Clustering de metrique et clustering de graphe

Fabien de Montgolfier, Mauricio Soto, Laurent Viennot

Projet ANR Aladdin & Équipe-projet INRIA GANG & LIAFA (CNRS-Université Paris Diderot)

AlgoTel, Cap Estérel / Mai 2011

Clustering : le contexte

1. On a un graphe de la vraie vie
2. On le saucissonne ($\rightarrow$ partition en clusters)
3. On raconte des choses dessus

Domaines : algorithmique (2) et
socio-bio-physico-géo-networko-métrologo-tétrapilectomie (3)

Qu'en dire d'un point de vue structurel ?

Modularité

[Newman & Girvan 2002, 2004] *The* qualité standard de clustering.

Définition

$$\mathcal{Q}(\mathcal{C}) = \sum_{i=1}^{k} \frac{m(C_i)}{m(G)} - \sum_{i=1}^{k} \frac{\mathrm{vol}(C_i)^2}{\mathrm{vol}(G)^2}$$

avec :

- $\mathcal{C} = \{C_1 ... C_k\}$: clustering à mesurer
- $m(A)$: nombre d'arêtes de $G[A]$
- $\mathrm{vol}(A)$: somme des degrés. $vol(A) = \sum_{x \in A} deg(x)$

Modularité

[Newman & Girvan 2002, 2004] *The* qualité standard de clustering.

Définition

$$\mathcal{Q}(\mathcal{C}) = \sum_{i=1}^{k} \frac{m(C_i)}{m(G)} - \sum_{i=1}^{k} \frac{\mathrm{vol}(C_i)^2}{\mathrm{vol}(G)^2}$$

- Terme de gauche : **densité** d'arêtes internes $\in [0, 1]$

Modularité

[Newman & Girvan 2002, 2004] *The* qualité standard de clustering.

Définition

$$\mathcal{Q}(\mathcal{C}) = \sum_{i=1}^{k} \frac{m(C_i)}{m(G)} - \sum_{i=1}^{k} \frac{\mathrm{vol}(C_i)^2}{\mathrm{vol}(G)^2}$$

- Terme de gauche : **densité** d'arêtes internes $\in [0, 1]$
- Terme de droite ?

Modularité

[Newman & Girvan 2002, 2004] *The* qualité standard de clustering.

Définition

$$\mathcal{Q}(\mathcal{C}) = Internes - \frac{\sum_{u,v} deg(u) * deg(v) * MêmeCluster(u,v)}{\text{vol}(G)^2}$$

- Terme de gauche : **densité** d'arêtes internes $\in [0,1]$
- Terme de droite ? Densité d'arêtes internes qu'aurait ce clustering d'un graphe **aléatoire** avec même distribution des degrés (shuffle des arêtes). $\in [0,1]$

Valeurs de $\mathcal{Q}$?

- Terme de gauche : densité d'arêtes internes $\in [0, 1]$
- Terme de droite : densité d'arêtes internes après shuffle $\in [0, 1]$
- Donc $\mathcal{Q} \in [-1, 1]$

Valeurs de $\mathcal{Q}$?

- Terme de gauche : densité d'arêtes internes $\in [0, 1]$
- Terme de droite : densité d'arêtes internes après shuffle $\in [0, 1]$
- Donc $\mathcal{Q} \in [-1, 1]$
- Interprétation :
  - Négatif : mauvais
  - Nul : non signifiant
  - Positif : bon
  - $1 =$ clustering parfait ?

Valeurs de $\mathcal{Q}$?

- Terme de gauche : densité d'arêtes internes $\in [0, 1]$
- Terme de droite : densité d'arêtes internes après shuffle $\in [0, 1]$
- Donc $\mathcal{Q} \in [-1, 1]$
- Interprétation :
  - Négatif : mauvais
  - Nul : non signifiant
  - Positif : bon
  - $1 =$ clustering parfait ?
- En fait $\mathcal{Q} \in [-0.5, 1[$
- En particulier $\mathcal{Q} \leqslant 1 - \frac{DegréMax^2}{4m^2}$

Modularité d'un graphe

Définition

$$\mathcal{Q}(G) = \max_{\mathcal{C}}\{\mathcal{Q}(\mathcal{C})\}$$

Modularité d'un graphe

Définition

$$\mathcal{Q}(G) = \max_{\mathcal{C}}\{\mathcal{Q}(\mathcal{C})\}$$

Tout graphe a modularité $\geqslant 0$

Démonstration : c'est la qualité d'un clustering en un seul cluster

Au passage, le clustering du graphe complet est 0...

Modularité d'un graphe

Définition

$$\mathcal{Q}(G) = \max_{\mathcal{C}}\{\mathcal{Q}(\mathcal{C})\}$$

Tout graphe a modularité $\geqslant 0$

Démonstration : c'est la qualité d'un clustering en un seul cluster

Au passage, le clustering du graphe complet est 0...

Théorème [Brandes & alii 2008]

Calculer $\mathcal{Q}(G)$ est NP-complet

Vaste littérature sur la question. Ce n'est pas notre propos.

Modularité asymptotique d'une classe de graphes

Définition

Une classe de graphes $\mathcal{G}$ a modularité asymptotique au moins ℓ si, pour toute suite de graphes $\{G_i \in \mathcal{G}\}_{i \in \mathbb{N}}$

$$\lim_{i \to \infty} (\mathcal{Q}(G_i)) \geqslant \ell$$

Modularité asymptotique d'une classe de graphes

Définition

Une classe de graphes $\mathcal{G}$ a modularité asymptotique au moins ℓ si, pour toute suite de graphes $\{G_i \in \mathcal{G}\}_{i \in \mathbb{N}}$

$$\lim_{i \to \infty} (\mathcal{Q}(G_i)) \geqslant \ell$$

Question

Existe-t-il des classes de modularité asymptotiques 0 ?

Modularité asymptotique d'une classe de graphes

Définition

Une classe de graphes $\mathcal{G}$ a modularité asymptotique au moins ℓ si, pour toute suite de graphes $\{G_i \in \mathcal{G}\}_{i \in \mathbb{N}}$

$$\lim_{i \to \infty} (\mathcal{Q}(G_i)) \geqslant \ell$$

Question

Existe-t-il des classes de modularité asymptotiques 0 ?

Réponse

Oui : la classe des étoiles $K_{1,n}$

Modularité asymptotique d'une classe de graphes

Définition

Une classe de graphes $\mathcal{G}$ a modularité asymptotique au moins ℓ si, pour toute suite de graphes $\{G_i \in \mathcal{G}\}_{i \in \mathbb{N}}$

$$\lim_{i \to \infty} (\mathcal{Q}(G_i)) \geqslant \ell$$

Question

Existe-t-il des classes de modularité asymptotiques 1 ?

Modularité asymptotique d'une classe de graphes

Définition

Une classe de graphes $\mathcal{G}$ a modularité asymptotique au moins ℓ si, pour toute suite de graphes $\{G_i \in \mathcal{G}\}_{i \in \mathbb{N}}$

$$\lim_{i \to \infty} (\mathcal{Q}(G_i)) \geqslant \ell$$

Question

Existe-t-il des classes de modularité asymptotiques 1 ?

Réponse

Oui : un clustering de k copies du même graphe connexe a modularité (exactement) $1 - 1/k$

Modularité asymptotique d'une classe de graphes

Définition

Une classe de graphes $\mathcal{G}$ a modularité asymptotique au moins ℓ si, pour toute suite de graphes $\{G_i \in \mathcal{G}\}_{i \in \mathbb{N}}$

$$\lim_{i \to \infty} (\mathcal{Q}(G_i)) \geqslant \ell$$

Question

Combien de clusters faut-il ?

Modularité asymptotique d'une classe de graphes

Définition

Une classe de graphes $\mathcal{G}$ a modularité asymptotique au moins ℓ si, pour toute suite de graphes $\{G_i \in \mathcal{G}\}_{i \in \mathbb{N}}$

$$\lim_{i \to \infty} (\mathcal{Q}(G_i)) \geqslant \ell$$

Question

Combien de clusters faut-il ?

Réponse

Beaucoup : en effet un clustering en k clusters
a modularité **au plus** $1 - 1/k$

Graphes bien décomposables

Définition

Un graphe G est (k, c, e)-*décomposable* si on peut le découper en k clusters avec :

- La densité d'arêtes externes est e, et
- Chaque cluster a volume au plus $c\frac{\text{vol}(G)}{k}$,

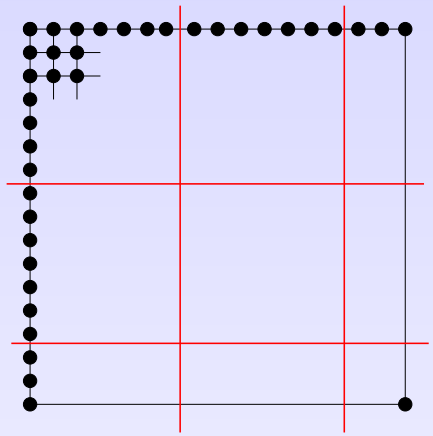
Lemme

Si G est (k, c, e)-décomposable alors $\mathcal{Q}(G) \geqslant 1 - e - \frac{c^2}{k}$.

But : faire tendre e vers 0, k vers ∞, $c = o(\sqrt{(k)})$
$\rightarrow$ si c'est faisable la classe a modularité asymptotique 1 !

Grilles

On coupe la grille de taille $a \times a$ en k carrés.



a=17 k=9

- Nombre d'arêtes externes $\leqslant \frac{\sqrt{k}-1}{a-1}$
- Volume d'un cluster $\leqslant 4\frac{\mathrm{vol}(G)}{k}$
- Couper en carrés de côté $\sqrt[3]{n} \rightarrow \ k = \Theta(a^{-2/3})$
- La modularité de ce clustering est de $1 - \Theta\left(\frac{1}{\sqrt[3]{n}}\right)$.
- Les grilles ont donc modularité asymptotique 1 (intuitif ?)

Tores en dimension d

- G = Hypertoretore (variante de grille) de dimension d
- On prend la plus grande dimension et on coupe en b tranches
- Crée au plus $\frac{2b}{dn^{1/d}}$ arêtes externes (petit devant les dn arêtes)
- Volume d'un cluster au plus $2\frac{\mathrm{vol}(G)}{k}$
- On coupe en $b = \sqrt{2d}n^{1/2d}$ tranches
- Alors modularité $1 - \Theta(n^{-1/2d})$.

Hypercube

- G = hypercube de dimension d. sommets $\Leftrightarrow$ nombres à d bits
- Ressemble à un hypertore de dim. non constante et côté 2
- Clustering :
  - Prendre une longueur de préfixe b
  - Cluster C_a = nombres commençant par le préfixe a (à b bits)
- Chaque sommet touche b arêtes externes et $d - b$ internes
- Volume cluster $d2^{d-b}$ et volume hypercube $d2^d$
- On prend $b = \log_2(d)/2$
- Alors $\mathcal{Q}(G) \geqslant 1 - \dfrac{\log_2(d)}{d} = 1 - \Theta\left(\dfrac{\log\log n}{\log n}\right)$

Espace métrique

- Souvent on n'a pas un graphe mais
  - un nuage de points V dans $\mathbb{R}^d$,
  - et une distance *dist*
- Pleins d'algos (k-means, ACP,...)
- Outils de graphes inapplicables ? Modularité ?

Espace métrique

- Souvent on n'a pas un graphe mais
  - un nuage de points V dans $\mathbb{R}^d$,
  - et une distance *dist*
- Pleins d'algos (k-means, ACP,...)
- Outils de graphes inapplicables ? Modularité ?
- **Graphe de boules** $G_L = (V, \{uv \mid dist(u, v) \leqslant L\})$

Croissance bornée

Grid dimension [Karger & Ruhl 2002]

Dim γ : pour toute boule, rayon $\times 2 \implies$ volume $\times \gamma$ au plus

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

On parle alors de croissance bornée (*bounded growth*)

Exemple

En espace euclidien, norme $||.||_\infty$ (boules cubiques) $\quad \gamma = 2^d$

S'applique aux graphes ou a tout espace métrique

R-net

Définition

$U \subset V$ R-net ($\varepsilon - net$) si

- indépendance : $\forall u, u' \in U \ dist(u, u') > R$
- couverture : $\forall v \in V \ \exists u \in U \ dist(u, v) \leqslant R$

Construction

Facile par un glouton (n'optimise pas $|U|$ mais pas utile)

R-net et clustering

Cluster $C_u :=$ points de V le plus proche de $u \in U \rightarrow |U|$ clusters

On remarque que le rayon des clusters est borné par R

Graphes de croissance bornée

Hypothèses :

- V nuage de n points
- grid dimension γ
- R tel que le volume de $B(x, R/2)$ est entre 2 et $o(\sqrt{n})$
- G_R graphe de boules de rayon R

Théorème

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

Preuve en utilisant le clustering par R-net

Preuve 1/2

Par construction $B(u_i, R/2) \subseteq C_i \subseteq B(u_i, R)$. Soit $b_i = |B(u_i, R/2)|$. Croissance bornée $\implies b_i \leqslant |C_i| \leqslant \gamma b_i$. $B(u_i, R/2)$ est une **clique** de G_R de taille b_i incluse dans C_i. On a :

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

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

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

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

Preuve 2/2

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

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

avec $\bar{b} = \max_i \{b_i\}$.
Mais l'hypothèse sur la taille des boules implique $b_i \geqslant 2$ et $b_i = o(\sqrt{n})$.

$$\text{On a alors : } \frac{\sum_{i=1}^{k} b_i(b_i-1)}{\sum_{i=1}^{k} b_i^2} = 1 - \frac{\sum_{i=1}^{k} b_i}{\sum_{i=1}^{k} b_i^2} \geqslant 1 - \frac{\sum_{i=1}^{k} b_i}{\sum_{i=1}^{k} 2b_i} = \frac{1}{2}$$

$$\text{Et l'inégalité précédente devient : } \mathcal{Q}(\mathcal{C}) \geqslant \frac{1}{2\gamma^3} - \frac{\gamma^6 \bar{b}^2}{m}$$

Pour conclure, comme $\bar{b} = o(\sqrt{n})$ et $m \geqslant \frac{n}{2}$, on a $\mathcal{Q}(\mathcal{C}) \geqslant \frac{1}{2\gamma^3} - o(1)$ et finalement $\mathcal{Q}(G_R) \geqslant \frac{1}{2\gamma^3} - o(1)$

Arbres de degré borné

On considère des graphes de degré **maximum** Δ

Arête centrale

Def : Arête qui coupe un arbre en composantes aussi équilibrées que possible (la taille de la plus petite composante est maximale).
Lemme : il en existe toujours une, incidente au(x) centroïde(s), qui coupe l'arbre en composantes de taille au moins n/Δ

Clustering Glouton$_{\leqslant h}$

Algo : tant qu'il existe un sous-arbre de taille $> h$ le couper par une arête centrale
Lemme : on obtient des clusters de taille entre h/Δ et h.

Arbres de degré borné

Ainsi les arbres sont bien décomposables

Choisir $h \to$ obtenir k clusters, $n-k$ arêtes internes, et $\text{vol}(C_i) \leqslant \Delta^2.\frac{\text{vol}(G)}{k}$

Théorème

Les arbres de degré $o(\sqrt[5]{n})$ ont modularité asymptotique 1

- On calcule $\mathcal{Q}(\mathcal{C}_{\leqslant h}) \geqslant 1 + \frac{1}{n-1} - \frac{\Delta n}{h(n-1)} - \frac{\Delta^4}{n}h$
- On choisit subtilement $h = \frac{n}{\Delta\sqrt{\Delta(n-1)}}$.
- Alors $\mathcal{Q}(\mathcal{C}_{\leqslant h}) \geqslant 1 + \frac{1}{n-1} - \frac{2\Delta^{2.5}}{\sqrt{n-1}}$
- Si $\Delta = o(\sqrt[5]{n})$ on a $\frac{2\Delta^{2.5}}{\sqrt{n-1}} = o(1)$.

Graphes de degré borné

Encore le clustering Glouton$_{\leqslant h}$

- Soit G de degré max Δ et degré moyen d
- On prend un arbre couvrant
- son degré borné est par Δ
- On applique Glouton$_{\leqslant h}$ dessus (couper l'arbre sur arête centrale tant que les paquets ont taille $> h$)
- les paquets ont toujours taille entre h/Δ et h
- $n - k$ arêtes sont internes, toujours (les autres : ?)

Semble un algo trop simple de clustering, pourtant...

Graphes de degré borné

Théorème

Les graphes connexes de degré moyen d et degré max $\Delta = o(\sqrt[5]{d^3 n})$ ont modularité asymptotique $\geqslant \frac{2}{d}$

- Lemme : $\mathrm{vol}(C_i) \leqslant \frac{\Delta^2}{d} \frac{\mathrm{vol}(G)}{k}$
- Le lemme des bien décomp. donne $\mathcal{Q}(\mathcal{C}_{\leqslant h}) \geqslant \frac{2}{d} - \frac{\Delta}{dh} - \frac{\Delta^4}{nd^2} h$
- On prend $h = \frac{\sqrt{dn}}{\Delta^{1.5}}$
- $\mathcal{Q}(G) \geqslant \frac{2}{d} - \frac{\Delta}{d} \frac{\Delta^{1.5}}{\sqrt{dn}} - \frac{\Delta^4}{4nd^2} \frac{\sqrt{dn}}{\Delta^{1.5}} \geqslant \frac{2}{d} - \frac{2\Delta^{2.5}}{d^{1.5}\sqrt{n}} \geqslant \frac{2}{d} - o(1)$

Application : power-law graphs

Power-law graph

Paramètre α. $Pr(X > k) = \Theta(k^{-\alpha})$
Exemples : Zeta law (on remplace $\Theta()$ par constante), Zipf law, Discrete Pareto law...

Lemme

Si $\alpha > 5$ alors degré max $\Delta = o(\sqrt[5]{n})$.

Corollaire

Les graphes de power-law de paramètre $\alpha > 5$ et de degré moyen d ont modularité asymptotique au moins $\frac{2}{d}$.

Dans les graphes «de la vraie vie» d est tout petit (mais α est en général <5...)

Conclusion

Il existe de (grandes) classes de bonne modularité

- Le R-net donne une borne ne dépendant que de la dimension pour les graphes de croissance bornée
- Un arbre couvrant + coupures au centroïde donne une borne ne dépendant que du degré moyen pour les graphes de degré maximum borné
- Pour certaines (grilles, hypertores, hypercubes, arbres de degré borné) on peut même donner la vitesse de convergence vers 1

Conclusion : modularité et treewidth

Relation entre modularité et treewidth d'une classe ?

	Modularité 0	Modularité asymptotique 1
Treewidth 1	Étoiles $K_{1,n}$	Arbres de degré max $o(\sqrt[5]{n})$
Treewidth n	Cliques K_{n+1}	Grilles $n \times n$

Intérêt algorithmique de la tree-decomposition ?

Pas clair qu'une tree-decomposition aide à calculer la modularité

NB : même chose avec la cliquewidth

Conclusion

Relativiser la modularité

- Préjugé usuel : une bonne modularité indique un clustering «naturel», éventuellement «masqué» par l'observation, qu'un algorithme doit «retrouver».
- Or, des graphes très réguliers (grille, hypercube) ont une modularité (asymptotique) maximum...
- Et les graphes power-law vont *tous* avoir bonne modularité dès que $\alpha > 5 \rightarrow$ études empiriques non surprenantes
- Pour nous, la modularité ne peut donc être que la fonction objectif d'un algo d'optimisation, pas une mesure intrinsèque du graphe