L'image couleur pour visualiser des données multidimensionnelles

Color Image to Visualize Multidimensional Data

Frédéric Blanchard , Michel Herbin

CReSTIC, Université de Reims, LERI, IUT, rue des Crayères, BP 1035, 51687 Reims Cedex 2
frederic.blanchard@univ-reims.fr, michel.herbin@univ-reims.fr

Manuscrit reçu le 15 juin 2004

Résumé et mots clés

La visualisation de données multidimensionnelles est un problème important. Nous proposons dans cet article d'utiliser l'image couleur pour obtenir une visualisation immédiate et synthétique des données initiales. L'apport de la couleur permet d'exhiber les principales structures de ces données complexes. Après avoir réduit la dimension du problème, notre méthode génére des pixels couleur en utilisant une transformation non triviale inspirée des travaux d'Ohta *et al.* Une dernière étape de tri et d'arrangement de ces pixels dans une image nous permet alors de visualiser nos données multidimensionnelles sur une image couleur.

Image couleur, données multidimensionnelles, visualisation.

Abstract and key words

Image is often considered as the fundamental perceptual unit of a visualization. In this paper, we suggest using one color image to allow an immediate and synthetic visualization of data. The color permits to exhibit the main structures of dataset. After reducing the dimensionality of the dataset, we generate color pixel using a transformation deduced from the work of Ohta *et al*. The last step consists in sorting and arranging pixel into a squared image to provides the final color image that summurizes initial data.

Color image, multimensional data, visualization.

1. Introduction

L'analyse des données multidimensionnelles devient particulièrement complexe quand la dimension et le nombre des données augmentent. Les informations pertinentes sont alors cachées et certains auteurs parlent d'analyse exploratoire [Lebart]. Dans ce cadre, la visualisation est une partie importante du traitement des données et de nombreux outils de visualisation ont été proposés dans la littérature [Minnotte], [Grinstein], [Bonnet]. Certains outils utilisent des représentations dynamiques [Asimov], [Buja] mais, dans ce papier, nous nous restreindrons à des représentations planes et statiques. Dans ce contexte, les techniques de représentation des données demeurent encore très variées et elles font appel à différentes métaphores, icones ou graphes (voir l'état de l'art dans [Grinstein]). On peut noter que la plupart d'entre elles utilisent la couleur comme élément visuel informatif. L'objectif de cet article est de présenter l'utilisation de la couleur pour définir une représentation plane des données multidimensionnelles. La technique de visualisation que nous proposons utilisera l'image numérique et la couleur pour visualiser un ensemble de données multidimensionnelles appartenant à un espace de dimension supérieure ou égale à trois. Si cette dimension est strictement plus grande que trois, une réduction de dimension sera nécessaire. Le choix des couleurs utilisées sera objectif et non supervisé. La technique proposée sera particulièrement adaptée pour des échantillons d'effectifs relativement importants (plus de 10.000 données) mais elle restera utilisable pour des échantillons de taille plus modeste. Elle est présentée pour des données quantitatives quelconques mais se voit simplifiée lorsque l'échantillon est constitué de l'ensemble des pixels d'une image multicomposante.

La visualisation de grands échantillons de données par des techniques orientées image n'est pas une approche nouvelle (voir [Keim]). Elle nécessite de définir une correspondance bijective entre l'ensemble de données et les pixels de l'image. La méthode classique de construction de l'image consiste à classer les données dans un certain ordre, l'ensemble des données forme alors une ligne de pixels successifs (*i.e.* de données successives), il faut ensuite définir le parcours que doit suivre cette ligne pour remplir toute l'image. Parmi les techniques de remplissage d'une image par une courbe, le parcours de Peano est l'un des plus connus (voir [Moon]). C'est cette technique qui sera utilisée dans cet article.

Si l'on souhaite utiliser l'image couleur, il faut aussi affecter une couleur (c'est-à-dire un triplet (R,V,B), Rouge Vert Bleu) à chaque pixel, donc à chaque donnée. Nous procéderons en deux étapes. La première consiste à réduire la dimension des données à trois par projection dans un espace de dimension trois, c'est une étape de réduction de dimension. La deuxième permet d'affecter à chaque triplet (X,Y,Z) qui correspond à une donnée projetée, un nouveau triplet (R,V,B) qui sera la couleur affectée à la donnée initiale. Le choix des couleurs est souvent très subjectif et a pour but de mettre en évidence les informations que l'on pense être pertinentes pour l'observateur [Healey2].

Dans cet article, nous proposons une approche plus objective et non supervisée basée sur une méthode statistique. Il s'agit d'utiliser la tranformée inverse de celle de Otha, Kanade et Sakai [Ohta]. La transformation de Ohta, Kanade et Sakai appliquée aux triplets (R,V,B) d'une image couleur permet d'obtenir de nouveaux triplets (C_1,C_2,C_3) qui approxime la transformation de Karhunen-Loève (TKL). Dans la situation inverse de celle étudiée par Ohta *et al.*, si nous avons trois composantes obtenues par la TKL sur un échantillon de données, nous proposons d'utiliser la tranformation inverse de celle de Otha *et al.* pour d'approximer la couleur. Appliquée à un triplet (C_1,C_2,C_3), cette transformation inverse approximera la couleur (R,V,B) de la donnée à l'origine de (C_1,C_2,C_3). Cette approximation nécessite d'abord d'avoir calculé les trois premières composantes (C_1,C_2,C_3) de l'échantillon de données, ensuite la transformée inverse de Ohta *et al.* donnera la couleur à affecter à chaque donnée.

Dans le paragraphe suivant, nous étudierons la réduction de dimensionalité et montrerons sa nécessité indépendemment des contingences liées à notre méthode de visualisation. Le paragraphe 3 exposera notre méthode d'affectation des couleurs à un ensemble de données projetées dans un espace de dimension trois. Ensuite le paragraphe 4 décrira la construction de l'image présentant l'ensemble des données. Le paragraphe 5 présentera quelques applications. Enfin nous terminerons par une discussion et nous proposerons la conclusion de ces travaux.

2. Réduction de dimensionalité

L'analyse de données multidimensionnelle nécessite une réduction de dimensionalité pour des raisons pratiques liées aux représentations des données mais aussi pour des raisons théoriques que nous rappellerons brièvement.

L'être humain peut concevoir un espace de dimension trois et les progrès des systèmes de visualisation permettent des présentations et des manipulations d'objets 3D. L'exploration d'espaces de dimension supérieure à trois nécessite l'introduction de métaphores comme le temps pour produire des visualisations dynamiques donnant des indications sur une quatrième dimension. L'utilisation d'icones, de graphes, de la couleur ou de textures permet aussi d'augmenter les possibilités d'exploration d'espaces de dimension supérieure à trois [Healey1]. Cependant la dimension trois reste une limitation naturelle de la perception humaine de l'espace. Pour cette raison pratique, on peut considérer qu'une réduction de dimensionalité à trois est optimale par rapport à la perception humaine de l'espace mais cette raison pratique n'est pas la seule.

Notre expérience humaine des espaces euclidiens de dimension trois n'est pas confirmée dans des espaces de dimension plus élevée [Landgrebe]. Les métriques classiques ne permettent pas

de révéler les structures spatiales des données multidimensionnelles dont la dimension est élevée (supérieure à cinq) [Aggarwal]. Par exemple, considérons la distance euclidienne dans $\mathbb{R}^n$ où la dimension est égale à n. Soit trois points de $\mathbb{R}^n$: l'origine A qui a pour coordonnées $(0,0,...0)$, le point B qui a pour coordonnées $(1,0,...0)$ et un point A' de coordonnées $(\varepsilon,\varepsilon,...\varepsilon)$ où ε est un nombre positif très petit. Comparons les distances AA' and AB. Quand la dimension n est inférieure à trois, on a AA' $\ll$ AB (AA' est plus petit ou égal à $\varepsilon\sqrt{3}$ et AB est égale à 1). Quand la dimension n croit, la distance AB reste égale à 1 mais la distance AA' peut devenir plus grande que AB. Nous ne pouvons pas extrapoler notre perception des espaces 2D ou 3D aux espaces de dimension supérieure. Prenons un exemple plus classique (voir par exemple [Verleysen] ou [Lennon]). En dimension n, l'hypervolume de la boule de rayon r est donné par l'expression :

$$Vol(B_r^{2p}) = \frac{r^{2p}\pi^p}{p!} \quad \text{si } n = 2p$$

$$Vol(B_r^{2p+1}) = \frac{r^{2p+1}2^{2p+1}\pi^p}{1.3.5\ldots(2p+1)} \quad \text{si } n = 2p+1$$

Quand n est plus grand que cinq, le volume de l'hypersphère décroit vers 0 si la dimension augmente. En revanche, si par exemple $r = 0.5$, cette hypersphère est incluse dans un hypercube dont l'hypervolume reste égale à 1. Quand la dimension augmente, le ratio entre l'hypervolume de l'hypercube englobant l'hypersphère et celui de l'hypershère tend rapidement vers l'infini. Ces observations sont regroupées sous le vocable *malédiction de la dimensionalité* (curse of dimensionality [Bellman]) et le phénomène de Hughes [Hughes] montre l'impossibilité d'analyser un échantillon de données quand la dimension de l'espace dans lequel se situe ces données devient trop importante. En d'autres termes, pour un échantillon d'observations de taille fixe, à partir d'un certain point il devient contreproductif de multiplier les mesures ou les variables pour observer un phénomène. Ajouter des variables n'ajoute alors pas d'information mais au contraire en fait disparaître. Ces aspects théoriques montrent la nécessité de réduire la dimension quand celle-ci est trop importante. La recherche de la dimension intrinsèque (la dimension idéale) est l'objet de nombreux travaux [Dohono], [Verveer], [Camastra] mais la plupart des auteurs réduisent la dimension sans rechercher une hypothétique dimension intrinsèque [Debacker], [Devijver].

Dans cet article nous proposons une approche très classique, simple et généralement efficace de la réduction de dimensionalité : nous conservons les trois premières composantes générées par une Analyse en Composantes Principales (ACP) [Rao]. Le principe de l'ACP est de projeter les données initiales dans un sous espace de dimension réduite k (dans ce papier $k = 3$). Ce sous-espace est optimisé pour maximiser l'inertie du nuage des données projetées. Le sous-espace est engendré par les k vecteurs propres correspondants aux k plus grandes valeurs propres de la matrice de covariance. Si l'on utilise la matrice de corrélation, on parle alors de la transformée de Karhunen-Loève (TKL). Par la projection par ACP, les axes de projections sont orthogonaux (et décorrélés dans le cas particulier de la TKL). Les techniques dites de *projection pursuit* [Nason] constituent un autre moyen d'obtenir des projections orthogonales en optimisant un index de projection. Si l'orthogonalité n'est pas nécessaire, l'Analyse en Composantes Indépendantes (ACI) [Hyvären] peut aussi permettre de projeter les données dans un sous-espace obtenu en maximisant une fonction de contraste [Comon]. D'autres techniques non linéaires comme l'Analyse en Composantes Curvilignes [Demartines] ont été développées ; toutes ces techniques ne seront pas comparées dans cet article, nous nous restreindrons à l'utilisation de la TKL pour définir un sous-espace de dimension trois.

3. Les couleurs d'un ensemble de données

Une couleur est définie dans cet article par un triplet de valeurs : Rouge, Vert et Bleu. Comme classiquement en imagerie, ces trois composantes sont codées sur un octet (un entier non signé entre 0 et 255). Dans cette partie, nous proposons une solution pour associer une couleur à chaque donnée d'un échantillon. Les données ayant été projetées dans un espace de dimension trois, chaque donnée a un représentant (X,Y,Z) dans cet espace. Nous décrivons comment nous associons une couleur (R,V,B) à chaque représentant (X,Y,Z) d'une donnée. Une correspondance naïve du type X pour R, Y pour V et Z pour B peut conduire à des interprétations erronées des couleurs présentées. Aussi nous expliquons d'abord pourquoi ces approches naïves sont à éviter puis nous proposons une approche objective non triviale du calcul de la couleur d'une donnée sans aucun *a priori* sur la palette des couleurs qui sera utilisée.

La méthode naïve consiste à rééchantillonner X, Y et Z pour obtenir trois valeurs entre 0 et 255. Cette normalisation est obtenue par :

$$\begin{cases} X' &= 255 \times \dfrac{X - X_{min}}{X_{max} - X_{min}} \\ Y' &= 255 \times \dfrac{Y - Y_{min}}{Y_{max} - Y_{min}} \\ Z' &= 255 \times \dfrac{Z - Y_{min}}{Z_{max} - Z_{min}} \end{cases}$$

où X_{min}, Y_{min} et Z_{min} sont respectivement les valeurs minimales de X, Y et Z sur l'échantillon de données et X_{max}, Y_{max} et Z_{max} sont respectivement les valeurs maximales de X, Y et Z sur ce même échantillon. Nous pourrions considérer que X', Y' et Z' représentent R, V et B à une permutation près. Malheureusement une approche de ce type n'est pas satisfaisante, les couleurs obtenues ne permettent pas de percevoir les structures ou les classes de données présentes sur l'échantillon.

Même l'ordre dans lequel sont considérés X', Y' et Z' n'est pas satisfaisant, aucun argument ne permet de préférer (X',Y',Z') à (Y',Z',X'). Prenons un exemple pour éclairer ce point. Les triplets $(255,0,0)$ et $(255,255,0)$ seront affichés et perçus comme respectivement du rouge et du jaune, alors que les triplets $(0,0,255)$ et $(0,255,255)$ seront affichés et perçus comme respectivement du bleu et du cyan. En considérant une métrique comme la distance euclidienne sur ces données projetées, on constate que les distances d'une part entre le rouge et le jaune et d'autre part entre le bleu et le cyan sont égales. En revanche la perception humaine de ces couleurs n'est pas la même. Le bleu et le cyan sont en effet plus proches entre eux que le rouge et le jaune. Ce type de représentation colorimétrique conduit donc à des rapprochements de données qui ne seront pas justifiés mais uniquement liés à une méthode de visualisation et au choix des couleurs.

Les triplets (R,G,B) et (G,R,B) ne représentant pas les mêmes couleurs, l'ordre des composantes est significatif vis-à-vis de la perception que l'on en aura. Hélas les valeurs normalisées Z', Y' and Z' n'ont aucun ordre *a priori* et, sans connaissance complémentaire, ces coordonnées ont le même poids au sens statistique. Pour ordonner ces trois variables, nous proposons une approche statistique en calculant les trois premières composantes C_1, C_2 et C_3 obtenues par la TKL appliquée à l'ensemble des données (voir partie *réduction de dimensionalité*). Ces composantes sont ordonnées, la première étant plus informative au sens statistique que la deuxième et celle-ci plus informative que la troisième. Le triplet (C_1,C_2,C_3) donne des informations statistiques qui ne peuvent pas être considérées directement comme des informations colorimétriques.

Ce papier propose de choisir les couleurs par des résultats connus en analyse d'images couleur. Otha, Sakai et Kanade [Ohta] ont proposé une transformation linéaire de l'espace (R,V,B) qui simule la tranformation de Karhunen-Loève pour les pixels d'une image couleur. À partir de (R,V,B), ils proposent une approximation des trois composantes de la TKL : (C_1,C_2,C_3). Nous sommes dans la situation inverse où nous disposons des composantes (C_1,C_2,C_3) obtenues par TKL sur un échantillon de données. En considérant la transformation inverse de celle de Ohta *et al.*, nous obtenons les triplets (R,V,B) qui approximent *Rouge*, *Vert* et *Bleu* de l'image couleur virtuelle qui aurait conduit à (C_1,C_2,C_3). Ces triplets sont définis par :

$$\begin{cases} R &= (6C_1 + 3C_2 - 2C_3)/6 \\ V &= (3C_1 + 2C_3)/3 \\ B &= (6C_1 - 3C_2 - 2C_3)/6 \end{cases}$$

À partir d'un échantillon de données décrites dans un espace de dimension trois par les triplets (X,Y,Z), nous obtenons après normalisation, TKL et transformation inverse de celle de Otha, les triplets (R,V,B) que nous proposons comme définition des couleurs des données relativement à l'échantillon considéré.

4. Construction d'une image

Dans le cas particulier où l'échantillon de données est constitué de l'ensemble des pixels d'une image multicomposante, on obtient immédiatement une image couleur à la fin de l'étape précédente. Chaque pixel multidimensionnel est représenté par un pixel couleur de composantes (R,V,B) dans l'image résultat. Dans le cas général, des données multidimensionnelles quelconques ne sont pas les pixels d'une image multicomposante et ne possèdent donc pas d'information de localisation dans l'espace image. Autrement dit, les données obtenues à l'étape précédente ne possèdent pas, *a priori*, de position dans l'image couleur finale. Cette partie de l'article propose une construction d'image pour visualiser ces données non initialement spatialisées.

Pour construire une image de l'échantillon de données, on associe un pixel de l'image à chaque donnée (voir *pixel-oriented visualization* [Keim]). Un ensemble de N données sera donc représenté par une image de N pixels. Cette approche de la visualisation orientée-pixel permet de représenter des échantillons de grande taille. Par exemple, une image 1000×1000 permet de représenter un million de données. On suppose que chaque donnée est représentée soit par un niveau de gris (un entier non signé entre 0 et 255) soit par une couleur (un triplet (R,V,B)) qui lui est associé. Les pixels (donc les représentations des données) doivent être placés spatialement dans l'image, ce paragraphe de l'article présente comment les coordonnées de chacun des pixels sont déterminées.

Les pixels ne peuvent pas être placés arbitrairement dans l'image. En effet, ceci produirait une image où les données seraient dispersées et l'oeil humain ne pourrait que très difficilement identifier des groupes ou classes de données. Les similarités ou dissimilarités entre données seraient alors difficiles à déterminer, les rapprochements entre données dispersées étant difficiles à effectuer. Pour que l'image soit lisible au premier coup d'oeil de manière très intuitive, il faut que les données similaires soient spatialement très proches, les données dissimilaires éloignées et que les classes de données soient le plus connexes possible. Cette information spatiale contenue dans l'image sera redondante avec l'information colorimétrique que nous avons proposée précédemment. En effet, plus les données seront similaires et plus leurs couleurs seront proches mais aussi plus leurs localisations dans l'image seront proches. Ce n'est qu'à cette condition de redondance que la lecture de l'image sera intuitive et immédiate. Nous proposons une construction de ce type d'image avec deux étapes : les pixels (*i.e.* les couleurs associées aux données) seront d'abord rangés de manière à former une ligne de pixels successifs, ensuite cette ligne sera utilisée pour remplir l'image.

Ranger les pixels (c'est-à-dire trouver un ordre pour l'ensemble des pixels) équivaut à projeter les données sur un espace de dimension un. Les données sont alors ordonnées ou rangées par

l'ordre naturel sur l'espace $\mathbb{R}$ (espace de projection de dimension un). Il serait possible de refaire une étude sur la meilleure projection sur un espace 1D mais nous avons déjà proposé une projection 3D par la TKL. Avec ce type de transformation, la meilleure projection 1D sera obtenue par la première composante principale. Cependant, pour tenir compte des trois premières composantes principales et non uniquement de la première, nous rangeons les données par un tri avec trois clefs successives qui sont les trois premières composantes principales de la TKL. Les représentations de ces données (*i.e.* les pixels) se trouvent ainsi rangés formant une ligne de pixels successifs.
Dans la deuxième étape de la construction de l'image représentative de l'ensemble des données, il faut remplir l'image par cette ligne de pixels successifs. Les courbes de Peano constituent le moyen le plus classique pour effectuer cette construction [Moon], les courbes en « U » de Hilbert ou les courbes en « Z » de Morton sont les plus utilisées. Le principal avantage de ces courbes est de préserver au mieux la connexité des classes de données [Sasov]. La procédure de construction de telles courbes est récursive, nous utiliserons dans cet article les courbes en « U » de Hilbert (voir Figure 1).
Les deux étapes de rangement des données puis de parcours de l'image par une courbe d'Hilbert évitent de disperser les pixels dans l'image construite. Cette approche tend à préserver la cohérence spatiale des données dans leur représentation par une image permettant ainsi une visualisation très intuitive des ensembles de données.

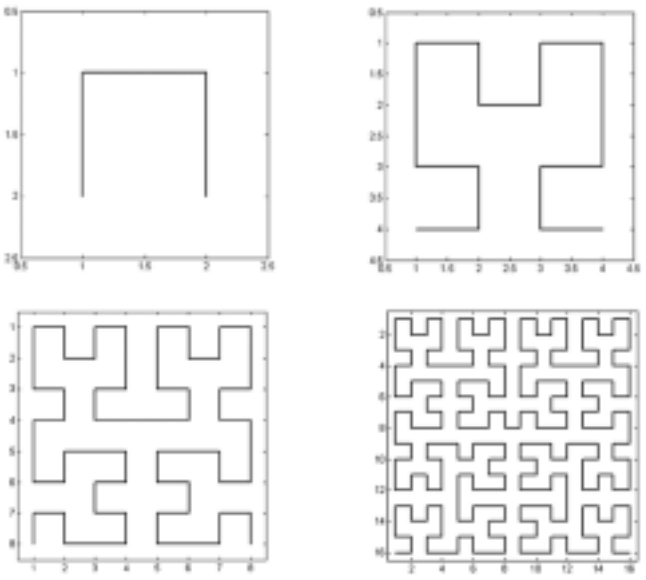

Figure 1. Etapes de la construction d'une courbe de Hilbert.

5. Applications

5.1. Visualisation des classes de données

La méthode de visualisation décrite dans cet article a d'abord été appliquée en conditions contrôlées. Pour cela, des échantillons de données ont été simulés. Ces données synthétiques permettent de vérifier l'efficacité de la méthode de visualisation.
Dans l'exemple que nous proposons (voir Figure 2), nous avons simulé 12 classes de données (non bruitées pour faciliter la lecture) et nous vérifions que notre méthode de visualisation permet de visualiser ces 12 classes rapidement de manière très intuitive. Détaillons cet exemple. Nous considèrons 65.536 données dans un espace de dimension six. Les six variables décrivant ces 65.536 observations prennent quatre valeurs possibles. Les 65.536 données ou observations sont préalablement rangées formant ainsi une image 256×256 de six composantes. Ainsi chaque composante est visualisée par une image de 65.536 pixels ayant quatre niveaux de gris (voir Figure 2). Chaque donnée étant déjà positionnée dans l'espace image, la simple application de la méthode de détermination des couleurs permet d'obtenir directement une image couleur de 256×256 pixels (voir Figure 2).
Ces données ayant six composantes prenant quatre valeurs chacune, l'ensemble des données peut potentiellement présenter 4^6 classes de données. Une bonne méthode de visualisation devrait

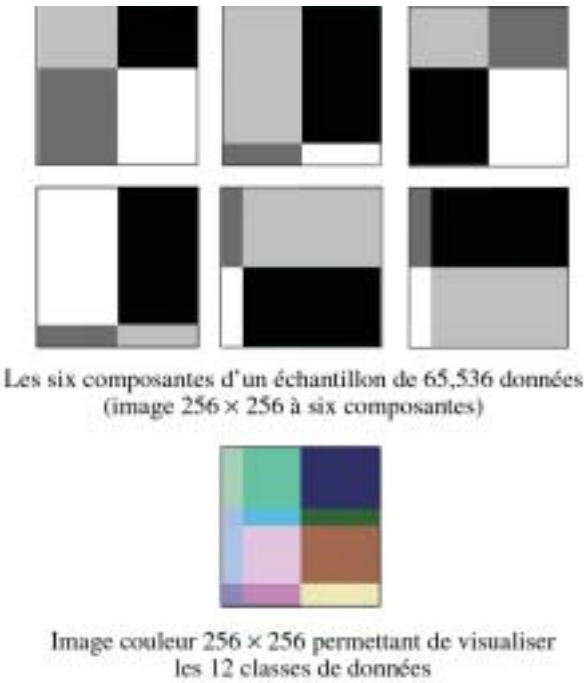

Figure 2. Visualisation couleur de 65.536 données d'un espace de dimension six.

permettre d'identifier au premier coup d'oeil les différentes classes présentes parmi les données pour rendre compte des similarités entre ces données. Sur cet exemple il n'y a que 12 classes simulées. Ces 12 classes sont visualisées sur l'image couleur résultat (voir Figure 2). Cet exemple très simple montre l'efficacité de la méthode de visualisation pour présenter les différentes classes de données (*i.e.* les similarités de données) par des couleurs. Les résultats obtenus seraient trivialement confirmés par la plupart des méthodes de classification mais la visualisation présente l'avantage d'être immédiatement lisible par l'oeil humain et facile à interpréter. Le système de visualisation permet de classer les données et de leur affecter un label sous forme d'une couleur associée à chacune des 12 classes de l'échantillon. Ces couleurs indiquant une similarité entre données, elles donnent aussi une indication sur la proximité des différentes classes. Dans cet article, nous avons pris un exemple très simple pour montrer l'efficacité de la méthode, des exemples plus complexes avec des données simulées donnent des résultats similaires. Cet outil de visualisation se révèle bien adapté aux grands échantillons de données multidimensionnelles ; il permet de visualiser les principales structures de l'échantillon lorsque la réduction de dimension à trois conserve l'information sur ces structures.

5.2. Visualisation d'images multicomposantes

Dans le cas réel d'images multicomposantes, nous observons des résultats semblables : les différentes classes de données (dans ce cas des ensembles de pixels) sont le plus souvent immédiatement visualisées par notre méthode avec une simple image couleur. Pour illustrer ce propos nous considérons un ensemble de 14 cartes de concentration d'élements chimiques d'un spécimen de granite enregistrées en fluorescence X [Wekemans]. Cet ensemble constitue une image de 14 composantes et de 39×40 pixels. Nous pouvons appliquer notre méthode de visualisation aux 1560 données (ou pixels) appartenant à un espace de dimension 14 (voir Figure 3).

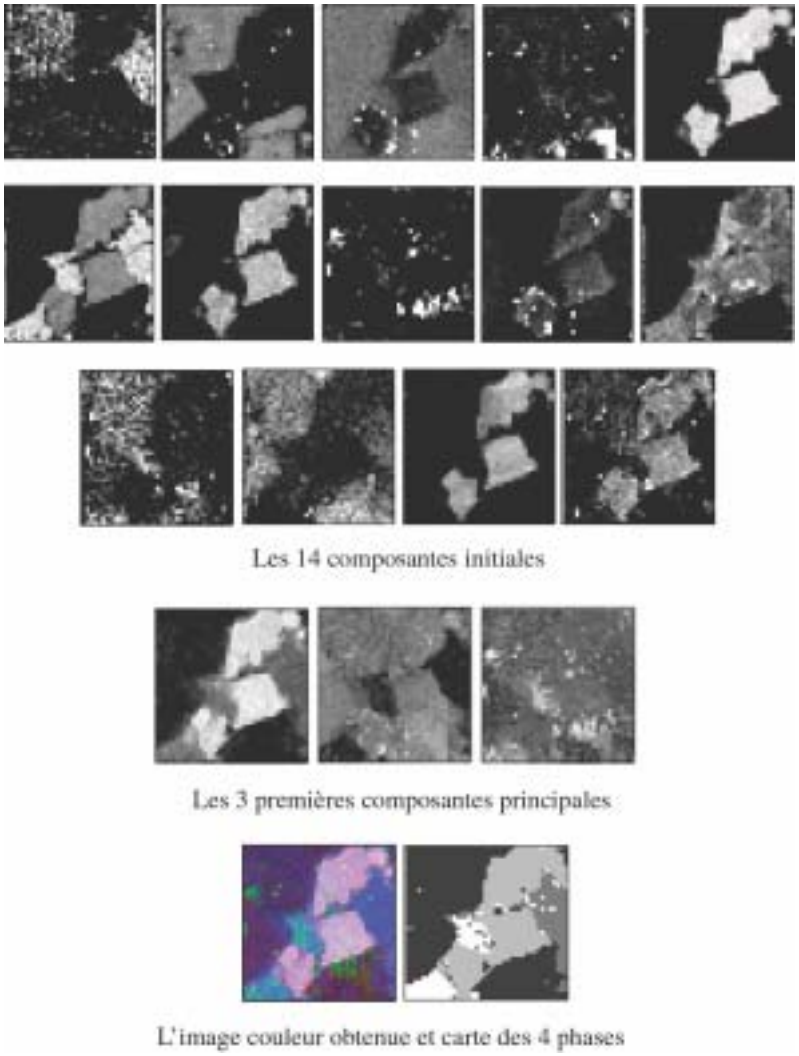

Figure 3. Image à 14 composantes visualisée par une seule image couleur et les quatre phases détectées par une méthode de classification.

L'image couleur que nous proposons permet de donner un résumé des 14 composantes de cette image de fluorescence X. Différentes méthodes de classification ont permis d'établir que quatre phases sont présentes sur cet exemple d'image multicomposante ([Wekemans], [Herbin]) ; ces quatre phases sont nettement observables sur l'image couleur que nous proposons pour visualiser l'image multicomposante. Cet exemple réel et non synthétisé confirme que notre méthode de visualisation par les couleurs permet bien de révéler les différentes classes de données multidimensionnelles. Notons que dans ce cas particulier des images multicomposantes, il est inutile d'utiliser l'étape de construction de l'image telle que nous l'avons décrite dans cet article.

5.3. Visualisation d'une base de données multidimensionnelle

Nous proposons d'appliquer notre méthode de visualisation à la classique base de données IRIS de Fisher [Blake]. Cet ensemble de 150 données en dimension 4 (longueur des sépales, largeur des sépales, longueur des pétales, largeur des pétales) est composé de trois classes : iris setosa, iris versicolor et iris virginica, avec 50 observations ou données par classe. À partir de ces 150 données en dimension 4 sont pouvons calculer une image couleur 16×16 (voir Figure 4) que l'on peut comparer avec l'image des labels des trois classes.
La classe des iris setosa est facilement différentiable des deux autres classes, en revanche ces deux classes sont difficiles à séparer. Ce résultat est bien connu sur cette base de données. Nous le retrouvons avec notre méthode de visualisation bien que celle-ci soit peu adaptée à un échantillon de faible dimension (dimension quatre) et de faible effectif (150 données).

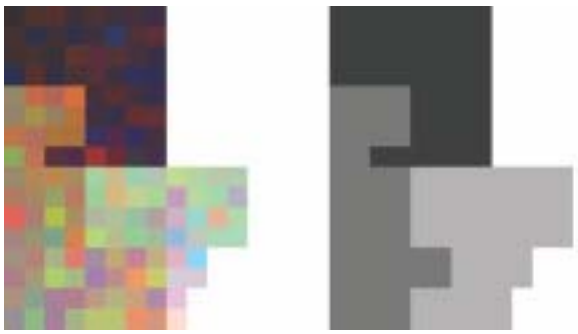

Figure 4. Visualisation des 150 données IRIS de dimension 4 (image couleur et label des trois classes).

6. Discussion et conclusion

La méthode que nous proposons pour visualiser un ensemble de données multidimensionnelles permet une représentation rapide de tout l'ensemble par une simple image couleur. Avec cet outil de visualisation, la couleur est un auxiliaire fondamental qui permet une lecture directe et très intuitive de l'image et donc de l'ensemble des données. L'utilisation d'une image permet ainsi d'obtenir une méthode de visualisation particulièrement bien adaptée aux grands échantillons de données multidimensionnelles.
Notre technique de visualisation a ses limites. Le fait d'avoir réduit la dimension à trois pour calculer les couleurs est un facteur limitatif à l'efficacité de cette approche. En effet, lorsque nous travaillons sur des données de grandes dimensions, il est nécessaire de réduire la dimension sans connaitre la dimension intrinsèque de nos données. Il est sûr que, si cette dimension intrinsèque est supérieure à trois, alors une partie éventuellement importante de l'information sera invisible dans notre représentation. Un autre inconvénient des travaux présentés dans cet article est l'utilisation d'une méthode linéaire de projection pour réduire la dimension. Ce modèle linéaire de réduction de dimensionalité n'est pas nécessairement le mieux adapté aux données à traiter. D'autres approches, par exemple par des cartes auto-organisatrices [Kohonen], pourraient être envisagées. Il faut cependant remarquer que les trois premières composantes d'une ACP (on peut d'ailleurs donner le pourcentage de variance expliquée dans les trois premières composantes) sont le plus souvent largement suffisantes pour que notre méthode de visualisation soit applicable à la plupart des cas.
Notre approche est très objective, l'observateur ne choisit pas les couleurs utilisées et n'apporte aucune connaissance *a priori* sur les données. Les couleurs dépendent uniquement de l'échantillon de données utilisé. Si l'échantillon change, les couleurs changent aussi. La couleur n'est donc interprétable que relativement aux autres couleurs de l'image (*i.e.* aux autres données). L'objectivité est un avantage considérable mais notre méthode est peu robuste. En effet, si l'échantillon de données contient de nombreuses données aberrantes, les couleurs des classes de données peuvent s'avérer difficiles à discriminer, l'image étant alors difficile à lire.
Les limitations citées précédemment ne sont pas spécifiques de cet outil de visualisation. Comme pour toutes les techniques de visualisation, il est nécessaire de confirmer les observations par d'autres moyens d'analyse et de traitement des données. Notre méthode de visualisation des données multidimensionnelles offre un outil nouveau qui, à notre avis, devrait être utilisé principalement dans deux cas : soit pour approche préliminaire à une exploration ou une classification non supervisée des données, soit pour controler ou confirmer le résultat d'une analyse des données obtenue par d'autres méthodes. La simplicité de cet outil de visualisation est un atout dû essentiellement à l'utilisation de la couleur, cela devrait en faire un auxiliaire précieux pour toute exploration des grands effectifs de données multidimensionnelles en particulier pour la fouille de données.

Références

[Aggarwal] C. C. AGGARWAL, A. HINNEBURG, and D. A. KEIM, "On the surprising behavior of distance metrics in high dimensional space". In *Proceedings of the 8th International Conference on Database Theory*. Springer-Verlag, 2001.

[Asimov] D. ASIMOV, "The grand tour: a tool for viewing multidimensional data", *Journal on Scientific and Statistical Computing*, Vol. 6, #1, p.128-143, 1985.

[Buja] A. BUJA, D. COOK, and D. F. SWAYNE, "Interactive high-dimensional data visualization", *Journal of Computational and Graphical Statistics*, Vol. 5, p.78-99, 1996.

[Bellman] R. BELLMAN, "Adaptive control processes : a guide tour". *Princeton University Press*, 1961.

[Bonnet] N. BONNET, M. HERBIN, and P. VAUTROT, "Extension of the scatterplot approach to multiple images", *Ultramicroscopy*, Vol. 60, 1995.

[Blake] C.L. BLAKE and C.J. MERZ, UCI repository of machine learning databases, 1998.

[Debacker] A. DE BACKER, A. NAUD, and P. SCHEUDERS, "Non linear dimensionality reduction techniques for unsupervised feature extraction", *Pattern Recognition Letters*, Vol. 19, p.711-720, 1998.

[Camastra] F. CAMASTRA, "Data dimensionality estimation methods: a survey", *Pattern Recognition*, Vol. 36, p.2945-2954, 2003.

[Comon] P. COMON, "Independant component analysis, a new concept?", *Signal Processing*, Vol. 36, #2, p.287-314, 1994.

[Demartines] P. DEMARTINES and J. HERAULT, "Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets", *IEEE Transactions on Neural Networks*, Vol. 8, p.148-154, 1997.

[Devijver] P.J. DEVIJVER and J. KITTLER, "Pattern Recognition : A statistical Approach", *Prentice-Hall*, Englewood Cliffs, NJ, 1982.

[Donoho] D. L. DONOHO, "High-dimensional data analysis: The curses and blessings of dimensionality", *Aide-Mémoire*, 2000.

[Grinstein] G. GRINSTEIN, M. TRUTSCHL, and U. CVEK, "High-dimensional visualizations". In *Proceedings of the Visual Data Mining workshop, KDD'2001*, San Francisco, California, 2001.

[Healey1] C. G. HEALEY and J. T. ENNS, "Large datasets at a glance: Combining textures and colors in scientific visualization", *IEEE Transactions on Visualization and Computer Graphics*, 5(2):145-167, 1999.

[Healey2] C. G. HEALEY, "Choosing effective colours for data visualization". In *7th IEEE Visualization '96 Conference*, p. 263, 1996.

[Hyvären] A. HYVAREN, J. KARHUNEN, and E. OJA, "Independant Component Analysis", *John Wiley and Sons*, 2001.

[Hughes] D.F. HUGHES, "On the mean accuracy of statistical pattern recognition", *IEEE Transaction on Information Theory*, Vol. 14, #1, p.55-63, 1968.

[Herbin] M. HERBIN, P. VAUTROT, and N. BONNET, "Estimation of the number of clusters and influence zones", *Pattern Recognition Letters*, Vol. 22, p.1557-1562, 2001.

[Keim] D. A. KEIM, "Pixel-oriented visualization techniques for exploring very large databases", *Journal of Computational and Graphical Statistics*, Vol. 5, #1, p.58-77, 1996.

[Kohonen] T. KOHONEN, "Self-Organizing Maps", *Springer*, Berlin, 1995.

[Landgrebe] D. LANDGREBE, "On information extraction principles for hyperspectral data". In *4th International Conference on GeoComputation*, Fredericksburg, Virginia, USA, p.25-28, 1999.

[Lennon] M. LENNON, « Méthodes d'analyse d'images hyperspectrales. Exploitation du capteur aéroporté CASI pour des applications de cartographie agro-environnemantale en Bretagne », *PhD thesis, Université de Renne I*, 2002.

[Lebart] L. LEBART, A. MORINEAU, and M. PIRON.,« Statistique exploratoire multidimensionnelle », *Dunod*, 2002.

[Moon] B. MOON, H. V. JAGADISH, C. FALOUTSOS, and J. H. SALTZ, "Analysis of the clustering properties of the hilbert space-filling curve", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 13, #1, p.124-141, 2001.

[Minnotte] M.C. MINNOTTE and R. W. WEST, "The data image: a tool for exploring high dimensional data sets". In *Proceedings of the ASA Section on Statistical Graphics*, Dallas, Texas, USA, 1998.

[Nason] G. NASON, "Three-dimensional projection pursuit", *Applied Statistics*, Vol. 44, #4, p.411-430, 1995.

[Ohta] Y. OHTA, T. KANADE, and T. SAKAI, "Color information for region segmentation", *Computer Graphics and Image Processing*, Vol. 13, p.222-241, 1980.

[Rao] C.R. RAO, "The use and interpretation of principal component analysis in applied research.", *Sankya serie A*, Vol. 26, 1964.

[Sasov] A. SASOV, "Non-raster isotropic scanning for analytical instruments", *Journal of Microscopy*, Vol. 165, 1992.

[Verveer] P. J. VERVEER and R. P.W. DUIN, "Estimators for the intrinsic dimensionality evaluated and applied", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17, #1, 1995.

[Verleysen] M. VERLEYSEN, "Machine learning of high-dimensional data: Local artificial neural networks and the curse of dimensionality", *Thesis, UCL University Catholique Louvain*, Belgium, 2000.

[Wekemans] B. WEKEMANS, K. JANSSENS, L. VINCZE, A. AERTS, and J. HEERTOGEN, "Automated segmentation of μ-xrf image sets", *X-ray Spectrometry*, Vol. 26, p.333-346, 1997.

Frédéric **Blanchard**

Frédéric Blanchard est titulaire d'un DEA d'Informatique et Recherche Opérationnelle de l'Université de Paris 6 obtenu en 2002 et est actuellement doctorant au CReSTIC-LERI de Reims. Ses centres d'intérets scientifiques sont l'imagerie, la classification et la visualisation.

Michel **Herbin**

Michel Herbin est docteur en biomathématiques et biostatistiques de l'Université de Paris 7 depuis 1989. Après avoir travaillé dans l'industrie dans le domaine de l'imagerie médicale, il est maintenant enseignant-chercheur à l'Université de Reims Champagne-Ardenne depuis 1999. Professeur d'Informatique à l'UFR des Sciences, ses travaux de recherche portent sur la vision par ordinateur, le traitement numérique d'images et l'analyse de données (classification et visualisation).