Classifications croisées de données de trajectoires contraintes par un réseau routier

Mohamed K. El Mahrsi [1, 2], Romain Guigourès [2, 3], Fabrice Rossi [2] et Marc Boullé [3]

[1] Télécom ParisTech - Département Informatique et Réseaux (Paris, France)
[2] Équipe SAMM EA 4543 - Université Paris I Panthéon-Sorbonne (Paris, France)
[3] Orange Labs (Lannion, France)

13e Conférence Francophone sur l'Extraction et la Gestion des Connaissances

Plan

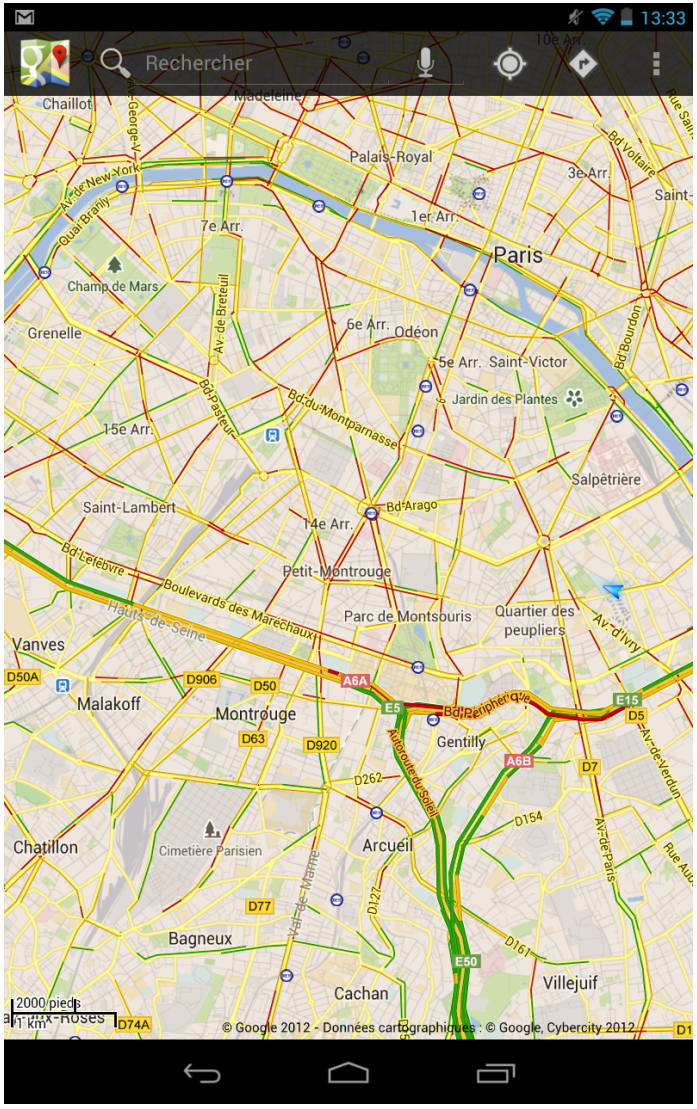
1 Contexte et motivations

2 Approches proposées

3 Résultats

4 Conclusion

Contexte et motivations



- Congestion routière :
  - Dégâts environnementaux
  - Énormes pertes économiques (5,55 G€/an en France)
- Monitoring "classique" du trafic :
  - Capteurs dédiés
  - Coûts élevés
  - Information incomplète
- Facilité de capture et de stockage de données GPS

Données et objectifs

- On dispose d'un ensemble $\mathcal{T}$ de trajectoires
- Qui se sont déplacées sur un réseau routier composé d'un ensemble de segments (noté $\mathcal{S}$)
- Une trajectoire est la suite de segments routiers traversés
  - $T = \langle id, \{s_1, s_2, ..., s_l\} \rangle$ ($\forall 1 \leq i \leq l-1$, s_i et s_{i+1} sont connectés)

Données et objectifs

Objectifs

Effectuer un clustering (classification non supervisée) de ces données pour :

- Dégager des groupes de trajectoires avec des profils similaires ;
- Découvrir des groupes de segments routiers souvent visités ensemble.

- Meilleure compréhension de la dynamique du trafic routier
- Meilleure prise de décision pour le réaménagement de l'infrastructure routière

Limitations des approches existantes

- Se focalisent sur le cas non contraint
- Utilisation du clustering basé sur la densité (DBSCAN et variantes)
  - pas de critère de qualité clair
  - sensibilité aux paramètres
  - problèmes en cas de densité hétérogène
- Clustering plat

Plan

1 Contexte et motivations

2 Approches proposées

3 Résultats

4 Conclusion

Modèle de données

- Idée clé : transposer le problème de clustering de données de trajectoires en un problème de clustering de graphes
- Considérer le graphe biparti $\mathcal{G} = (\mathcal{T}, \mathcal{S}, \mathcal{E})$
  - $\mathcal{T}$: nœuds représentant les trajectoires
  - $\mathcal{S}$: nœuds représentant les segments routiers
  - arête : passage d'une trajectoire par un segment

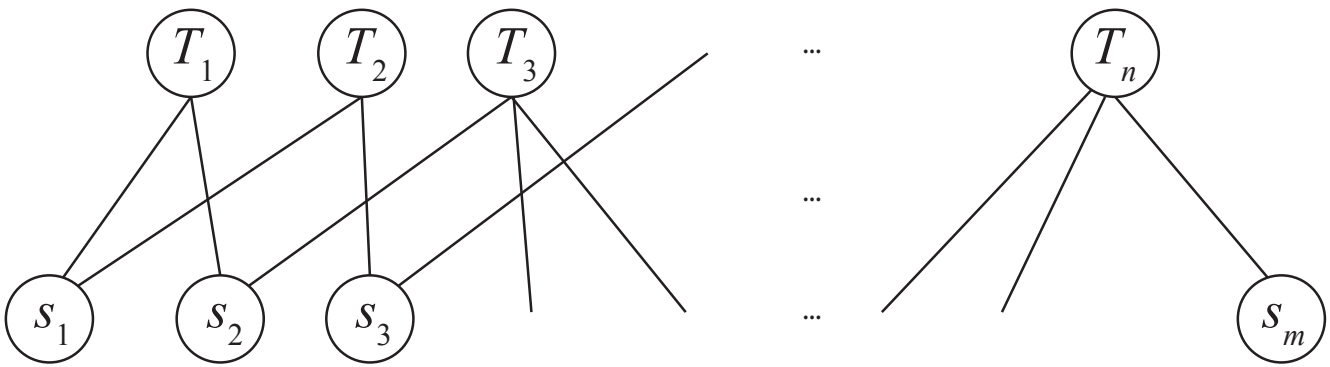
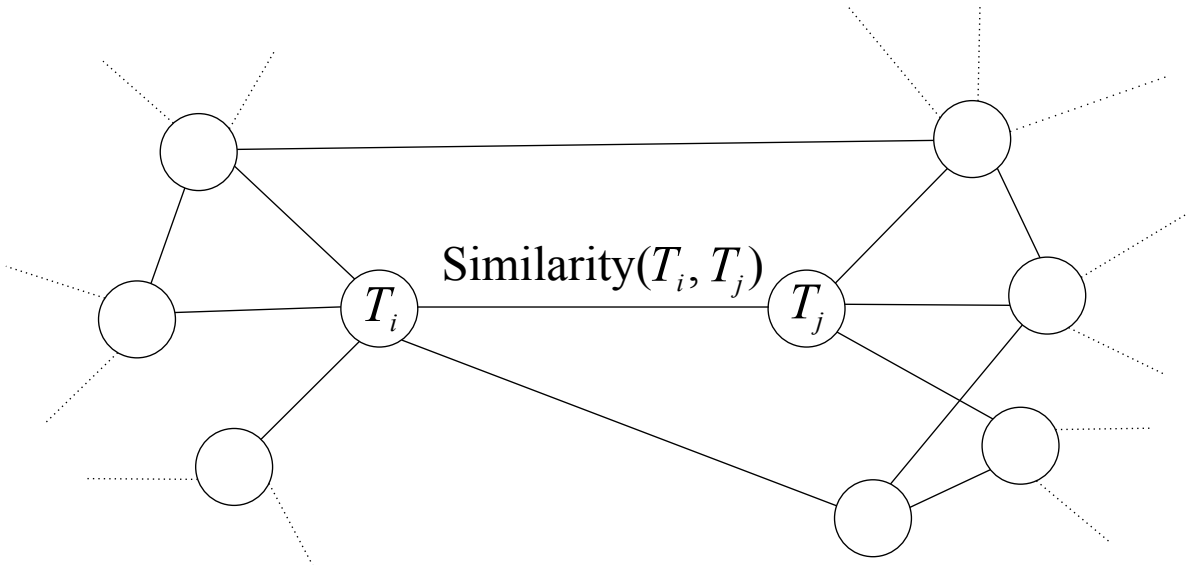


Figure: Graphe biparti entre trajectoires et segments

Projection sur les nœuds de trajectoires

- $\mathcal{G}_{\mathcal{T}} = (\mathcal{T}, \mathcal{E}_{\mathcal{T}}, \mathcal{W}_{\mathcal{T}})$: graphe non-orienté pondéré
  - Sommets : trajectoires
  - Arêtes : liens de similarité
  - Poids d'une arête : intensité de la similarité entre les deux trajectoires qu'elle relie



Similarité entre trajectoires

- Au lieu d'un comptage simple, une similarité cosinus avec tf-idf pondéré

$$\text{Similarity}(T_i, T_j) = \frac{\sum_{s \in \mathcal{S}} w_{s,T_i} \cdot w_{s,T_j}}{\sqrt{\sum_{s \in \mathcal{S}} w_{s,T_i}^2} \cdot \sqrt{\sum_{s \in \mathcal{S}} w_{s,T_j}^2}}$$

avec :

$$w_{s,T} = \frac{n_{s,T} \cdot \text{length}(s)}{\sum_{s' \in T} n_{s',T} \cdot \text{length}(s')} \cdot \log \frac{|\mathcal{T}|}{|\{T_i : s \in T_i\}|}$$

Projection sur les nœuds de segments

- Procéder par analogie au cas des trajectoires
- $\mathcal{G}_{\mathcal{S}} = (\mathcal{S}, \mathcal{E}_{\mathcal{S}}, \mathcal{W}_{\mathcal{S}})$: graphe de similarité entre segments routiers
  - $\text{Similarity}(s_i, s_j) = \frac{\sum_{T \in \mathcal{T}} w_{T,s_i} \cdot w_{T,s_j}}{\sqrt{\sum_{T \in \mathcal{T}} w_{T,s_i}^2} \cdot \sqrt{\sum_{T \in \mathcal{T}} w_{T,s_j}^2}}$
  - $w_{T,s} = \frac{n_{s,T}}{\sum_{T' \in \mathcal{T}} n_{s,T'}} \cdot \log \frac{|\mathcal{S}|}{|s \in \mathcal{S} : s \in T|}$

Clustering des graphes projetés

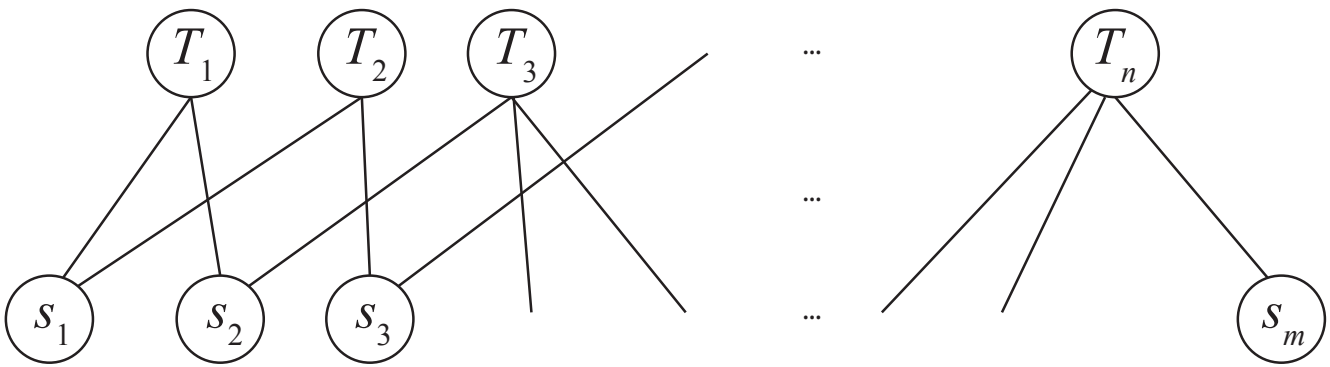
- Utilisation de l'algorithme de clustering hiérarchique par optimisation de la modularité décrit dans [Rossi and Villa-Vialaneix, 2011]
- Pour un graphe $G = (V, E, W)$ donné, trouve la partition en communautés $\mathcal{P} = \{C_1, C_2, ..., C_K\}$ qui maximise la modularité :

$$\mathcal{Q} = \frac{1}{2m} \sum_{k=1}^{K} \sum_{i,j \in C_k} \left(\omega_{ij} - \frac{d_i d_j}{2m} \right)$$

- Descendre récursivement sur le sous-graphe isolé correspondant à chaque communauté
- S'arrêter lorsque la partition détectée est non significative

Bi-clustering simultané trajectoires/segments

- Considérer directement le graphe biparti $\mathcal{G} = (\mathcal{T}, \mathcal{S}, \mathcal{E})$
- Appliquer un biclustering sur la matrice d'adjacence de $\mathcal{G}$
- Réarranger les lignes et les colonnes pour faire apparaitre des blocs de densité homogène



Approche MODL

- Utilisation de l'approche MODL [Boullé, 2011]
  - non-paramétrique
  - ne nécessite pas une pondération (une mesure de similarité)
  - capable de supporter le passage à l'échelle
  - repose sur un critère de Maximum A Posteriori

$$\mathcal{M}^* = \underset{\mathcal{M}}{\text{argmax}}\, P(\mathcal{M})P(\mathcal{D}|\mathcal{M})$$

Plan

1 Contexte et motivations

2 Approches proposées

3 Résultats

4 Conclusion

Données utilisées

- Jeu de données synthétiques
  - 85 trajectoires
  - réparties sur 5 clusters "naturels"
  - 485 segments routiers visités
- Générées en utilisant la carte d'Oldenburg
  - 6105 nœuds (intersections routières)
  - 14070 segments routiers

Données utilisées : clusters naturels

(a) 14 trajectoires (b) 19 trajectoires (c) 20 trajectoires

(d) 20 trajectoires (e) 12 trajectoires

Figure: Jeu de données utilisé

Clusters de trajectoires obtenus par projection

- 3 clusters seulement au plus haut niveau de l'hiérarchie
- certains clusters naturels regroupés en fonction de leur interaction

(a) (b) (c)

Clusters de trajectoires obtenus par projection

- tous les clusters naturels retrouvés au niveau suivant
- certains sont même "raffinés"

(d)

(e)

(f)

(g)

(h)

(i)

(j)

(k)

Clusters de segments obtenus par projection

- Moins interprétables que les clusters de trajectoires

(l)

(m)

(n)

(o)

(p)

(q)

Clusters de trajectoires obtenus par bi-clustering vs. obtenus par projection

- Résultats très similaires à ceux par projection

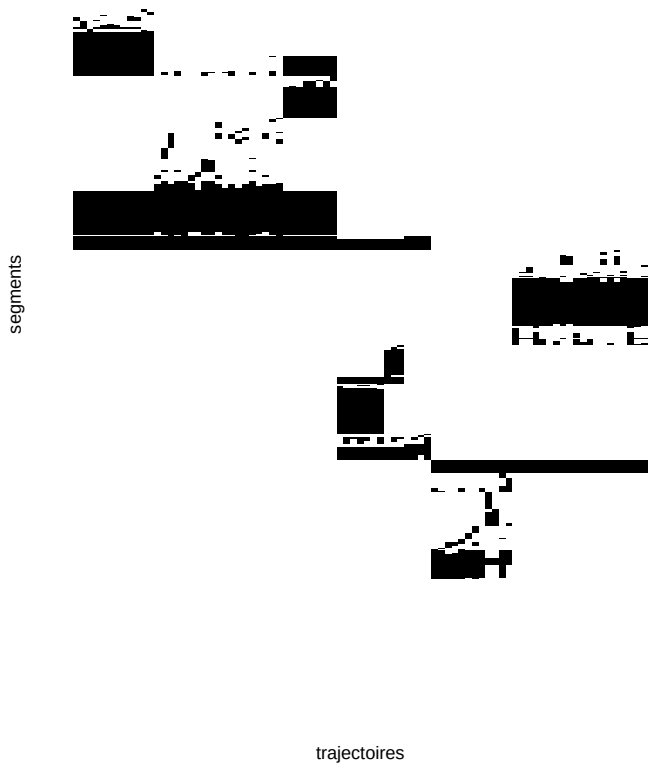
	1	2	3	4	5	6	7	8
1	0	0	0	7	3	4	0	0
2	0	19	0	0	0	0	0	0
3	12	0	8	0	0	0	0	0
4	0	0	0	0	0	0	0	20
5	0	0	0	0	0	0	12	0

Table: Matrice de confusion entre clusters naturels (sur les lignes) et ceux obtenus par optimisation de la modularité (sur les colonnes).

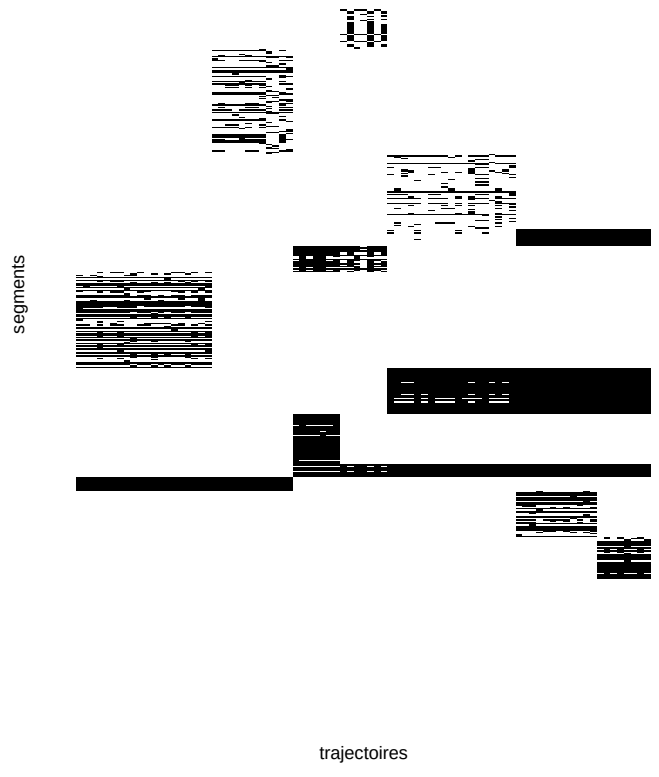
	1	2	3	4	5	6	7
1	0	0	7	7	0	0	0
2	0	0	0	0	19	0	0
3	0	0	0	0	0	12	8
4	20	0	0	0	0	0	0
5	0	12	0	0	0	0	0

Table: Matrice de confusion entre clusters naturels de trajectoires (sur les lignes) et clusters optimaux obtenus par biclustering (sur les colonnes).

Clusters croisés



(r) Par projection



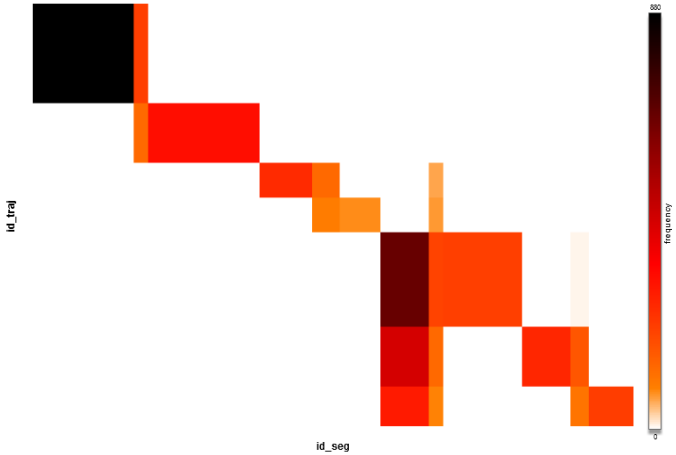
(s) Biclustering

Figure: Matrices d'adjacence des clusters croisés.

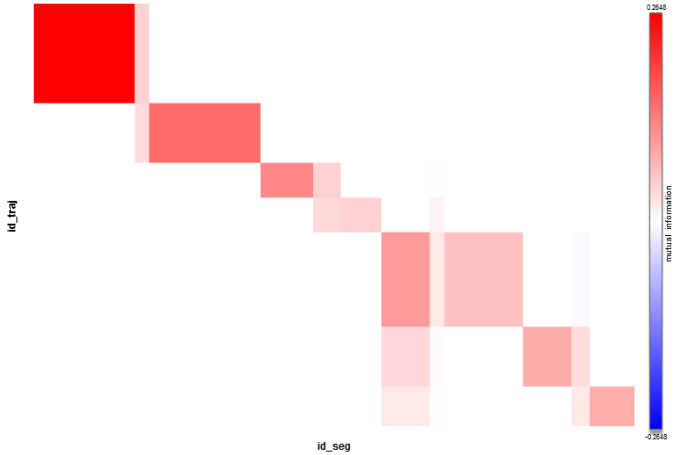
Utilisation de l'information mutuelle

- Information mutuelle

$$mi(c_S, c_T) = P(c_S, c_T) \log \frac{P(c_S, c_T)}{P(c_S)P(c_T)}$$



(a) Fréquences des biclusters de segments et trajectoires

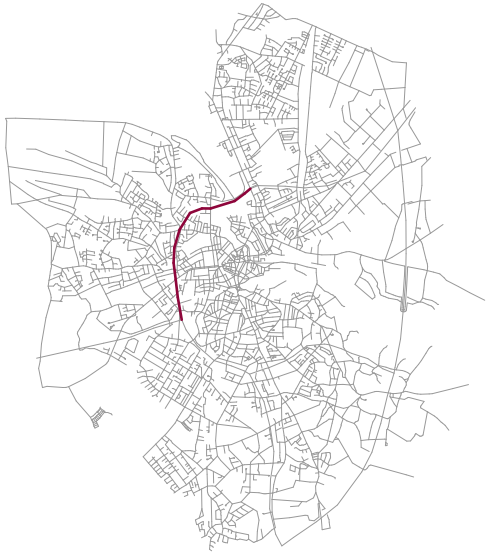


(b) Information mutuelle entre clusters de trajectoires et de segments

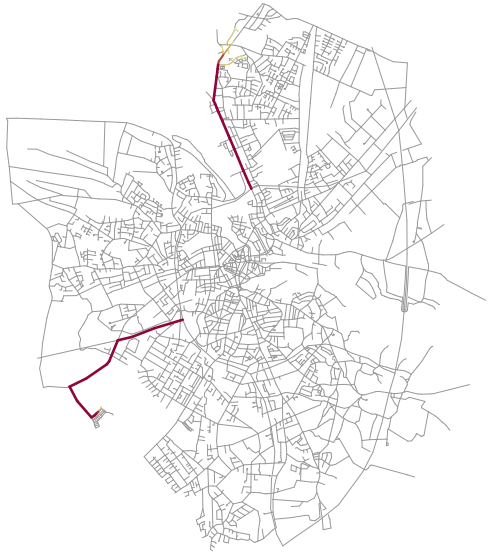
Figure: Fréquence et information mutuelle pour les biclusters découverts.

Utilisation de l'information mutuelle

- Permet de faciliter l'interprétation des clusters



(a) Hub autoroutier



(b) Axes secondaires menant vers la périphérie de la ville

Conclusion

- Intérêt des approches de type clustering de graphes pour résoudre le problème de clustering de trajectoires
- Avantages :
  - critères de qualité clairs
  - non-paramétriques
  - hiérarchies favorisant une exploration multi-niveaux
  - résultats interprétables
- Inconvénients :
  - complexités élevées
  - inutilisables dans un contexte "flux de données"
  - comportement face au bruit

Perspectives

- Cohérence des clusters en présence de bruit
- Autres approches de bi-clustering
- Tests sur données réelles

References I

Boullé, M. (2011).

Data grid models for preparation and modeling in supervised learning.
In *Hands-On Pattern Recognition : Challenges in Machine Learning, vol. 1*, pages 99--130. Microtome.

Rossi, F. and Villa-Vialaneix, N. (2011).

Représentation hiérarchique d'un grand réseau à partir d'une classification hiérarchique de ses sommets.
Journal de la Société Française de Statistique, 152 :34--65.