Segmentation non-supervisée de champs de données multi-composantes

L. Galluccio[1,2], O. J.J. Michel[2], P. Comon[1]

[1]Laboratoire I3S UNSA-CNRS - UMR 6070
2000, route des Lucioles - Les Algorithmes - Bât. Euclide B - BP.121 06903 Sophia Antipolis - Cedex, France

[2]LUAN, UNSA-CNRS - UMR 6525
Faculté des Sciences, Parc Valrose - Bât. H Fizeau, 06108 Nice Cedex 2, France

laurent.galluccio@i3s.unice.fr, olivier.michel@unice.fr
pcomon@i3s.unice.fr

Résumé – Dans cet article, nous revisitons plusieurs méthodes de segmentation non supervisée de champs de données multi-composantes. La première de ces méthodes repose sur l'algorithme de Lloyd plus connu sous le nom d'algorithme Kmeans ; nous proposons une approche originale d'initialisation de cet algorithme par des données déterminées lors la construction de l'arbre de représentation minimale (Minimum Spanning Tree : MST) [7]. Le même type d'initialisation est appliqué pour une seconde méthode inspirée des travaux de Shi et Malik [11] exploitant le Laplacien d'un graphe totalement connecté. Le problème très important du choix de la métrique pour caractériser la différence entre deux points du champ multidimensionnel est discuté. Nous présentons des résultats obtenus par les deux méthodes de segmentation sur des images multispectrales et sur des spectres de réflectance d'astéroïdes.

Abstract – In this paper, we revisit some methods of unsupervised segmentation on multi-component datasets. The first method lies on the Llyod's algorithm, also known as Kmeans algorithm; we suggest an original approach to initialize this algorithm, by using data obtained during the building of the minimum spanning tree (MST) [7]. The same kind of initialization has been applied for a second method inspired by works of Shi and Malik [11] based upon the Laplacian of a fully connected graph. The very important problem of the choice of the metric for characterizing similarities between two points of multi-dimensionnal datasets is also discussed. We present results obtained with both segmentation methods on multispectral images and on asteroid reflectance spectra.

1 Introduction

Dans ce papier, on considère un ensemble de N données dans $\mathbb{R}^d$ dont on cherche une partition en K classes, sans apporter d'*a priori* (approche complètement non supervisée). Chaque donnée (vecteur de $\mathbb{R}^d$) peut être considérée comme le sommet d'un graphe. Nous nous limitons à ne considérer que le graphe acyclique totalement connecté de longueur minimale : MST (Minimum Spanning Tree). Une méthode simpliste de segmentation de données consiste à réaliser des coupures dans un graphe de connexion. Créer K groupes (« clusters ») pourrait donc consister à supprimer les $K-1$ liens les plus longs du graphe. Cette méthode, basée sur le « single linkage clustering », est cependant connue pour être instable, principalement en présence d'outliers et/ou de « clusters proches ». Une alternative consistant à prendre en compte le voisinage des points pour améliorer la segmentation des données a été proposée par Griskschat et al. [5]. La distance qui y est proposée est construite sur le temps nécessaire pour que les « fronts de croissance »associés à deux processus de construction de MST par l'algorithme de Prim se rencontrent. Ainsi, la distance obtenue dépend de l'environnement proche de chacun des points et de la présence ou de l'absence de groupes denses dans toutes les directions. Cependant, le coût algorithmique de ces approches reste très élevé. Une alternative consiste à exploiter plus directement et à moindre coût l'algorithme de Prim pour la construction du MST afin d'initialiser des algorithmes de segmentation plus classiques. Les méthodes citées seront appliquées à deux types de données multi-composantes : des images satellitaires et des spectres de réflectances d'astéroïdes (Section 3).

2 Méthodes de segmentation

2.1 Utilisation du MST

L'utilisation d'un MST dans une méthode de segmentation de données a déjà été proposée dans un contexte astrophysique [7]. Son principal avantage est de passer d'un problème de segmentation multidimensionnelle de données à un problème de partitionnement de graphe.
Parmi les algorithmes permettant de construire un MST de manière exacte, le plus efficace en terme de temps de calcul est celui de Prim [10] ($O(N \log N)$).
Cet algorithme connecte au graphe partiellement connecté à l'itération i le sommet non encore connecté le plus proche du graphe (au sens d'une métrique arbitraire). Avec N points, l'algorithme de Prim construit le MST en $N-1$ itérations. Le graphe déterminé est acyclique, unique (i.e. indépendant du point de départ de construction du graphe)

et de longueur minimale : i.e. soit L_g la longueur totale des segments e parcourant le graphe, avec $L_g = \sum_i e_i$, le MST est le graphe qui parmi tous les graphes connectés est celui dont la longueur L_g est minimale.
La trajectoire de Prim est la fonction f qui exprime la longueur d'un nouveau segment construit à l'itération i, $e_i = f(i)$. f permet de transformer la fonction de densité de probabilité (fdp) des points du champ de données d-dimensionnel, en une fonction à une dimension. Cette dernière exhibe des « vallées »dans les voisinages de forte densité : un ensemble de points très proches est connecté par un ensemble d'itérations consécutives associées à la création de segments courts. La longueur du segment mesure la similarité entre deux points, au sens d'une mesure arbitraire.
La détection de « vallées »dans la courbe représentant la distance entre les points successivement sélectionnés et le graphe en cours de construction, permet d'identifier le nombre de modes principaux de la fdp, chaque mode étant associé à un des groupes de points recherchés. Les informations déduites de la construction du graphe permettent de définir le nombre de groupes présents dans l'ensemble de données et leurs positions (barycentres des groupes). Ces informations servent d'initialisation à l'algorithme Kmeans, par exemple. Cette méthode classique de segmentation est connue pour dépendre fortement de ces paramètres d'initialisation (nombre de groupes, position des centres de ces groupes). Notre approche originale permet à Kmeans de converger plus rapidement vers le minimum global.
On peut visualiser sur la Fig.1 la procédure décrite ci-dessus.

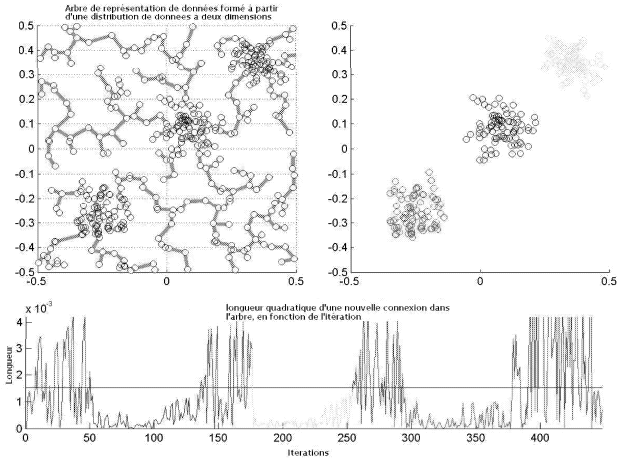

FIG. 1 – *Exemple de visualisation de la construction d'un MST, de la longueur séparant chaque sommet successif du graphe et des trois clusters formés.*

Le choix du seuil appliqué à $f(i)$ qui va déterminer le nombre de modes principaux de la fdp est un paramètre important. Dans [7], nous avons proposé d'identifier chaque groupe en associant une même étiquette à tous les sommets connectés consécutivement sans que $f(i)$ ne repasse au dessus du seuil. Cette approche nécessiterait pour être complète de connaître la fonction de densité de probabilité des $|e_i|$ pour être optimale. Une solution est proposée et discutée dans [7]. Dans le cas présent, le nombre de modes et les barycentres estimés au sein de ces modes nous sont nécessaires pour initialiser un algorithme classique. La position exacte du seuil est moins sensible. On fixe donc arbitrairement le seuil à e.g. $\eta = std\{e_i\}$.
Une autre méthode nécessitant un temps de calcul plus important, mais plus rigoureuse, peut être proposée. On détermine la valeur du seuil optimal *a posteriori* en analysant les indices de validité de segmentation (Davies-Bouldin, Dunn [6]) pour chaque ensemble de valeurs (nombre de clusters, barycentres).

Un autre paramètre fondamental à prendre en compte dans la construction du MST est la métrique choisie pour pondérer les différentes connexions du graphe. La norme euclidienne est la métrique la plus utilisée pour calculer la distance entre deux sommets d'un graphe. Lorsque les données étudiées « vivent »dans $(\mathbb{R}^+)^d$, il s'avère intéressant (dans nos applications) d'associer à chaque vecteur de $(\mathbb{R}^+)^d$ une quantité normalisée ($\sum_i x_i = 1$, avec $X = [x_1, ..., x_d]^T$) que l'on interprète comme une densité de probabilité. On apporte une notion d'information dans notre construction d'un MST en définissant la « similarité »entre deux sommets comme la divergence de deux fonctions de densité de probabilité.
A chaque point X, on associe une loi de probabilité P.

$$P(\lambda_l) = p_l = x_l / \sum_{l=1}^{d} x_l$$

Pour calculer la similarité entre deux lois de probabilité P et Q associées aux sommets X et Y, on utilise les divergences de type Csiszàr [9] définies sous la forme suivante :

$$h\left[\ E_q\left\{\ g(\tfrac{p_l}{q_l})\ \right\}\ \right]$$

où E_q représente l'espérance mathématique liée à la loi de probabilité q.
Si h est l'Identité et $g(u) = u\log(u)$, on trouve la divergence de Kullback-Leibler.

$$D_{KL}(X||Y) = \sum_{l=1}^{d} p_l \log \frac{p_l}{q_l} \qquad (1)$$

La divergence de Rényi est définie comme suit.
Si $h(u) = \frac{1}{\alpha - 1}\log(u)$ et $g(u) = u^\alpha$:

$$D_R(X||Y) = \frac{1}{\alpha - 1}\log \sum_{l=1}^{d} p_l^\alpha q_l^{1-\alpha} \text{ si } 0 < \alpha < 1 \qquad (2)$$

Il est facilement démontrable que pour $\alpha = 1$ les divergences de Rényi et de Kullback-Leibler sont équivalentes.
L'arbre construit est non-dirigé ; par conséquent, les divergences définies ci-dessus sont symétrisées :
$D_{sym} = D(X||Y) + D(Y||X)$
Le problème du coût de calcul n'est pas le moins important : la construction d'un MST par l'algorithme de Prim requiert d'ordonner les longueurs de l'ensemble des segments (il y en a $N-1$ pour un MST qui connecte ensembles N points). Cette opération peut être réalisée en $O(N\log(N))$ opérations logiques (e.g. par « bubble

sort »). L'opération la plus coûteuse est donc le calcul des distances pour chaque paire de points du champ de données ($O(N^2)$ opérations). Une méthode de présélection par arbre de classification hiérarchique des données est proposée, qui permet de réduire le nombre de calculs de distances en $O(Nk\log(N))$, $k << N$. Dans la suite de la communication, on appelle cette méthode Nearest Neighbor MST [1].

La mesure construite à l'aide des $f-$divergences n'est pas une distance. Bien que ce point n'ait pas été étudié en détails (à notre connaissance), il semble que cela ne soit pas pénalisant dans le cas que nous envisageons ici. Cependant, d'autres mesures sont actuellement étudiées [5] : Dual rooted hitting time, Direct Prim. Ces deux mesures définissant des distances (non négatives, symétriques, respectant l'inégalité triangulaire) sont en cours d'étude. Pour chacune de ces mesures de proximité entre deux points (sommets) induites par le graphe envisagées, le choix de la métrique utilisée pour construire le graphe reste libre (norme euclidienne, divergences, etc ...).

2.2 Segmentation spectrale

Afin de pallier les difficultés liées à la segmentation simple par « single linkage clustering », mais sans contraindre les résultats à être convexes (ce qui est imposé par l'algorithme Kmeans), nous proposons d'utiliser une méthode simple de segmentations multiples, $NCUT$, introduite par Shi et Malik [11]. Dans ce contexte, le MST et l'algorithme de Prim ne sont plus utilisés que pour estimer le nombre de modes principaux de la fdp de la distribution de points qui constituent le champ de données. Toutes les distances entre paires de points sont considérées : le graphe de connexion contient donc $(N(N-1)/2) - N$ segments. Le problème de segmentation, i.e. minimisation de la coupure normalisée d'un graphe ($NCUT$) ainsi posé est NP-Complet [11] ; une solution relaxée a été proposée par Jordan et Bach [1]. Cette méthode conduit à exploiter le spectre de valeurs propres du Laplacien normalisé du graphe qui connecte tous les points entre eux.
Le graphe totalement connecté est pondéré par la matrice d'affinités W. Celle-ci peut prendre diverses formes, mais la plus commune reste la suivante

$$W_{ij} = e^{-\frac{M_{ij}^2}{2\sigma^2}}$$

M correspond à une matrice de distances (au sens où elle respecte les axiomes de définitions mathématiques d'une distance). Par conséquent, nous pouvons utiliser les métriques induites par le MST définies précédemment. Le choix du paramètre d'échelle σ^2 suscite quelques polémiques dans la littérature. En général, il est laissé libre au choix de l'utilisateur, ou validé *a posteriori* par le calcul d'indice de qualité de la segmentation obtenue.
Soit L le Laplacien d'un graphe totalement connecté (chacun des N points est connecté aux $N-1$ autres par un segment de poids w_{ij}), caractérisé par la matrice $W = [w]_{ij}$: $L = D - W$ où $D = \sum_j w_{ij}$.
Le Laplacien normalisé du graphe $\widetilde{L}$ est défini comme : $\widetilde{L} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}WD^{-1/2}$. On ne se préoccupe que de la seconde partie de l'équation, la matrice Identité n'interfère dans les résultats qu'au niveau des valeurs propres ($1 - \lambda_i$ à la place de λ_i), et donc n'a pas d'influence sur les vecteurs propres[2]. $\widetilde{L}$ est symétrique et semi-définie positive.
Belkin et al. [2] ont montré la relation entre le problème de recherche de valeurs propres du Laplacien du graphe et la réduction de dimensions, les deux conduisant aux mêmes équations. Le problème de la segmentation spectrale revient donc à projeter le graphe sur une variété de dimension réduite de sorte que les points adjacents dans le graphe restent aussi proches les uns des autres que possible. Une méthode classique de segmentation de données de type Kmeans est ensuite appliquée à ces données projetées. Cette famille d'approches présente un grand intérêt pour la segmentation de champs de données très redondants ou pour lesquels il peut être supposé que l'ensemble des données évolue sur une variété de dimension très inférieure à la dimension de l'espace dans lequel sont observées les données. Cela est souvent le cas en imagerie hyperspectrale par exemple.

3 Résultats

L'image satellitaire dont nous disposons est de taille 512×512. Pour déterminer les paramètres d'initialisation à l'algorithme Kmeans, on construit un MST sur un échantillon des données (e.g. facteur de réduction de 4). On utilise également la méthode de présélection par classification hiérarchique (Nearest Neighbor MST). Le premier exemple d'application consiste à segmenter en aveugle un champ de données contenant 3 images de Paris acquises à 3 λ différentes ($P(\lambda_l)$ représente ici la probabilité que les photons reçus en X aient une longueur d'onde λ dans la bande associée au canal l). L'algorithme proposé (Nearest Neighbor MST : métrique de type divergence de Kullback-Leibler symétrisée, segmentation des $f(i)$, Kmeans) identifie 5 groupes : 3 sont facilement interprétables, les 2 autres nécessitent d'être confrontés à une expertise.
On détecte la Seine, ainsi que les différents plans d'eau (lacs du bois de Boulogne et de la Butte Chaumont) dans le cluster 5 mais également dans le cluster 4 toutes les parties boisées se trouvant aux alentours de Paris (bois de Boulogne, Vincennes, Pré St-Gervais, Buttes Chaumont). Sur le cluster 3, on détecte toutes les artères routières qui entourent Paris entre autres.
Nous avons également à notre disposition plusieurs champs de données correspondant à des spectres de réflectances d'astéroïdes. Les astéroïdes sont des corps n'émettant pas de lumière ; on mesure par conséquent la façon dont les particules minéralogiques présentes sur la surface des astéroïdes absorbent ou diffusent la lumière (suivant la longueur d'onde). Nous réalisons donc une classification minéralogique de surface. Les mesures viennent de l'étude appellée SMASSII : Small Main Belt Asteroid Spectroscopic Survey II, où le rapport entre la lumière incidente

[1] code matlab disponible : contacter gallucci@i3s.unice.fr.

[2] Pour plus de renseignements sur la théorie des graphes et sur les différentes formes de Laplacien de graphe, le lecteur se référera à [4].

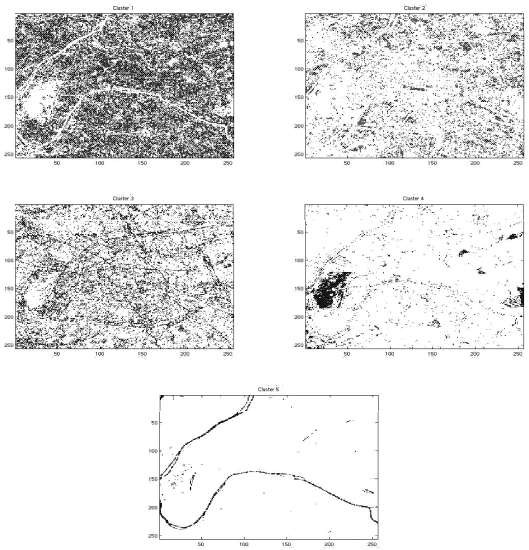

FIG. 2 – *Test sur une image satellitaire représentant Paris et ses environs*

et la lumière réfléchie du soleil (i.e. la réflectance) a été mesuré sur 1329 astéroïdes. Ces mesures ont été échantillonnées sur 9 valeurs de longueurs d'onde. Les résultats sont comparés à une taxonomie réalisée par Bus & Binzel [3]. L'indice de validité de segmentation utilisé est un indice de référence dans le domaine [12]. Il est défini à partir de ces deux variables.
$\mathcal{N}_{Tax}$ = nombre d'astéroïdes présents dans la classe taxonomique de référence Tax,
$\mathcal{N}_{inter}$ = nombre d'astéroïdes se trouvant dans le groupe C formé et dans Tax.

L' indice de validité de segmentation appliqué à des données astrométriques est défini comme :

$$\text{Score} = \frac{\mathcal{N}_{inter}}{\mathcal{N}_{Tax}}$$

Score caractérise le pourcentage d'astéroïdes d'une classe taxonomique ayant été détectés et bien classés à l'issue de la classification. Ce résultat permet de définir l'efficacité de la méthode utilisée.

Méthodes de segmentation	Score
Nearest Neighbor MST (Euclidienne) + Kmeans	815/1329 = 61,32%
Nearest Neighbor MST (Kullback-Leibler) + Kmeans	976/1329 = 73,44%
Segmentation Spectrale [11]	878/1329 = 66,06%
Segmentation Spectrale (Direct Prim)	897/1329 = 67,49%
Segmentation Spectrale (Dual Rooted Hitting Time)	913/1329 = 68,69 %
K plus proche voisin [12]	777/1329 = 58,46 %
Multidimensional Scaling	898/1329 = 67,56 %
Expectation Maximization	826/1329 = 62,15%

TAB. 1 – *Synthèse du SMASSII avec les différentes méthodes de segmentation définies précédemment.*

En analysant les résultats des méthodes de segmentation développées dans cette étude, on constate que le nombre d'astéroïdes identifiés comme appartenant à leur classe de Bus & Binzel est plus important que dans la méthode utilisée dans [12], ou avec des algorithmes classiques tels que Expectation Maximization ou Multi-Dimensional Scaling.

4 Conclusion

Le domaine de la segmentation de données est un domaine large où de nombreuses méthodes (suivant différentes approches) ont été développées ces dernières années. Nous nous plaçons dans le cadre de la classification de données, dans l'imagerie (groupement de pixels) et dans les mesures astrométriques (spectres de réflectance d'astéroïdes). Notre approche est d'utiliser les étapes de la construction d'un arbre de représentation minimale des données afin de déterminer des paramètres d'initialisation (nombre de sur-densités de points présents, centres de ces groupes) à des algorithmes standards de classification tels que Kmeans et Segmentation Spectrale (Spectral Clustering). L'utilisation du MST permet également de définir des distances prenant en compte le voisinage local des points, mais également une notion d'information si on utilise la divergence de Kullback-Leibler pour caractériser la similarité entre deux sommets du graphe. Les résultats en termes de segmentation sont concluants. Concernant la segmentation spectrale, un lien établi avec la réduction de dimension dans [2] peut être très intéressant.

Références

[1] F. R. Bach, M. I. Jordan. *Learning Spectral Clustering.* EECS Department, University of California, Berkeley. Technical Report No. UCB/CSD-03-1249. June 2003.

[2] M. Belkin, P. Niyogi. *Laplacian eigenmaps for dimensionality reduction and data representation.* Technical Report TR 2002-01, Univ. Chicago, Dept. Comp. Sci. and Statistics, January 2002.

[3] S. J. Bus, R. P. Binzel, *Phase II of the Small Main-Belt Asteroid Spectroscopic Survey : The Observations* Icarus - Volume 158, Issue 1, July 2002, Pages 146-177.

[4] F.R.K. Chung, *Spectral Graph Theory*, Am. Math. Soc., 1997.

[5] S. Griskschat, J. A. Costa, A. O. Hero, O. J. J. Michel. *Dual rooted-diffusions for clustering and classification on manifolds.* 2006 IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, Toulouse France.

[6] M. Halkidi, Y. Batistakis, M. Vazirgiannis, *On clustering validation techniques*, Journal of Intelligent Information Systems, 17 :2/3, 107-145, 2001.

[7] O. J.J. Michel, P. Bendjoya, P. RojoGuer. *Unsupervised clustering with MST : Application to asteroid data.* PSIP 2005, Toulouse, paper 01_03, 2005.

[8] A. Y. Ng, M. I. Jordan, Y. Weiss. *On Spectral Clustering : Analysis and an algorithm.* Advances on Neural Information Processing Systems 14, MIT Pres, 2001.

[9] F. Österreicher, *Csiszár's f-divergences-Basics properties*, Institut of mathematics, Universitity of Salzburg, Austria, 2002.

[10] R. Prim. *Shortest connection networks and some generalizations*, Bell Syst. Tech. J.36, 1389-1401, 1957.

[11] J. Shi, J. Malik. *Normalized cuts and image segmentation.* IEEE Transactions on pattern analysis and machine intelligence, Vol. 22, No. 8, August 2000.

[12] J. Warell, C.-I. Lagerkvist, *Asteroid taxonomic classification in the Gaia photometric system*, Astronomy and Astrophysics, Volume 467, Issue 2, May IV 2007, pp.749-752.