Détection de familles d'astéroïdes: un problème de segmentation non supervisée

Olivier J.J. MICHEL, Philippe BENDJOYA, Pablo ROJO GUERRA

LUAN, UMR 6525-CNRS
Université de Nice-Sophia Antipolis, Parc Valrose, 06108 Nice cedex 2 , France
olivier.michel@univ-nice.fr, philippe.bendjoya@univ-nice.fr

Résumé – Nous présentons une méthode de segmentation de données qui utilise les étapes de construction d'un graphe de représentation minimal par l'algorithme de Prim. Les relations entre l'entropie de la distribution des sommets du graphe et la longueur de ce dernier permettent de proposer un critère entropique de détermination du seuil nécessaire à la détection des familles d'objets. Nous présentons une application à la détermination de famille d'astéroïdes à partir de données physiques.

Abstract – An unsupervised clustering method is presented. It makes use of intermediate steps of the Prim algorithm for the minimal spanning tree construction. The links exhibited between the entropy of the vertices of the graph and the entropy of their distribution is reminded. A entropic criterion for setting a new detection threshold is presented.

1 Introduction

1.1 Contexte astrophysique

Une famille d'astéroïdes est un groupe d'objets issus de la fragmentation d'un astéroïde "père" lors d'une collision hyper énergétique avec un autre astéroïde. L'identification de ces familles et leur étude constituent un enjeu important pour la compréhension de l'évolution collisionnelle des astéroïdes dans la ceinture principale (typiquement entre Mars et Jupiter). L'ensemble des objets d'une même famille forme une sorte de puzzle 3D dont la compréhension contient des informations sur la structure interne de l'astéroïde père, la physique de la fragmentation, la formation de nuages proto-planétaires...

Les familles ainsi définies ne sont cependant pas toutes identifiables. Un choc trop violent peut pulvériser l'astéroïde père en une myriade d'objet très petits non observables dont les caractéristiques dynamiques (vitesse d'éjection, paramètres orbitaux) sont trop largement dispersées. Dans le cas de familles a priori facilement observables, les éléments orbitaux des fragments issus du choc sont perturbés au cours du temps par des perturbations du champ gravitationnel perçu (présence des planètes, essentiellement Jupiter); ces derniers amplifient les différences initiales au point de ne pas permettre de reconstruire l'histoire commune de ces fragments. Ces familles ne peuvent donc être détectées que dans un espace de paramètres constituant des 'pseudo-invariants' du mouvement. Ces "éléments propres", notés (a, e, i) [1] peuvent être calculés pour chaque objet observé. Le problème de détection des familles d'astéroïdes se résume alors à un problème de segmentation de zones de l'espace 3D (a, e, i); un aspect spécifique à ce problème tient aussi à la nature non euclidienne de la distance proposée (à partir d'arguments physiques) entre objets, et qui sous-tend la mesure par rapport à laquelle il faut chercher des sur-densités. Cette métrique, développée depuis le début des années 1990 [1, 15] est aujourd'hui très largement acceptée.

1. a est le demi grand axe de l'orbite elliptique, e son excentricité, et i l'angle d'inclinaison par rapport au plan de l'écliptique.

1.2 Le Problème de segmentation

La détermination de familles "naturelles" au sens de la proximité ou du partage de caractéristiques communes implique les taches consécutives de détection d'existence d'un groupe pertinent, puis de décision (quels en sont les membres?). Ces deux tâches sont largement compliquées par la présence d'objets "parasites" dont les caractéristiques peuvent être très proches de celles des objets qui nous intéressent. De plus, aucune information n'est ici a priori disponible sur le nombre de familles à détecter, pas davantage que sur leur taille, ou la séparation (dans l'espace (a, e, i)) attendue entre les familles. Les techniques de segmentation usuelles (analyse en composantes principales, factorielle des correspondances, méthodes "K-Means", algorithme de Forgy-Jancey...) [9] ne peuvent apporter de réponse satisfaisante dans ce contexte. De plus, ces approches concentrées sur le rassemblement de points autour d'un centre ne permettent pas d'aborder les situations pour lesquelles les familles recherchées peuvent avoir des formes compliquées, non forcément convexes.

Divers travaux se sont attachés à apporter des solutions à ce problème, s'appuyant sur des algorithmes de classification hiérarchique [16], ou de détection de sur-densité mettant en jeux plusieurs échelles de résolution [2]. Ces approches sont coûteuses en temps de calcul et nécessitent la définition de nombreux paramètres de réglages. Nous proposons une alternative, exploitant les propriétés des graphes de représentation minimaux (MST pour *Minimal Spanning Tree*), l'existence d'algorithmes de calculs efficaces (en $O(n \log n)$ et leurs propriétés asymptotiques, qui en font des estimateurs efficaces de l'entropie de la distributions des points reliés par le MST [7, 8]. Cette dernière propriété est conditionnée par l'utilisation d'une me-

sure de distance asymptotiquement euclidienne (dans la limite des petites distances); une expression alternative de distance est donc proposée à partir d'arguments relevant d'analyse dimensionnelle; ces points sont l'objet de la section suivante.

2 MST et entropie de Rényi

2.1 Entropie de Rényi

Dans cette section, nous rappelons quelques résultats établis dans [7, 11, 8, 12], et utilisés dans la suite. L'entropie de Rényi d'ordre α de la loi de probabilité P complète est définie par :

$$H_\alpha(P) = \frac{1}{1-\alpha} \log_2 \left(\sum_{i=1}^{n} p_i^\alpha \right) \quad (1)$$

La propriété de convexité de H_α n'est obtenue que pour $0 < \alpha < 1$. Seul ce dernier cas sera considéré dans la suite. Cette définition se généralise sans difficulté au cas de fonctions de densité de probabilité continues λ:

$$H_\alpha(\lambda) = \frac{1}{1-\alpha} \log_2 \left(\int \lambda^\alpha d\mu \right) \quad (2)$$

où μ est la mesure.

2.2 MST, estimation de H_α

Soit $\mathcal{X} = \{X_1, \ldots, X_n\}$ une réalisation de n vecteurs aléatoires indépendants et identiquement distribués où $X_i \in A_0 \subset \mathbb{R}^d$ suit une distribution notée P, de densité de Lebesgue λ. Un arbre de représentation est un graphe $\mathcal{T}$ non dirigé, défini par un ensemble de sommets $\mathcal{X}$ et un ensemble de liens (X_i, X_j) de mesure $d_{i,j}$, connectant les sommets entre eux. La longueur totale d'ordre γ du graphe est définie par

$$L_{n,\gamma} = \sum_{d_{i,j} \in T} |d_{i,j}|^\gamma$$

Le graphe acyclique minimal de représentation (MST) est parmi tous les graphes totalement connectés, le graphe $\mathcal{T}^\star$ dont la longueur $L_{n,\gamma}$ est minimale. $\mathcal{T}^\star$ peut être calculé de façon exacte à l'aide d'algorithmes dont le coût varie en $n \log n$.
Dans [7, 8] nous avons établi que la quantité

$$\hat{H}_\alpha(\lambda) = \frac{1}{1-\alpha} \log_2(n^{-\alpha} L_{n,\gamma}^\alpha) + \beta(\alpha, d) \quad (3)$$

où $\gamma = (1-\alpha)d$, $0 < \alpha < 1$ (donc $0 < \gamma < d$), est un estimateur consistant de l'entropie de Rényi de la densité de Lebesgue de la distribution des sommets de $\mathcal{T}^\star$. β est une constante, dépendant de d et γ, mais non de la densité λ.
Cette propriété ne tient cependant que si (réf. dans [7])

$$\lim_{|e| \to 0} |d_{i,j}| \simeq ||X_i - X_j||$$

où $||e||$ est la norme Euclidienne.
Soit $\Pi_{A_0} = \{A_{01}, \ldots, A_{0k}\}$ une partition de A_0. Soient $\pi = (p_{A_0}, \ldots, p_{A_k})$ la distribution de probabilité discrète associée aux évènements $x \in A_{0j}$, $j \in \{1, \ldots, k\}$. Nous avons établi dans [11]

$$\hat{H}_\alpha(\lambda \times \pi) = \frac{1}{1-\alpha} \log_2 \left(\frac{1}{n^\alpha} \sum_{j=1}^{k} L_{n_{A_{0j}},\gamma} \right) + \beta(\gamma, d) \quad (4)$$

La propriété de super-additivité des MST [14] assure que $\hat{H}_\alpha(\lambda \times \pi) \leq \hat{H}(\lambda | A_0)$

3 Méthode de détection, seuillage

3.1 Description

L'utilisation des MST offre certains avantages. Cela permet de commuer un problème de segmentation multidimensionnelle en un problème de partition de graphe; une telle approche est connue et ses principales limitations ont déjà fait l'objet d'études (voir [5, 10]). Cependant, si le coût de calcul (en flops) reste faible, la complexité de l'analyse combinatoire à mettre en place s'avère rédhibitoire dés lors que le problème est reformulé pour permettre une estimation robuste des familles à l'aide des k-MST (sous graphes minimaux de k points parmi n); le problème posé est alors NP-complet. Les temps de calcul mis en jeux par les algorithmes de construction approchée de k-MST [7] ne permettent que difficilement d'apporter une solution sur des données expérimentales sur les astéroïdes, pouvant compter jusqu'à 5.10^5 objets (dans un futur proche).

Une solution suggérée récemment par Olman [13] dans un contexte bio informatique, exploite les étapes intermédiaires de calculs de MST dans l'algorithme d'agrégation récursive de Prim[2].

Soit $L(p)$ la fonction qui exprime la longueur du segment qui connecte un nouveau point au graphe lors de l'itération p de l'algorithme de Prim. $L(p)$ présente des "vallées" lorsqu'une sur-densité de points est détectée : de fait, cet algorithme va connecter préférentiellement les points proches les uns des autres consécutivement dés lors qu'un premier point de la "famille" sera connecté au graphe (voir figure 1). La difficulté est la détermination du seuil à appliquer à $L(p)$ pour déterminer les frontières de la famille ainsi détectée (section suivante). Le comportement de cet algorithme est illustré sur la figure 1, sur des données de simulation calculées à partir d'un modèle physique de collisions.

Les résultats de ce détecteur sont développés dans [12] et les courbes opérationnelles de réception (CORes) y sont présentées. Cette approche ne fait aucune hypothèse sur la convexité des familles à détecter.

3.2 Estimation du seuil

La section précédente met en évidence la nécessité de définir un seuil sur la courbe $L(p)$; ce seuil permet de détecter les limites des vallées qui sont la signature de l'existence de surdensités dans l'espace (a, e, i), que l'on associe à l'existence de famille d'astéroïdes. L'objet de cette section est de proposer une méthode de détermination du seuil, en exploitant les relations entre la longueur d'un MST et l'entropie de la distribution des points qui en forment les sommets. Cette relation n'existe cependant que si la métrique utilisée est asymptotiquement euclidienne, dans la limite des petites distances. La métrique utilisée dans [1, 16], homogène à une vitesse (ms^{-1}), ne possède pas cette propriété. Nous avons proposé dans [12] une distance

2. Soit $\mathcal{T}_p$ le graphe obtenu après p itérations. L' algortihme consiste à construire le MST complet par adjonction du plus proche voisin de $\mathcal{T}_p$, non encore connecté au graphe. Le graphe résultant est alors $\mathcal{T}_{p+1}$. Le processus d'agrégation est itéré jusqu'à ce que tous les points soient connectés, soit après $n-1$ itérations. L'algorithme peut être initialisé aléatoirement en n'importe quel point. Cet algorithme peut être implanté avec une complexité en $O(n \log n)$ et conduit au graphe acyclique minimal unique [10]

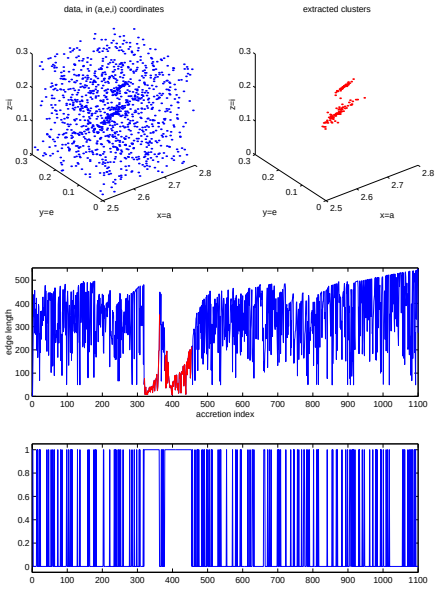

FIG. 1: *Exemple : 2 familles (totalisant 250 et 150 objets) un environnement de 1250 objets; Chaque objet est caractérisé par (a, e, i). Au centre : $L(p)$ en fonction de l'itération p dans la construction du MST. Il apparaît nettement deux vallées. Un seuil η appliqué à $L(p)$ ($S(L(p)) = 1$ si $L(p) < \eta$, 0 sinon) conduit au dernier graphique. Les objets connectés entre eux lors des iitérations successives et telles que $S(L(p)) = cste = 1$ sont représentes sur la figure supérieure droite.*

euclidienne un espace transformé (x, y, z) :

$$x = k'_2.\frac{e}{\sqrt{a}} \quad y = k'_3.\frac{\sin i}{\sqrt{a}} \quad z = k'_1 a$$

où (k'_1, k'_2, k'_3) sont des constantes dimensionnées. La distance euclidienne reste alors homogène à une vitesse. Cette transformation présente une singularité en $a = 0$; cependant, pour les ensemble d'astéroïdes de la ceinture principale, $a > 2.3$[3], et cette transformation fortement non linéaire va préserver la notion de voisinage (deux voisins dans (a, e, i) resteront voisins dans (x, y, z)). L'algorithme de segmentation par détection des vallées de la fonction $L(p)$ définie au paragraphe précédent s'applique également dans l'espace (x, y, z). Les courbes CORes obtenues expérimentalement sur des données de simulation sont équivalentes sinon meilleures à celles obtenues dans (a, e, i) (voir [12]).

On définit l'entropie associée à l'ensemble des K familles détectées en considérant que pour l'ensemble A_0 des objets appartenant à une famille (les "outliers" sont exclus), la définition des familles A_i, $i = 1, \ldots, K$ définit une partition π_{A_0} au sens de l'équation 4. Le MST est construit dans l'espace (x, y, z), avec $d^E_{i,j} = ||X_i - X_j||$, $X_i = (x_i, y_i, z_i)$. La longueur de graphes ou de sous graphes permet alors d'estimer l'entropie de Rényi d'ordre $\alpha = \frac{d-\gamma}{d} = \frac{3-1}{3} = \frac{2}{3}$ à une constante additive près :

$$L_j = \sum_{i,k/(X_i,X_k)\in A_j^2} d^{(E)}_{i,k}$$

$$H_{int} = 3\log_2\left(\frac{\sum_{j=1}^K L_j}{(\sum_{j=1}^K \mathrm{Card}(A_j))^{2/3}}\right)$$

3. $a = 0UA$ correspond à un astéroïde au centre du soleil...

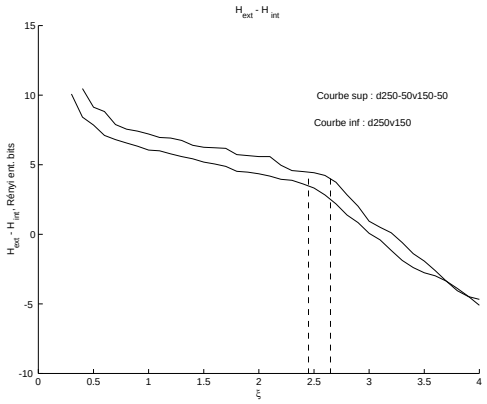

FIG. 2: *Différence entre l'entropie de la distribution des outliers (H_{ext} et l'entropie de la distribution des objets regroupés en familles (H_{int}), pour 2 fichiers de données simulées (d250-50v150-50 et d250v150).*

De même, l'entropie associée à la distribution des 'outliers' (objets n'appartenant à aucune famille identifiée) peut être estimée par :

$$A_{ext} \subset \mathcal{X} \qquad A_{ext} \cup A_0 = \mathcal{X}$$
$$H_{ext} = 3\log_2\left(\frac{L_{ext}}{(\mathrm{Card}(A_{ext}))^{2/3}}\right)$$

Une définition raisonnée de ce qu'est une famille d'objets s'appuie sur la proximité (au sens d'une métrique d ou $d^{(E)}$ définie précédement) des valeurs prises par les paramètres caractéristiques des objets qui constitue cette famille. Par opposition, les "outliers" sont des objets dont les caractéristiques sont dispersées. Naturellement, l'entropie de la distribution des caractéristiques est faible au sein d'une famille, et significativement plus importante pour les "outliers". Nous proposons alors le seuil de construire le seuil suivant: Soit ξ le paramètre permettant de définir le seuil η à appliquer à la fonction $L(p)$ (voir section 3.2), via la relation

$$\eta = \xi.\mathrm{std}\left[d^{(E)}_{i,j}), d^{(E)}_{i,j} \in \mathcal{T}^\star\right]$$

Une augmentation de ξ (donc de η) conduit à élargir la ou les familles déjà existantes, voire à en proposer de nouvelles. D'après les équations précédentes, l'entropie H_{int} doit rester faible tant que des outliers, de paramètres distants de ceux des familles et fortement dispersés, ne sont pas adjoints aux familles d'objets. Dans cette perspective, l'entropie H_{int} doit augmenter et corrélativement, H_{ext} doit diminuer. La figure 2 représente la différence $H_{int} - H_{ext}$ en fonction de ξ. La détection du coude de la courbe $H_{int} - H_{ext} = f(\xi)$ conduit sur cet exemple à retenir $\xi = 2.65$ pour le fichier d250-50v150-50. Ce fichier contient 2 familles d'objets assez proches (voir figure 1); Le fichier d250v150 ne contient que la plus grosse de ces deux familles.

3.3 Validation

Il existe dans la littérature un nombre important de définition d'indice de validité d'une segmentation. Parmi les plus souvent mentionnés, les indices de Dunn, Davies Bouldin et plus récemment Chou et Sun ont été testés sur la partition de A_0 obtenue par notre algorithme. Il est important de souligner que les outliers (constituant une sorte de "famille" en un sens étendu)

ξ optimal		
-	d250v150	d250_50v150_50
Dunn's measure	2.3	0.6
Davies-Bouldin's measure	2.6	2.7
CS's measure	2.7	3.4
Entropy's measure	2.45	2.65
CORE (Bayes Minimax)	2.5	2.5

FIG. 3: *Values estimées de ξ optimal*

ne sont pas pris en compte dans le calcul de ces indices. Soit C l'ensemble des indices associé à l'ensemble des familles définies par la partition π_{A_0}, et n_C le nombre de ces familles.

Indice de Dunn[6]

$$DI(A_0) = \min_{i \in C} \left\{ \min_{j, j \neq i} \left\{ \frac{\delta(A_i, A_j)}{\max_{k \in C}\{\Delta(A_k)\}} \right\} \right\} \quad \text{où}$$

$$\delta(A_i, A_j) = min\{d(\underline{x}_i, \underline{x}_j) \mid \underline{x}_i \in A_i, \underline{x}_j \in A_j\}$$

$$\Delta(A_j) = max\{d(X_i, X_j) \mid X_i, X_j \in A_j\}$$

d est la métrique utilisée, i.e. $d^{(E)}$.

Indice de Davies Bouldin

Dans [4], un ensemble de propriétés qui doivent être vérifiées par un indice de validité de segmentation est proposé. Un choix simple conduit à définir l'indice suivant:

$$DB(\pi_{A_0}) = \sum_{i=1}^{n_C} R_i \quad \text{où}$$

$$R_i = \max_{j \in C, j \neq i} \left\{ \frac{s_i + s_j}{d_{ij}} \right\} \quad \text{avec}$$

$$d_{ij} = \| V_i - V_j \| \quad \text{and} \quad s_i = \frac{1}{\mathrm{card(A_i)}} \sum_{X \in A_i} \| X - V_i \|$$

et V_i est le barycentre de A_i.

Indice de Chien et Sun[3]

Plus adapté aux cas où les familles peuvent être de tailles très différentes, cet indice est défini par:

$$CS(\pi_{A_0}) = \frac{\sum_{i=1}^{n_C} \{ \frac{1}{\mathrm{Card A_i}} \sum_{X_j \in A_i} \max_{X_k \in A_i} \{d_{j,k}\} \}}{\sum_{i=1}^{n_C} \{ \min_{j, j \neq i} \{d(V_i, V_j)\} \}} \quad \text{où}$$

$$V_i = \frac{1}{\mathrm{card(A_i)}} \sum_{x_j \in A_i} X_j$$

Ces indices ne sont pas équivalents et sont plus ou moins sensibles à la géométrie des familles. L'étude de leur sensibilité dépasse le cadre de cet article. Cependant, les résultats obtenus sont cohérents avec le critère entropique proposé: Pour chacun des ces définition, le seuil ξ_{opt} retenu est celui qui optimise l'indice de validité de la segmentation. L'ensemble des résultats est présenté dans le tableau 3. La dernière ligne indique la valeur de ξ à l'intersection de la courbe CORe du détecteur et de la seconde diagonale dans le graphe P_{FA}, P_D, qui correspondrait donc à un critère Minimax dans un test d'hypothèse binaire Bayesien (interloper vs famille).

4 Résultats, Conclusions

Des résultats obtenus sur des fichiers de simulation (calculés à partir de modèle physique de collisions) permettent de situer les performances de la méthode proposée par rapport aux méthodes pré-existante, à travers l'étude de courbes CORe, dans le cas où une ou plusieurs familles d'objets sopnt présentes dans les données. Quelques indices de validité de segmentation sont calculés et exploités pour déterminer le seuil optimal et permettent de valider et établir l'intérêt de la méthode, par construction insensible à la non convexité des familles et à une 'éventuelle structure 'filamentaire' de ces familles.

Références

[1] Ph. Bendjoya, A. Cellino, Cl. Froeschlé, V. Zappala: "Asteroid dynamical families : a reliabity test for different new identification methods." A & A. vol. **272**, pp.651-670, 1993.

[2] P. Bendjoya : "A classification of 6479 asteroids into families by means of the wavelet clustering method." A & A. Supp. vol. **102**, pp.25-55, 1993.

[3] Chien-Hsing Chou, Mu-Chun Sun, Eugene Lai: "A new cluster validity measure for clusters with different densities.", 2003.

[4] Davies, DL, Bouldin, D.W. "A cluster separation measure". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.1, No2,1979

[5] R.O.Duda, P.E.Hart, D.G.Stork : "Pattern Classification", Wiley Interscience Ed., 2001.

[6] Dunn, J. C. "Well separated clusters and optimal fuzzy partitions", *J. Cybern.* Vol.4, pp.95-104, 1974.

[7] A.O.Hero, O.Michel : "Asymptotic theory of greedy approximations to minimal K-point random graphs", *IEEE Trans. on Information theory*,vol. IT-45, no.6, pp.1921-1939, Sept. 1999.

[8] A.O. Hero, B.Ma, O. Michel and J.D. Gorman, "Applications of Entropic Spanning Graphs," *IEEE Signal Processing Magazine*, vol.19, No.5, pp.85-95, 2002.

[9] A.K.Jain, M.N.Murty, P.J.Flynn :"Data clustering: a review", *ACM Computing Surveys*, vol.31-3,pp.264-323., 1999.

[10] D. Marchette : "Random graphs for statistical pattern recognition", Wiley Series in Probabilioty and Statistics, 2004.

[11] 0.Michel, A.O.Hero, P.Flandrin : "Entropie conditionnelle de Rényi et Segmentation", Gretsi'2001, Toulouse, France, paper 254.

[12] O.Michel, P.Bendjoya : "Unsupervised clustering with MST: Applicationto asteroid data", PSIP'05, Toulouse, pp.29-33.

[13] http://csbl.bmb.uga.edu/ olman, mentionned by A.O.Hero (private communications)

[14] J.E.Yukich :"Probability Theory of Classical Euclidean Optimization problems.", *Lecture Notes in mathematics*, 1675, Springer, 1998.

[15] V. Zappala, A. Cellino, P. Farinella, and Z. Knezevic: "Asteroid families. I - Identification by hierarchical clustering and reliability assessment." Astronomical Journal, vol. **100**, pp.2030-2046, 1990.

[16] Zappala V., Bendjoya Ph., Cellino A., Farinela P., Froeschlé Cl.: "Asteroid families : search in a 12,487 asteroid sample whith two different clustering techniques" *Icarus*, vol. **116**, pp.291-314, 1995.