Compendium de données multidimensionnelles par une image couleur

Frédéric Blanchard*, Michel Herbin** et Francis Rousseaux***
CReSTIC, Université de Reims,
LERI, IUT, rue des Crayères, BP 1035,
51687 Reims Cedex 2, FRANCE
*frederic.blanchard@univ-reims.fr,
**michel.herbin@univ-reims.fr
***francis.rousseaux@univ-reims.fr

1 Introduction

L'extraction de connaissances nécessite des outils de visualisation pratiques et rapides qui sont essentiels pour la fouille des données. Dans ce contexte, une présentation des principales techniques de visualisation est proposée dans [Grinstein 2001] et dans [Card 1999]. Dans cet article, nous nous intéressons plus particulièrement à la visualisation statique et plane de données multidimensionnelles quantitatives. L'utilisation des outils de visualisation se heurte alors à deux difficultés principales : la dimension et l'effectif des échantillons de données. D'une part la dimension de l'espace dans lequel se situent les données peut être importante et, dans certain cas, dépasser largement cent. Ceci conduit à un ensemble de phénomènes qui dissimulent l'information pertinente que l'on recherche. Ces phénomènes sont connus sous le vocable de *malédiction de la dimensionalité* (*curse of dimensionality* [Bellman 1961]). D'autre part, l'effectif de l'échantillon peut être considérable et dépasser le million. Les techniques de visualisation ont alors tendance à masquer l'information pertinente du fait de cet effectif considérable. Dans ce cadre, nous rappelons quelques techniques classiques utilisées pour résoudre ces difficultés puis nous proposons un nouvel outil de visualisation pour résumer un échantillon des données, en fournir un compendium sous forme d'une image couleur. La mise en oeuvre d'une stratégie d'exploration des données ou d'extraction de connaissances n'est évoquée qu'en discussion de cet article.

Les techniques de réduction de dimension permettent de s'affranchir de la première difficulté signalée précédemment. Dans la seconde partie de cet article, nous présentons donc brièvement les principales techniques de réduction de dimension et leur justification (voir à ce propos [Landgrebe 1999]). Pour l'estimation de la dimension intrinsèque d'un échantillon de données, on se réfèrera aux travaux de F. Camastra [Camastra 2003].

Dans la troisième partie, nous utilisons une technique orientée pixel pour visualiser des échantillons de grands effectifs. Cette technique permet de présenter autant de données que de pixels affichables sur un écran de visualisation. On peut noter que cette technique n'est pas spécifique des échantillons de grande taille et les travaux de Keim [Keim 2000] font référence pour une présentation détaillée des différentes approches des méthodes de visualisation orientées pixel.

Dans la plupart des techniques de visualisation, la couleur est un élément important. Dans la quatrième partie de cet article, nous proposons une méthodologie originale pour

affecter une couleur à chaque donnée de l'échantillon. Notre approche est basée sur l'étude statistique de la couleur de Otha, Kanade et Sakai [Ohta 1980]. Ils proposent d'approximer la tranformation de Karhunen-Loève (TKL) d'une image couleur par une tranformation linéaire. A partir des données (R,V,B) des pixels couleur, ils calculent les triplets (C_1,C_2,C_3) qui approximent les trois composantes de la TKL. Nous nous plaçons dans la situation inverse, nous calculons la TKL des données, nous obtenons les composantes (C_1,C_2,C_3) de chaque donnée, la transformation inverse de celle de Ohta et al. nous permet alors de proposer une approximation de la couleur (R,V,B) associée à une donnée. Ce type d'approche présente l'avantage d'être objectif et non supervisée contrairement aux méthodes traditionnelles de détermination de palettes ou d'échelles de couleurs.

La conjonction des techniques de réduction de dimension et de visualisation orientées pixel avec une nouvelle méthode pour déterminer la couleur de chaque donnée constitue une approche originale et nouvelle de la visualisation de données multidimensionnelles par une image couleur. Nous proposons dans la cinquième partie quelques applications pour évaluer cet outil de visualisation.

En discussion, nous proposons une stratégie d'utilisation de ce nouvel outil pour une recherche de points singuliers pour explorer les données. Enfin nous concluons cet article.

2 Réduction de dimensionalité

L'analyse de données multidimensionnelles nécessite une réduction de dimensionalité pour des raisons pratiques liées aux représentations des données [Healey 1999] et théoriques liées à la *malédiction de la dimensionalité*. Dans cet article, les données sont dans un espace de dimension supérieure à trois.

Notre perception humaine de l'espace 3D n'est pas généralisable à des espaces de dimension plus élevée. Prenons par exemple la notion classique de distance euclidienne. Considérons la distance euclidienne dans $\mathbb{R}^n$ où la dimension est égale à n. Soient 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 (ϵ,ϵ,...ϵ) où ϵ est un nombre positif très petit. Comparons les distances AA' et AB. Quand la dimension n est inférieure à trois, on a AA' $\ll$ AB (AA' est plus petit ou égal à $\epsilon\sqrt{3}$ 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. Cet exemple montre que la perception de distance dans les espaces 2D ou 3D ne peut pas être extrapolée aux espaces de dimension supérieure. Ceci illustre un aspect de la *malédiction de la dimensionalité*, de nombreux exemples sont donnés dans les travaux de [Donoho 2000].

Dans cet article nous proposons une approche 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 1964]. Le principe est de projeter les données dans un sous-espace de dimension trois, les axes de projections étant orthogonaux et décorrélés dans le cas particulier de la Transformée de Karhunen-Loève (TKL) que nous utilisons. De nombreux travaux proposent des approches alternatives pour réduire la dimension. Des techniques dites de *projection*

pursuit [Nason 1995] 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) peut aussi permettre de projeter les données dans un sous-espace obtenu en maximisant une fonction de contraste [Comon 1994]. Dans cet article, nous utilisons la TKL pour projeter les données dans un sous-espace de dimension trois mais la méthode de visualisation que nous proposons pourrait être adaptée en utilisant d'autres techniques de réduction de dimension.

3 Une image de l'échantillon

Pour construire une image d'un échantillon de données, nous proposons d'associer chaque donnée à un pixel de l'image. Cette approche de la visualisation orientée-pixel permet de représenter des échantillons de grande taille [Keim 2000]. Un ensemble de N données sera représenté par une image ayant N pixels. Dans ce papier, chaque donnée est représentée par une couleur : un triplet (R,V,B)). Dans le paragraphe 4, nous présenterons comment obtenir une représentation couleur d'une donnée. Dans ce paragraphe, la représentation est supposée fixée. La construction de l'image consiste à déterminer les coordonnées des pixels (i.e. des représentations des données) dans l'espace image. En effet si les pixels sont placés arbitrairement dans l'image, il est alors impossible de percevoir des groupes ou classes de données. Quand les données sont dispersées dans l'image, il devient difficile d'effectuer des rapprochements entre données. Pour que l'image soit un outil de visualisation efficace, lisible au premier coup d'oeil de manière très intuitive, il faut que les similarités ou dissimilarités entre données soient faciles à déterminer. Pour cela, il faut que des données similaires soient spatialement très proches et des données dissimilaires éloignées. Pour cela nous proposons une méthode de construction de l'image en deux étapes : les pixels (i.e. les représentations des données) seront d'abord triés de manière à former une suite de pixels successifs, ensuite cette ligne sera utilisée pour remplir l'image.

3.1 Tri des données

Il nous faut d'abord trier les pixels. Trouver un ordre sur l'ensemble des pixels équivaut à projeter les données sur un espace de dimension un. Les représentations des données sont alors ordonnées ou rangées par l'ordre naturel sur l'espace $\mathbb{R}$ ($\mathbb{R}$ étant l'espace de projection de dimension un). Nous avons déjà réduit la dimension en projetant les données dans un espace 3D (voir réduction de dimensionalité). Les trois composantes obtenues nous donnent trois clefs pour effectuer un tri sur l'ensemble des données de l'échantillon. Cette étape de tri des pixels est effectuée en utilisant les résultats de la réduction de dimension du paragraphe précédent. Nous obtenons alors une ligne de pixels successifs.

3.2 Courbe de Peano-Hilbert

L'étape suivante consiste à remplir l'image avec cette ligne de pixels successifs. La courbe de Peano-Hilbert constitue le moyen le plus classique pour effectuer cette

construction [Moon 2001] (voir sur la Fig.1 la description de la procédure récursive de construction d'une telle courbe). Le principal avantage de cette courbe est de préserver au mieux la connexité des classes de données [Sasov 1992]. En effet, cette courbe tend à minimiser les écarts entre d'une part les distances entre pixels dans l'image (distances dans l'espace image 2D) et d'autre part les distances entre pixels sur la ligne initiale (distances sur la ligne 1D c'est à dire différences de rang). Ces écarts seront plus importants si l'on utilise un parcours de l'image ligne par ligne ou bien colonne par colonne.

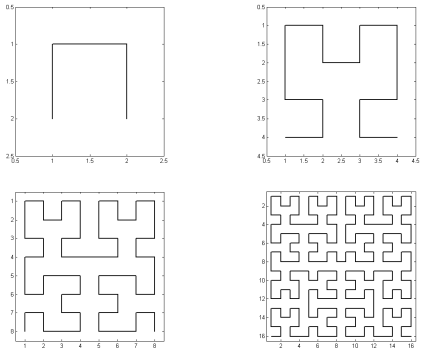


FIG. 1 – *Etapes de la construction d'une courbe de Peano-Hilbert.*

Avec ces deux étapes de tri des données puis le remplissage de l'image par une courbe de Peano-Hilbert, nous évitons de disperser les pixels dans l'image construite. Cette approche tend à préserver la cohérence spatiale des données permettant ainsi une visualisation très intuitive des échantillons de données comme nous le verrons dans les exemples du paragraphe 5.

4 La couleur d'une donnée

En imagerie, la couleur est définie le plus souvent par un triplet (R,V,B) de trois valeurs Rouge, Vert et Bleu, codées sur 8 bits (entre 0 et 255). Après avoir réduit la dimension de l'échantillon par projection des données dans un sous-espace de dimension trois, chaque donnée est représentée par un triplet (X,Y,Z). Il serait très maladroit de considérer que X représente le Rouge, Y le Vert et Z le Bleu. Pour construire la couleur de chaque donnée, nous utilisons d'une part le fait que X, Y et Z sont les trois premières composantes de la TKL calculée sur l'échantillon. Si ce n'est pas le cas (si une méthode de réduction de dimensionalité autre que la TKL a été mise en oeuvre), il sera alors nécessaire d'utiliser préalablemant la TKL pour obtenir ces trois premières composantes principales. D'autre part, nous utilisons une transformation classique de l'analyse d'image couleur : celle de Otha, Kanade et Sakai [Ohta 1980]. Cette transformation consiste à approximer par une application linéaire les trois composantes de la TKL pour une image en couleurs naturelles (il ne s'agit pas d'une image couleur de

synthèse). Nous nous plaçons dans la situation inverse où nous disposons des trois composantes de la TKL, la transformation inverse de celle de Ohta et al. nous permet de proposer une approximation de ce que devraient être les couleurs naturelles de chaque donnée. Cette approche originale que nous proposons sera illustrée par des exemples au paragraphe suivant. La tranformation inverse de celle de Ohta et al. est alors définie par :

$$\begin{cases} R &= (6X + 3Y - 2Z)/6 \\ V &= (3X + 2Z)/3 \\ B &= (6X - 3Y - 2Z)/6 \end{cases}$$

Ohta et al. ont proposé leur transformation pour permettre une segmentation plus aisée d'une image de couleurs naturelles. La transformation inverse doit permettre d'obtenir des couleurs respectant *naturellement* les classes des représentants des données. Cette approche de la couleur dépend de l'échantillon de données. Si l'échantillon change, les couleurs changent. Notre approche de la couleur ne propose qu'un résumé coloré associé à un échantillon, ce constat ne constitue qu'un préliminaire à une exploration ou une interprétation plus guidée des données.

5 Exemples et applications

Nous proposons trois exemples pour illustrer notre méthode de visualisation. Le premier a pour but de montrer l'efficacité d'une image pour visualiser des données. Le second permet de mieux comprendre l'utilité de la couleur pour appréhender rapidement l'organisation d'un échantillon de données. Le troisième propose une application où les classes de données sont connues, il permet de vérifier que notre approche de la visualisation révèle bien les principales structures présentes dans un échantillon de données.

5.1 L'image pour visualiser des données

Dans cet exemple nous disposons d'un échantillon de 65.536 données simulées quantitatives dans un espace de dimension six (non bruitées pour simplifier la lecture). Chaque donnée est un vecteur ayant six composantes. Les Fig.2.a représentent les six composantes de cet échantillon. En appliquant notre technique, nous pouvons affecter une couleur à chaque donnée. La Fig.2.b représente les 65.536 pixels couleurs obtenues. Enfin la Fig.2.c donne le résumé de l'échantillon par une image couleur de 256×256 pixels. Cette représentation nous révèle les 49 classes de données simulées dans cet échantillon.

5.2 La couleur pour visualiser des données

Dans ce deuxième exemple nous disposons à nouveau d'un autre échantillon de 65.536 données dans un espace de dimension six. Ces données sont cette fois spatialement organisées et l'on peut considérer que cet échantillon est déjà une image multi-composante. La Fig.3.a représente les six composantes de cette image. Les 49 classes présentes dans l'échantillon sont immédiatement perceptibles sur l'image couleur de la

Fig.3.c. La couleur est un auxiliaire précieux pour la visualisation et la compréhension de cet échantillon de données.

5.3 Classification visuelle

Nous proposons d'appliquer notre méthode de visualisation à un échantillon extrait de manière aléatoire de la base de données "Forest CoverType"[Blake 1998]. Cet échantillon de 4096 observations ayant 54 attributs est composé de 7 classes. Chaque donnée représente l'observation d'une parcelle de 30×30 mètres. Les variables décrivent l'altitude, la distance au point d'eau le plus proche, les coordonnées géographiques, etc... On sait qu'il existe 7 classes représentant chacune un "type" de forêt. A partir de ces 4096 données en dimension 54 nous pouvons calculer une image couleur 64×64 (voir Fig.4) que l'on peut comparer avec l'image des labels des 7 classes. Si les frontières des classes ne sont pas immédiatement visible sur notre image couleur, on peut cependant y observer les principales structures des données. Notre outil présente donc un grand intéret pour une première exploration des données sans connaisances a priori, et constitue ainsi une première étape dans un processus d'exploration.

6 Discussion et Conclusion

Les exemples précédents illustrent d'une part l'utilité de l'organisation spatiale des données pour une visualisation par une image et d'autre part l'utilité de la couleur pour visualiser un échantillon de données. Dans notre approche, c'est la redondance de ces deux informations spatiales et chromatiques qui permet une perception immédiate et très intuitive des données représentées.

Dans l'état d'avancement de nos travaux, cet outil est actuellement utilisé pour une prévisualisation d'un échantillon ou une confirmation visuelle d'une analyse des données obtenue par une autre méthode. En réduisant la taille de l'image couleur obtenue, on peut proposer des vignettes couleur de chaque échantillon permettant ainsi une identification visuelle rapide de chacun des échantillons. Ce type de signature d'un échantillon reste à étudier plus précisément.

L'outil de visualisation que nous proposons utilise la TKL qui centre les données sur un point moyen de l'espace (point dont les coordonnées sont les moyennes des coordonnées des données). Nous proposons dans des travaux futurs de recalculer une nouvelle image couleur en centrant les données sur un point quelconque de l'espace. Ce point sera un observateur de l'ensemble des données. Un échantillon d'observateurs et les caractéristiques colorimétriques des images obtenues pour chaque observateur nous permettrons de déterminer des points singuliers dans l'espace des données. Nous envisagerons alors des scénarii d'exploration où le ou les observateurs se déplaceront en engendrant ainsi des images couleurs ou des films permettant une exploration plus approfondie les données.

Références

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

[Card 1999] S. K. Card, J. D. MacKinlay, and B. Shneiderman. *Readings in Information Visualization–Using Vision to Think*. Series in Interactive Technologies. Morgan Kaufmann, Morgan Kaufmann Publishers, San Fransisco, USA, first edition, 1999.

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

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

[Camastra 2003] F. Camastra. Data dimensionality estimation methods: a survey. *Pattern Recognition*, 36(12):2945–2954, 2003.

[Keim 2000] D. A. Keim. Designing pixel-oriented visualization techniques: theory and applications. *IEEE Trans. on Visualization and Computer Graphics*, 6(1):59–78, 2000.

[Ohta 1980] Y. Otha, T. Kanade, and T. Sakai. Color information for region segmentation. *Computer Graphics and Image Processing*, 13:222–241, 1980.

[Healey 1999] 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.

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

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

[Nason 1995] G. Nason. Three-dimensional projection pursuit. *Applied Statistics*, 44(4):411–430, 1995.

[Comon 1994] P. Comon. Independant component analysis, a new concept? *Signal Processing*, 36(2):287–314, 1994.

[Moon 2001] 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*, 13(1):124–141, 2001.

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

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

Figures

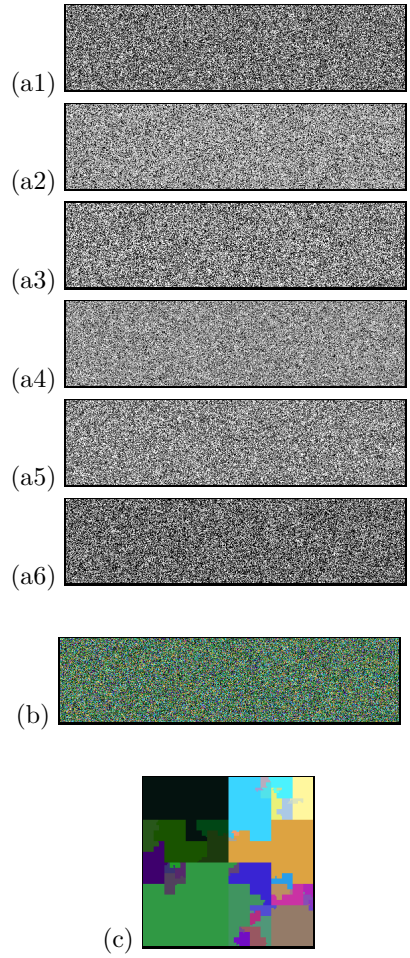
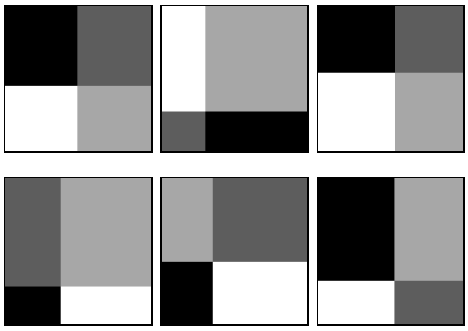
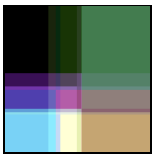


FIG. 2 – *Organisation spatiale d'un échantillon de* 65.536 *données simulées en dimension* 6 *avec 49 classes : (a) Six composantes de l'échantillon, (b) La couleur des données sans organisation spatiale, (c) L'image couleur organisée en* 256×256 *pixels.*



(a) Les six composantes d'un échantillon de 65.536 données spatialement organisées.



(b) L'image couleur présentant les 49 classes de données

FIG. 3 – *Visualisation couleur de 65.536 données de dimension 6 dans une image de* 256×256 *pixels.*

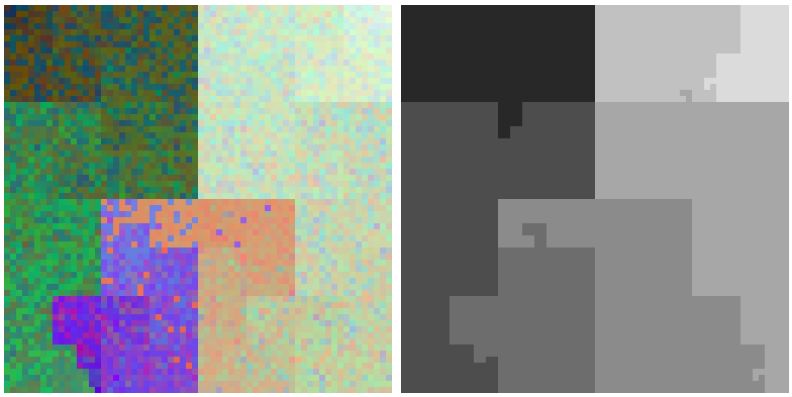


FIG. 4 – *Visualisation des 4096 données de la base "Forest Cover Type" de dimension 54 (image couleur et label des 7 classes).*