Clustering de métrique et clustering de graphe

Fabien de Montgolfier, Mauricio Soto, Laurent Viennot

Clustering de métrique et clustering de graphe

Fabien de Montgolfier, Mauricio Soto et Laurent Viennot

fm@liafa.jussieu.fr mausoto@liafa.jussieu.fr Laurent.Viennot@inria.fr
Équipe-Projet GANG entre l'INRIA Paris-Rocquencourt et le LIAFA, UMR 7089 CNRS - Université Paris Diderot.

Ce papier s'intéresse aux liens entre deux concepts : d'une part le clustering de graphes, mesuré par la modularité de Newman, et d'autre part le clustering d'un espace métrique, mesuré par la somme des carrés des distances au barycentre des clusters. Nous montrons qu'en passant de l'espace à un graphe de façon "naturelle" (par boules unitaires) et en appliquant un algorithme "naturel" de clustering (par r-net), on obtient un clustering de modularité bornée inférieurement (en fonction de la dimension de grille de l'espace et pour certains rayons de r-net). Quelques simulations en espace euclidien viennent illustrer le propos en comparant deux algorithmes pour les deux mesures de qualité.

1 Introduction

Nous essayons de tisser un lien entre clustering de graphe et clustering d'espace métrique. En effet, le problème du clustering a été largement étudié séparément sur les graphes et sur les métriques. Pour les graphes, la notion de modularité introduite par Newman [New06] permet de mesurer la qualité d'une partition des sommets du graphe en clusters. Il existe de nombreux algorithmes calculant une telle partition [Sch07]. D'autre part, calculer une partition de points dans un espace métrique a largement été étudié en géométrie (voir par exemple [Ber06] qui inclut bien d'autres domaines encore). Pourtant il est naturel d'associer une métrique à un graphe. Réciproquement, on peut associer à une métrique le graphe de boules unitaires. Il existe donc un pont naturel entre ces deux domaines qu'il nous semble intéressant d'explorer. Nous proposons un premier angle d'attaque en montrant qu'un clustering par recouvrement par des boules fournit un clustering de bonne modularité pour le graphe de boules unitaires.

1.1 Clustering géométrique et k-means

Dans la suite, on considère un espace V, ayant un nombre fini de points, et muni d'une métrique (ou distance) $dist$. Un *clustering* est toujours une partition de V. Un algorithme que nous appellerons Glouton-Simple de clustering est le suivant : on se fixe un rayon R. On choisit un point quelconque v_1, et le premier cluster C_1 est l'ensemble des points à distance inférieure à R de v_1. Puis on recommence sur $V - C_1$ pour obtenir un cluster C_2, et ainsi de suite jusqu'à vider le nuage. Dans la suite on verra que ce clustering n'est pas mauvais. Notons déjà que les clusters ont un rayon borné par R.

Le but de toute une famille d'algorithmes plus sophistiqués est aussi de calculer un partitionnement du nuage en clusters, qui sont alors des parties "aussi regroupées que possible" du nuage. On peut citer : k-means, k-medians, k-medoids... Dans ces algorithmes k, le nombre de clusters, est une donnée du problème, non un résultat comme avec Glouton-Simple.

Intéressons-nous à l'un des plus fameux. Un **k-means** est un clustering en au plus k clusters minimisant, parmi tous les clusterings, la distance au carré des points au barycentre de leur cluster. Plus formellement, soit $\mathcal{C} = \{C_1, ...C_k\}$ un clustering . On note μ_i le barycentre (centroïde) au sens de $dist$ du cluster C_i. La **fonction objectif** ou **WCSS** *(within-cluster sum of squares)* du clustering est $WCSS(\mathcal{C}) = \sum_{i=1}^{k} \sum_{v \in C_i} dist(v, \mu_i)^2$

Notez qu'on se place en fait dans un $\mathbb{R}$-espace vectoriel $\mathcal{E}$ contenant V, car les barycentres peuvent ne pas appartenir à V. Certains clusters peuvent être vides, ils comptent alors pour zéro dans la somme. Un k-means est un clustering d'objectif minimal. En calculer un est NP-complet. Toutefois l'**algorithme de Lloyd** est la façon standard de l'approximer (voir [KMN+02] pour une étude approfondie). Il répète jusqu'à stabilisation ($\mathcal{C}_i = \mathcal{C}_{i+1}$, qui n'est pas garantie) les deux étapes suivantes :

1. Calculer l'ensemble $\mathcal{B}_i$ des barycentres des clusters de $\mathcal{C}_i$
2. $\mathcal{C}_{i+1}$ est le clustering où tout sommet v est assigné au cluster du point de $\mathcal{B}_i$ le plus proche de lui.

Notons qu'à l'étape $i = 0$ on peut soit choisir $\mathcal{C}_0$ soit choisir $\mathcal{B}_0$ et en déduire $\mathcal{C}_0$ par le point 2.

1.2 Clustering de graphe et modularité

Une définition standard en théorie des graphes de la *qualité* d'un clustering est la modularité de Newman [New06]. Soit $G=(V,E)$ un graphe et $\{C_1,...C_k\}$ une partition de V en k clusters. La **modularité** de ce clustering est donnée par

$$Q(C)=\sum_{i=1}^{k}\left[\frac{|E(C_i)|}{m}-\frac{(\sum_{v\in C_i}d_v)^2}{4m^2}\right]$$

où $m=|E|$ est le nombre d'arêtes, d_u est le degré du sommet u et $E(C_i)=E\cap(C_i\times C_i)$ les arêtes internes au cluster. La *modularité d'un graphe G* est celle du clustering de G de modularité maximum. La modularité est comprise entre $-1/2$ et 1 [**?**]. Son but est de mesurer la connectivité des clusters par rapport au graphe aléatoire ayant la même distribution de degrés (qui a une modularité nulle a.f.p.)

1.3 Un pont entre les deux : graphe de boules, dimension de grille et r-net

Nous souhaiterions appliquer cette définition à un clustering de l'espace métrique V. Mais il nous faut pour cela un graphe. La façon la "plus naturelle" (ce propos n'engage que les auteurs...) est de fixer une longueur L d'arête et de créer l'ensemble d'arêtes $E=\{uv|dist(u,v)\leqslant L\}$. Nous appellerons *graphe de boules* unitaires ce graphe, car c'est aussi le graphe d'intersection des boules de centre $v\in V$ et de rayon $L/2$. Nous le noterons $G_L=(V,E)$. On peut donc maintenant parler de la "modularité de V" comme étant celle calculée dans G_L.

On note $B(u,r)\subset V$ (boule de centre u et de rayon r) les points de V situés à distance au plus r de u. On remarque que si les points sont distribués de façons quelconque dans l'espace, certains clusters auront beaucoup de sommets et d'autres peu, alors qu'un axiome d'homogénéité, même assez lâche, fournit des clusters de taille plus homogène. Nous allons étudier les espaces en fonction de leur *dimension de grille*, mesure de croissance ayant des conséquences sur la densité introduite par Karger et Ruhl :

Définition 1 (Dimension de grille [KR02] aussi appellée *bounded growth*) *L'espace V est de dimension de grille bornée par $\gamma>0$ si, en doublant le rayon d'une boule, on en multiplie le volume par moins de γ :*

$$\forall x\in V,\forall r>0,|B(x,r)|\leqslant\gamma\cdot|B(x,r/2)|$$

On remarque qu'un espace euclidien de dimension d est de dimension de grille $\gamma=2^d$ (pour la norme $|\ |_1$). Il y a un lien entre le clustering Glouton-Simple décrit ci-dessus et le graphe G_L, car les deux sont faits à partir de boules centrées sur les points du nuage. Présentons un outil algorithmique :

Définition 2 (r-net) *Un sous-ensemble U d'un espace V est un r-net si*

- $\forall u,u'\in U$, $dist(u,u')>r$ *(les points sont à distance au moins r)*
- $\forall v\in V\ \exists u\in U\ dist(u,v)\leqslant r$ *(couverture du nuage par les boules de rayon r centrées sur u)*

Un processus glouton peut construire un r-net. Ensuite, en définissant pour tout $u_i\in U$ son cluster C_i comme l'ensemble des points plus proches de u_i que de tout autre point du r-net (les cas d'égalité sont cassés de manière quelconque), on obtient un clustering qui généralise le Glouton-Simple, puisqu'il s'applique à tout r-net. Dans la section suivante, nous montrons qu'un tel clustering a une modularité strictement positive et bornée inférieurement par une fonction de la dimension.

2 Borne inférieure de modularité

Théorème 1 *Soit V un espace fini de n points, muni d'une métrique dist, de dimension de grille bornée par γ, et $R\geqslant 0$ tel que pour tout $v\in V$ on ait $|B(x,R/2)|>1$ et $|B(x,R/2)|=o(\sqrt{n})$. On a :*

$$Q(G_R)\geqslant\frac{1}{2\gamma^3}-o(1)$$

Tout espace métrique fini ayant une certaine dimension de grille γ, ce théorème s'applique pour tout rayon vérifiant la condition sur les tailles des boules. Cependant, la borne obtenue est d'autant plus élevée que la valeur de γ est faible.

Démonstration : Soit $U=\{u_1,\ldots,u_k\}$ un R-net de V (il en existe un puisqu'un algorithme glouton en produit). On construit le clustering $\mathcal{C}_U=\{C_i\}_{i\in\{1,\ldots,k\}}$ avec $C_i=\left\{v\in V|\forall j\neq i,\ dist(v,u_i)\leqslant dist(v,u_j)\right\}$.

Nous avons par construction $B(u_i, R/2) \subseteq C_i \subseteq B(u_i, R)$. Posons $b_i = |B(u_i, R/2)|$. La dimension de grille nous donne : $b_i \leqslant |C_i| \leqslant \gamma \cdot b_i$. De plus, comme U est un R-net ses points sont mutuellement à distance au moins R. Les points de b_i sont donc tous reliés dans G_R et forment une clique incluse dans C_i. On a alors :

$$Q(\mathcal{C}_U) = \sum_{i=1}^{k} \left[\frac{|E(C_i)|}{m} - \frac{(\sum_{v \in C_i} d_v)^2}{4m^2} \right] \geqslant \sum_{i=1}^{k} \left[\frac{b_i(b_i-1)}{2m} - \frac{(\sum_{v \in C_i} d_v)^2}{4m^2} \right]$$

pour tout $v \in C_i$, on a $d_v = |B(v,R)| \leqslant |B(u_i, 2R)| \leqslant \gamma^2 b_i$, donc :

$$Q(\mathcal{C}_U) \geqslant \sum_{i=1}^{k} \left[\frac{b_i(b_i-1)}{2m} - \frac{(|C_i|\gamma^2 b_i)^2}{4m^2} \right] \geqslant \sum_{i=1}^{k} \left[\frac{b_i(b_i-1)}{2m} - \frac{\gamma^6 b_i^4}{4m^2} \right]$$

Or,
$$\sum_{i=1}^{k} b_i(b_i-1) \leqslant 2m = \sum_{u \in V(G)} d_u \leqslant \sum_{i=1}^{k} |C_i|\gamma^2 b_i \leqslant \gamma^3 \sum_{i=1}^{k} b_i^2,$$

Et donc :
$$Q(\mathcal{C}_U) \geqslant \frac{\sum_{i=1}^{k} b_i(b_i-1)}{\gamma^3 \sum_{i=1}^{k} b_i^2} - \frac{\gamma^6 \sum_{i=1}^{k} b_i^4}{2m \sum_{i=1}^{k} b_i(b_i-1)}$$

Posons $\bar{b} = \max_i \{b_i\}$. Comme $\frac{\sum_{i=1}^{k} b_i(b_i-1)}{\sum_{i=1}^{k} b_i^2} = 1 - \frac{\sum_{i=1}^{k} b_i}{\sum_{i=1}^{k} b_i^2} \geqslant 1 - \frac{\sum_{i=1}^{k} b_i}{\sum_{i=1}^{k} 2b_i} \geqslant \frac{1}{2}$, on a :

$$Q(\mathcal{C}_U) \geqslant \frac{1}{2\gamma^3} - \frac{\gamma^6 \bar{b}^2 \sum_{i=1}^{k} b_i^2}{2m \sum_{i=1}^{k} b_i(b_i-1)} \geqslant \frac{1}{2\gamma^3} - \frac{\gamma^6 \bar{b}^2}{m}$$

Enfin, comme $\bar{b} \in o(\sqrt{n})$ et $m \geqslant \frac{n}{2}$, nous obtenons $Q(\mathcal{C}_U) = \frac{1}{2\gamma^3} - o(1)$ donc $Q(G_R) = \frac{1}{2\gamma^3} - o(1)$ □

3 Simulations en espace euclidien

Des simulations montrent que l'on obtient en pratique une bien meilleure valeur que la borne, dans le cas euclidien uniforme. Le choix de la géométrie euclidienne est dû à la nécéssité d'avoir un $\mathbb{R}$-espace vectoriel pour comparer avec k-means. Et la distribution uniforme assure la dimension de grille bornée, que trop d'hétérogénéité violerait.

On pose donc le modèle suivant : n points tirés au hasard, uniformément et indépendemment, dans un hypercube de côté 1, et de dimension d. On pose $R < 1$ et on construit le R-net gloutonnement, puis le clustering au plus proche (en brisant arbitrairement les égalités). La figure (page suivante en haut à gauche) montre la modularité en fonction de R. Notons que les conditions du théorème sur la valeur de R ne sont pas remplies pour R trop petit ou trop grand, ce qui peut amener à une modularité faiblement négative.

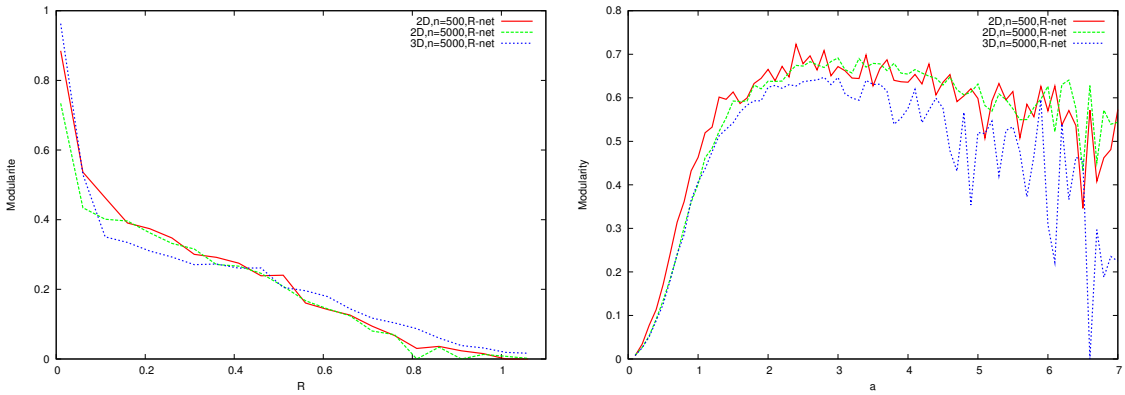
Dans la preuve nous avons posé l'égalité entre R, le rayon des clusters, et L, le rayon des boules de G_L. Cela permet de maximiser la taille de la clique b_i dont nous avons besoin. Il peut être intéressant de comparer ces deux paramètres indépendemment. C'est ce que nous faisons avec la deuxième figure, pour R fixé à $0,1$. Le paramètre est $a = R/L$. Partout ailleurs qu'en cette figure on a $R = L$.

Les différentes courbes font varier le nombre de points (500 ou 5000) et la dimension ($d = 2$ ou 3) de l'espace. Comme on voit ces deux paramètres sont sans influence claire. Nous avons aussi utilisé un tore pour supprimer les effets de bord : les résultats sont quasiment identiques.

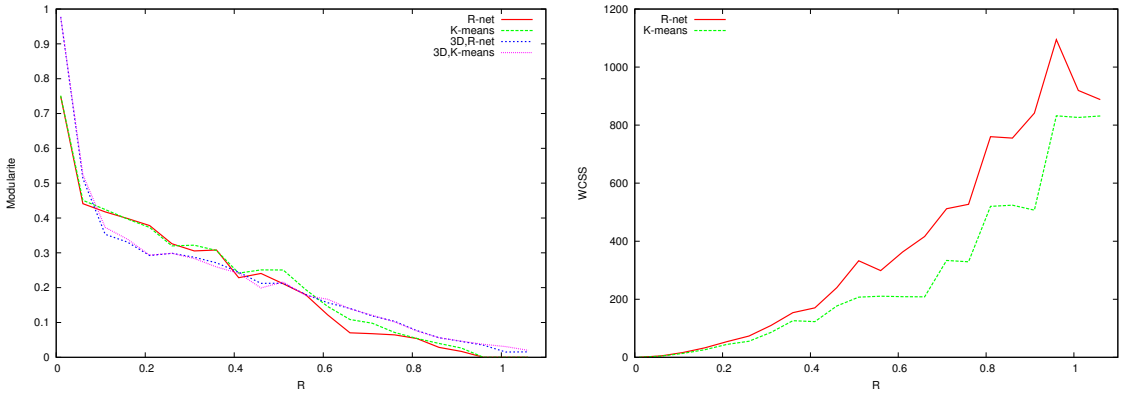
Nous avons comparé empiriquement avec le résultat de l'algorithme Lloyd pour les k-means, en l'initialisant de sorte que $\mathcal{B}_0$ est un R-net. La troisième figure reprend la première, mais appliquée à $\mathcal{C}_\infty$ (résultat de Lloyd), comparé à $\mathcal{C}_0$ (celui du R-net). Nous constatons que l'algorithme de Lloyd produit un clustering de modularité sensiblement égale : elle est quasi-invariante d'une itération à l'autre.

Remarquons que pour R trop petit le graphe est quasi sans arêtes. On obtient de plus énormément de clusters très petits (par exemple 2757 clusters pour 5000 sommets et $R = 0,01$) et donc Lloyd, pour k tellement grand, atteint un objectif presque nul. Le graphe devient connecté (avec forte proba.) quand $n\pi R^2 \geqslant \ln n$, soit $R \geqslant 0,023$ pour $n = 5000$. Pour $R = 0,11$ on n'a plus que $k = 59$ clusters, valeur qui tombe à $k = 10$

pour $R = 0,31$ et $k = 3$ pour $R = 0,71$. Les courbes sont donc signifiantes (clustering non trivial) seulement quand R est environ entre $0,1$ et $0,5$. Le nombre de clusters varie en $1/\pi R^2$.



Enfin, la figure en bas à droite montre que le R-net est déjà une bonne heuristique pour le calcul de k-means : utilisé pour initialiser Lloyd, ce dernier fait baisser WCSS mais pas énormément (pour $0,1 < R < 0,5$). Question nombre de clusters, Lloyd ne vide quasiment jamais de cluster si on l'initialise avec le R-net, alors qu'une initialisation de $\mathcal{C}_0$ comme partition aléatoire crée beaucoup de clusters vides dans $\mathcal{C}_\infty$.



4 Conclusion

Nous montrons comment construire un clustering d'un espace de dimension de grille bornée par γ dont la modularité, dans le graphe de boules, est au moins $\frac{1}{2\gamma^3} - o(1)$. Dans le cas euclidien aléatoire uniforme, des simulations nous montrent que nous sommes bien plus haut que la borne. Nous utilisons des r-nets soit directement, soit pour initialiser Lloyd. On constate que le résultat est assez identique pour la modularité, et que pour le WCSS un r-net est déjà très efficace, mais bien sûr Lloyd l'améliore mais pas sensiblement. Le r-net est donc utile à la fois pour prouver la borne, mais aussi pour calculer des clusterings de qualité.

Références

[BDG+08] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Trans. on Knowledge and Data Engineering*, 20 :172–188, 2008.

[Ber06] P. Berkhin. *A Survey of Clustering Data Mining Techniques*, pages 25–71. 2006.

[KMN+02] T Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu. An efficient k-means clustering algorithm : Analysis and implementation. *IEEE TPAMI*, 24 :881–892, 2002.

[KR02] D. Karger and M. Ruhl. Finding nearest neighbors in growth-restricted metrics. In *Proceedings of STOC, 34th Symposium on Theory Of Computing*, pages 741–750. ACM, 2002.

[New06] M Newman. Modularity and community structure in networks. *PNAS*, 103(23) :8577+, 2006.

[Sch07] S. Schaeffer. Graph clustering. *Computer Science Review*, 1(1) :27 – 64, 2007.