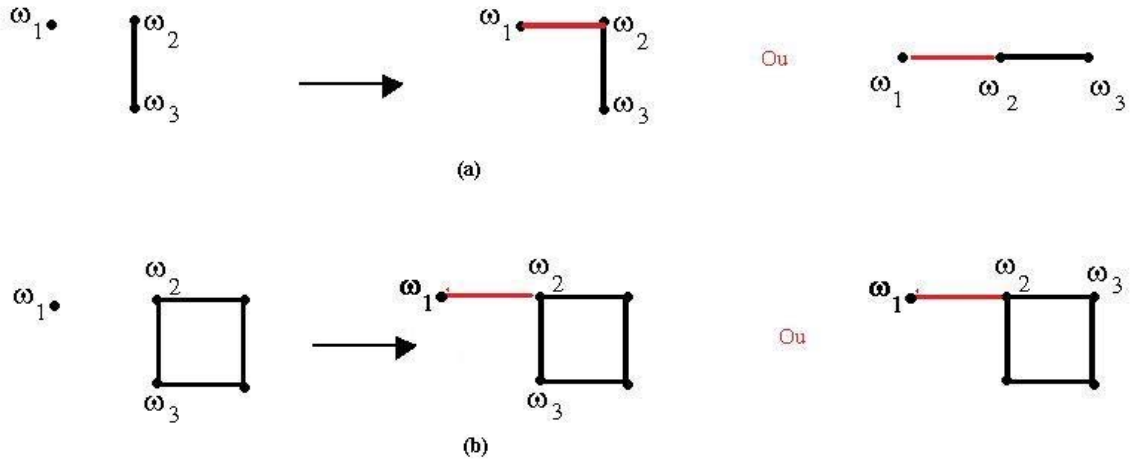

FIG. 3 – Illustration du cas d'agrégation 2 d'un singleton libre avec un singleton déjà agrégé

- g et g' ont la même taille
- $longueur(g) = longueur(g')$
- $largeur(g) = largeur(g')$
- $longueur(g) = largeur(g')$
- $largeur(g) = longueur(g')$

En effet, ces conditions garantissent la convexité sur la grille du nouveau palier résultant.

- **Agrégations convexe-connexe**
  Au début de l'algorithme de classification spatiale, il est nécessaire de construire des paliers segments qui ne seront pas représentés, étant donné qu'ils ne forment pas de convexe et ne respectent donc pas la quatrième propriété de la définition d'une pyramide spatiale. Néanmoins ces paliers doivent être pris en compte pour la construction des premier convexes, nous utiliserons pour ce faire les mêmes conditions d'agrégation proposées dans [7] pour l'agrégation de paliers temporaires, nous obtiendrons ainsi les premiers convexes de la pyramide, mais en plus puisque l'algorithme que nous proposons ici construit directement la pyramide spatiale, nous devons autoriser la possibilité qu'un palier segment puisse s'agréger avec un convexe pour en former un nouveau convexe, ce cas d'agrégation est illustré par la figure 5.

5 S'CAP un nouvel algorithme de classification spatiale

Nous décrivons dans cette section l'algorithme S'CAP de construction d'une pyramide spatiale qui est basé sur l'algorithme présenté dans [6]. Cet algorithme prend en entrée :

- une matrice D de taille $n \times n$ exprimant la distance - ou la dissimilarité - entre chaque couple d'éléments de l'ensemble à classer E
- la taille de la grille $\mathcal{G}_{l \times w}$ sur laquelle seront projetés les individus de E et les convexes

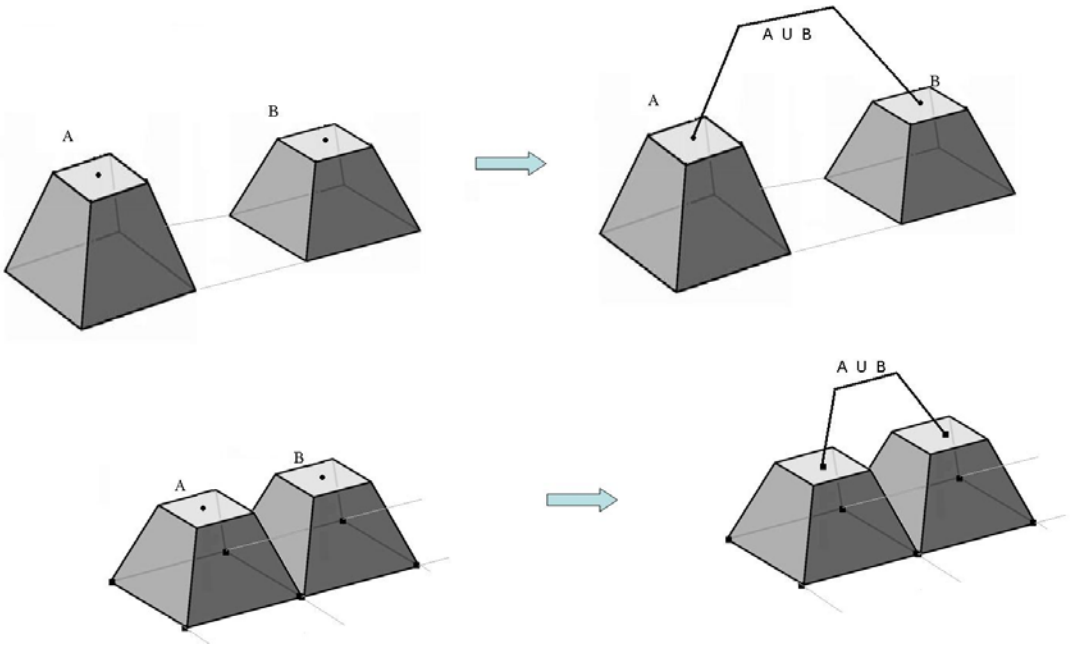

FIG. 4 – Illustration de deux cas d'agrégations convexes (rectangulaire)

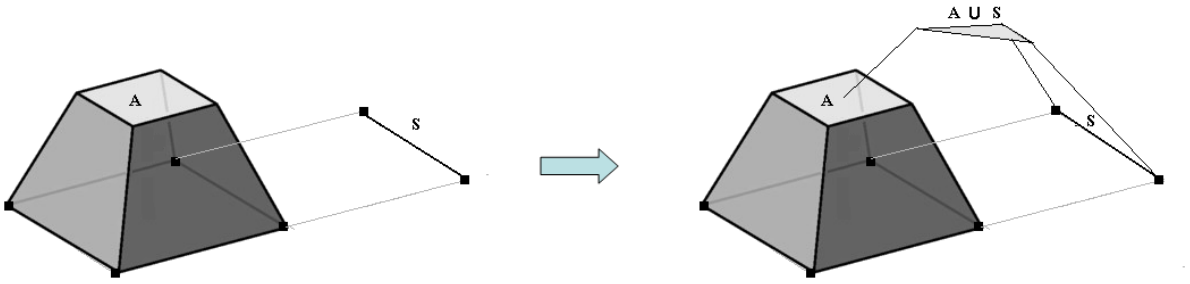

FIG. 5 – Illustration de l'agrégation d'un palier convexe avec un palier segment

associés aux paliers construits
- mesure d'agrégation ρ entre les parties de E.

Dans ce qui suit, nous présentons une formulation du pseudo-algorithme du nouvel algorithme de classification spatiale :

- **Étape 1 :** *Initialisation*
  - Les seuls paliers construits sont les singletons de E, $f(\{\omega\}) = 0, \forall \omega \in E$.
- **Étape 2 :** *Agrégation*
  - On réunit les deux paliers « actifs » h^* et h'^* tel que $\rho(h, h')$ soit le minimum de ρ parmi les paliers h et h' qui vérifient les conditions d'agrégation définies dans 3.
  - A Chaque fois qu'un palier convexe est construit, on le place sur la grille.
- **Étape 3 :** *Mise à jour*
  - On met à jour la liste des candidats
  - On met à jour la grille et les positions des nouveaux convexes s'il y a eu déplacement, rotations ou changement de coordonnées.
  - On met à jour la matrice de dissimilarité en fonction des nouveaux paliers construits.
- **Étape 4 :** *Test d'arrêt*
  - On revient à l'étape 2 tant que le palier coïncidant avec E n'est pas construit.

Lors de la phase d'initialisation, nous créons les singletons de la pyramide et nous initialisons les ensembles des segments Seg (ensemble vide au départ) et l'ensemble des convexes Cv. La pyramide en construction contient au départ de l'ensemble des singletons, on considère que les singletons qui occuperons des nœuds de la grille sont des convexes, ils seront donc intégrés dans l'ensemble Cv.

Étape d'agrégation

Dans cette étape, on cherche les couples de paliers actifs agrégeables qui respectent les conditions de la section 3 et les cas d'agrégation énumérées dans 4 et qui représentent la valeur minimale de la matrice de dissimilarité. On agrège donc ces deux paliers et on passe à l'étape de mise à jour.

Étape de mise à jours

Lors de l'étape de mises à jour
- on recalcule les positions des convexes en cas de déplacement ou de rotation,
- on met à jour l'ensemble des paliers actifs,
- on met à jour les ensembles de successeurs du palier créé et l'ensemble de prédécesseurs des paliers agrégés,
- on met à jour la matrice de dissimilarité.

Remarquons que nous utilisons deux ensembles Cv et Seg dans lesquels sont rangés respectivement les paliers convexes et segments construits au fur et à mesure de l'évolution de l'algorithme. La recherche de paliers agrégeables se fait dans les deux ensembles, puisque selon les conditions d'agrégation, on peut construire plusieurs types de paliers (voir 4). L'algorithme s'execute en trois étapes, une étape d'initialisation qui consiste en la création des paliers triviaux de la pyramide (i.e. les singletons) et qui initialise les ensembles Seg et Cv ainsi que la matrice de distance, une étape double d'agrégation et de mise à jour, lors de l'agrégation on construit les nouveaux paliers de la pyramide en se basant sur la matrice de dissimilarité d'une part et sur les positions et les tailles des convexes d'autre part, enfin l'étape de mise à jour qui permet la réorganisation des paliers nouvellement

construits sur la grille, la suppression des paliers ayant été agrégés quatre fois ainsi que le recalcul de la matrice de dissimilarité.

La différence majeure entre le nouvel algorithme de classification spatiale et PyrSpat est la construction de paliers convexes formant des rectangles sur la grille, ces paliers sont dus aux nouveaux cas d'agrégation introduits dans ce papier (voir les sections 3 et 4).

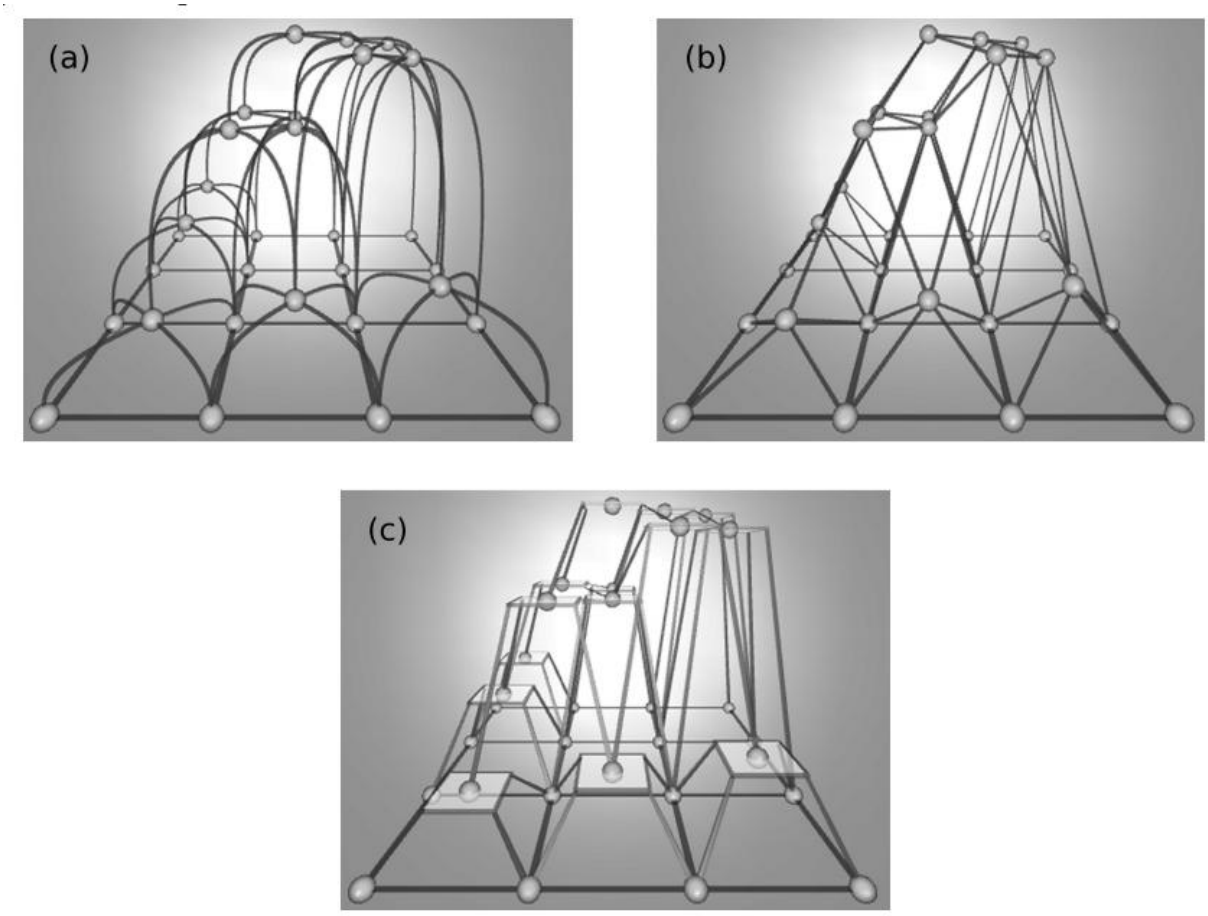

FIG. 6 – Exemple de trois types de visualisation possibles de pyramides spatiales développés dans cadre du Projet ANR SEVEN(LIMSI)

6 Conclusion

Dans ce papier nous avons présenter d'une manière non exhaustive un nouvel algorithme de classification spatiale en utilisant de nouvelles agrégations. Cet algorithme est une généralisation des travaux de [7] et [6], son originalité réside dans le fait qu'on puisse directement construire une pyramide spatiale sans passer par une hiérarchie ou par un ordre prédéfini sur les individus de départ (comme c'est le cas dans [7]) mais aussi permettre la création de paliers convexes rectangulaires ce qui n'est pas traité dans les algorithmes précédents. On cherchera à chaque étape les deux groupes qui minimise la distance deux à deux et dont l'union forme un convexe sur la grille qu'il soit carré ou rectangulaire.

Références

[1] Benzecri, J.-P. : L'analyse des données tome 1 : La taxonomie, Paris : Bordas (1980)

[2] Bertrand, P. et Diday, E. (1990). Une généralisation des arbres hiérarchiques : les représentations pyramidales. Revue de Statistique Appliquée, 38 :53,78.

[3] Bertrand, P. (1986). Etude de la représentation pyramidale. Thèse de 3e cycle, Université Paris Dauphine.

[4] Bertrand, P. (2008). Systems of sets such that each set properly intersects at most one other set-application to cluster analysis. Discrete Appl. Math., 156(8) :1220,1236.

[5] Diday, E. (1984). Une représentation visuelle des classes empiétantes. Rapport de recherche, INRIA.

[6] Diday, E. (2008). Spatial classification. Discrete Appl. Math., 156(8) :1271 ;1294.

[7] Pak, K. (2005). Classifications Hiérarchique et Pyramidale Spatiales. Thèse de doctorat, Université Paris Dauphine.