Segmentation géographique par étude d'un journal d'appels téléphoniques

Romain Guigourès [*,**], Marc Boullé [*], Fabrice Rossi [**]

[*]Orange Labs
2 avenue Pierre Marzin
22307 Lannion Cedex
{prenom.nom}@orange.com

[**]SAMM, Université Paris 1
70 rue Tolbiac
75013 Paris
{prenom.nom}@univ-paris1.fr

Résumé. Dans cet article, il est question de segmentation géographique par l'étude d'un journal d'appels aggrégés par ville. Au lieu de réaliser directement un clustering de nœuds, nous proposons ici de faire du coclustering sur les arcs, définis comme des instances bidimensionnelles décrites par deux variables : le nœud source et le nœud cible. Une fois la segmentation optimale obtenue, les clusters sont fusionnés successivement de manière à déteriorer le moins possible le modèle de clustering. Des expérimentations ont été menées sur un journal d'appel de l'opérateur de télécommunications Belge Mobistar.

1 Introduction

Avec la récente croissance des données de réseau tels que le web, les réseaux sociaux, les réseaux téléphoniques ou encore les graphes de collaborations scientifiques (Albert et Barabasi, 2002), on assiste à un regain d'intérêt pour les problèmes de partitionnement de graphes, plus particulièrement pour la découverte de communautés dans les gros graphes. De nombreuses approches ont été proposées telles que le clustering hiérarchique, les méthodes spectrales ou encore les marches aléatoires (Schaeffer, 2007). Pour évaluer la qualité d'un clustering de graphe, Girvan et Newman (2002) proposent un critère appelé modularité qui est aujourd'hui très largement utilisé dans la littérature, voir directement optimisé via des algorithmes de clustering (Blondel et al., 2008).

Dans cet article, nous proposons un moyen d'analyser et de résumer la structure de grands graphes, basé sur une estimation de densité jointe constante par morceaux. Les modèles en grille (Boullé, 2005) sont appliqués aux graphes, pour lesquels chaque arc est considéré comme une unité statistique, définie par deux variables, le nœud source et le nœud cible. Le but de la méthode est de réaliser un découpage en grilles dont les cellules, plus ou moins denses, représentent l'intersection des clusters de nœuds sources et cibles. En outre, cela revient à réaliser un coclustering. La meilleure grille est obtenue par la méthode MODL (Boullé, 2005) en utilisant des heuristiques combinatoires dont la complexité est linéaire et fonction du nombre d'arcs. Une fois la grille optimale obtenue, un post-traitement permettant de réduire le nombre de clusters est appliqué. Les clusters sont fusionnés de manière à dégrader le moins possible la grille optimale pour permettre une analyse exploratoire des données. Ce post-traitement peut

être assimilé à un clustering hiérarchique ascendant dont la mesure de dissimilarité correspond au coût de la fusion de deux clusters.

2 La segmentation

Clustering de graphe par la méthode MODL. Au lieu de réaliser un clustering de graphe simple comme les méthodes basées sur la modularité le font, le graphe sur lequel nous travaillons est biparti, orienté et avec des arcs multiples. Nous le traitons sous sa forme tabulaire, les nœuds du graphe représentant les villes et les arcs les appels. Deux villes seront regroupées si les distributions de leurs appels sont similaires. Cela signifie qu'au lieu de faire des clusters de villes qui s'appellent souvent, les villes seront dans un même cluster si elles appellent ou sont appelées par les mêmes villes et dans les mêmes proportions. L'objectif de la méthode est d'estimer la densité des arcs, ce qui conduit à faire un coclustering des villes émettrices et des villes réceptrices. La Figure 1 illustre le principe. Dans cet exemple, la probabilité des arcs de Bruxelles ou Liège vers Bruxelles ou Namur est de 50%.

	Brussels	Namur	Liège	Charleroi
Brussels	1	1	0	0
Liège	1	1	0	0
Charleroi	0	0	1	1
Namur	0	0	1	1

→

	{Brussels,Namur}	{Liège,Charleroi}
{Brussels,Liège}	4	0
{Charleroi,Namur}	0	4

FIG. 1: Exemple de coclustering

Un modèle M d'estimation de densité d'arc définit les propriétés de la segmentation : nombre de clusters, partitions des nœuds dans les clusters, distribution des arcs sur les clusters... Le modèle le plus grossier est composé d'un cluster de villes, le plus fin d'un cluster par ville. Un modèle grossier sera interprétable, tandis qu'un modèle fin sera informatif. L'idée de la méthode est de trouver le bon compromis entre informativité et interprétabilité. En appliquant une approche de sélection de modèle Bayésien, le meilleur modèle M^* est défini comme étant le plus probable connaissant les données D et est obtenu en maximisant un critère construit hiérarchiquement et uniformément à chaque étape, basé sur un terme de prior $P(M)$ favorisant les modèles simples et une vraisemblance $P(D|M)$ qui favorise les modèles informatifs.

$$M^* = \underset{M}{\operatorname{argmax}}\, P(M|D) = \underset{M}{\operatorname{argmax}} \left(\frac{P(M)P(D|M)}{P(D)} \right)$$

Le critère complet n'est pas détaillé ici pour des raisons de concision. Quant à l'algorithme d'optimisation, il est basé sur des heuristiques qui présentent de bonnes propriétés de scalabilité avec une complexité en espace de l'ordre de $O(m)$ et en temps de $\mathrm{O}(m\sqrt{m}\log m)$, avec m le nombre d'arcs. L'algorithme est étudié en détails dans (Boullé, 2011).

Fusion des clusters. Lorsque le nombre d'arcs du graphe est très élevé, la densité converge vers la vraie distribution des arcs. Ceci signifie que pour chaque ville, les distributions des appels sont suffisamment précises pour être différenciées les unes des autres. Ainsi la méthode construira un cluster par ville, ce qui est trop précis pour une interprétation raisonnable. Pour éviter ce problème, nous proposons de fusionner les clusters, de manière à dégrader le moins possible le critère optimisé précédemment, jusqu'à obtenir le nombre désiré de cluster. En étudiant en détail la variation du critère, il apparait qu'il s'agit de la somme des divergences de

Kullback-Leibler de chacun des clusters fusionnés par rapport au résultat de leur fusion. Ainsi, deux clusters seront d'autant plus sujets à fusion que les distributions de leurs appels seront similaires. Pour résumer, ce post-traitement est équivalent à une classification hiérarchique ascendante dont la mesure de dissimilarité est basée sur des divergences de probabilité.

3 Expérimentations

Des expérimentations ont été réalisées sur des journaux d'appels téléphoniques de l'opérateur de téléphonie Belge Mobistar. Le jeu de données compte 217 millions d'appels entre 589 villes. Une autre étude sur les mêmes données a été réalisée, conduisant à une segmentation en 17 clusters (Blondel et al., 2010). Comme dans cette étude, les clusters que nous obtenons sont connectés géographiquement. Cependant, notre étude permet une étude exploratoire des données à différentes échelles.

Le niveau le plus fin. Il conduit à segmentation des villes en 588 clusters, soit environ une ville par cluster. Cela signifie que la distribution des appels pour chaque ville est suffisamment fine pour être différenciée.

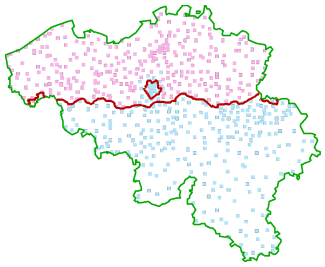

FIG. 2: Segmentation en 2 clusters.

Deux communautés linguistiques. Les 588 clusters ont été fusionnés successivement jusqu'à obtenir 2 clusters. Les deux clusters obtenus reflètent exactement les deux principales communautés linguistiques Belges : les Flandres et la Wallonie 2. Bruxelles se présente comme un îlot Wallon en territoire Flamand, ceci s'explique par la prédominance des Francophones dans la capitale.

Onze clusters qui ne correspondent pas aux Provinces Belges. Il y a 11 provinces Belges. Dans le but de comparer le tracé des bassins téléphoniques et les frontières des provinces, les clusters ont été fusionnés jusqu'à en obtenir onze. La Figure 3 présente les bassins téléphoniques et les frontières des provinces.

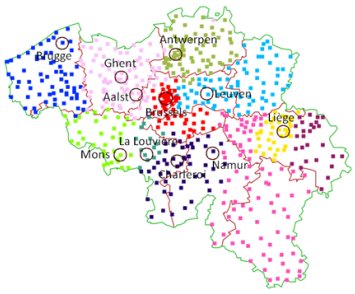

FIG. 3: Segmentation en 11 clusters

Pour 3 provinces flamandes, le tracé des bassins correspond à peu près aux frontières des provinces. Certaines d'entre elles sont divisées en 3 clusters. Pour le Hainault, il s'agit de la zone d'influence des 3 grosses villes de la province : Charleroi, Mons et La Louvière. Pour la province de Liège, on notera la zone d'influence de la capitale de province, ainsi qu'un cluster à l'est de la ville qui correspond à l'arrondissement de Verviers où 25% de la population est germanophone. Le cas de Bruxelles montre une corrélation entre les appels passés et la zone d'influence de la ville qui inclut une petite partie du Brabant Flamand et la quasi totalité du Brabant Wallon (région au sud-sud-est de la ville). Ceci peut être expliqué par l'attrait des Bruxellois pour les zones périphériques présentant les mêmes caractéristiques linguistiques, c'est-à-dire les zones méridionales francophones (Kesteloot et al., 2007).

Une étude locale de Bruxelles. Une étude des appels provenant des municipalités de la province de Bruxelles-Capitale vers l'ensemble de la Belgique permet de réaliser une segmentation locale à la capitale sans négliger les appels passés vers le reste du pays. Fusionner les clusters jusqu'à en obtenir trois permet de montrer une corrélation entre les clusters et la nature sociale des quartiers de la capitale Belge. En effet, sur la Figure 4, le cluster vert couvre les quartiers plutôt aisés de Bruxelles alors que le rose représente plutôt les quartiers populaires. Les deux villes du cluster orange sont d'importants centres étudiants.

FIG. 4: Segmentation de Bruxelles en 3 clusters

4 Conclusion

Dans cet article, nous avons proposé une méthode qui répond au problème du clustering de graphe appliqué aux données de téléphonie. En réalisant un coclustering sur les nœuds sources et les nœuds cibles, la méthode se comporte comme un estimateur non paramétrique de densité d'arcs. Dans le cas de grands graphes, le meilleur modèle a tendance à être trop fin pour être interprétable. C'est pourquoi un post-traitement consistant à fusionner pas à pas les clusters de manière à dégrader le moins possible le modèle optimal est proposé.

Des expérimentations sur un jeu de données de l'opérateur de téléphonie Belge Mobistar montrent la diversité des analyses possibles grâce à la méthode. D'une part, il est possible d'analyser la segmentation à plusieurs niveaux de grain mais également de le faire localement sur une ville ou une province. La méthode se basant sur une estimation de densité menant à un coclustering à deux variables, il est envisagé dans les prochains travaux d'ajouter une troisième variable continue, qui serait temporelle et permettrait d'une part d'étudier les graphes dynamiques mais également d'être discrétisée de manière optimale.

Références

Albert, R. et A.-L. Barabasi (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics 74*, 47.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks.

Blondel, V. D., G. Krings, et I. Thomas (2010). Regions and borders of mobile telephony in belgium and in the brussels metropolitan zone. *the e-journal for academic research on Brussels*.

Boullé, M. (2005). A bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research 6*, 1431–1452.

Boullé, M. (2011). Nonparametric edge density estimation in large graphs. Technical report, Orange Labs.

Girvan, M. et M. E. J. Newman (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences 99*(12), 7821–7826.

Kesteloot, C., C. Vandermotten, et B. Ippersiel (2007). Dynamic analysis of troubled neighbourhoods in the belgian urban regions.

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review 1*(1), 27 – 64.