Sur la normalisation de la matrice Laplacienne en partitionnement spectral

Julien Ah-Pine

Sur la normalisation de la matrice Laplacienne en partitionnement spectral

Julien Ah-Pine*

*Université de Lyon, Lyon 2, ERIC EA 3083
5 avenue Pierre Mendès France, 69500 Bron
julien.ah-pine@univ-lyon2.fr

Résumé. Les méthodes de partitionnement spectral permettent de mieux tenir compte de la géométrie intrinsèque des données contrairement aux approches classiques telle que les k-moyennes. Ceci est permis par l'analyse de la matrice Laplacienne du graphe de similarités. Il existe plusieurs types de matrice Laplacienne qui peut être normalisée ou pas. Dans cette contribution, nous introduisons une généralisation de la normalisation par division symétrique. Nous présentons les motivations et propriétés de notre approche et nous montrons les résultats obtenus sur plusieurs jeux de données.

1 Introduction

Le partitionnement spectral ("spectral clustering") modélise le problème de classification automatique par des graphes et utilise des résultats de la théorie spectrale des graphes. En effet, les problèmes de coupes de graphe reviennent à partitionner l'ensemble des sommets et sont donc connexes au problème de classification automatique. La représentation graphique permet alors de s'affranchir d'une représentation métrique des données qui serait nécessairement engendrée par un ensemble de variables. Ce qui importe en tout premier lieu, ce sont les valuations des relations de similarité entre sommets.

Nous rappelons en section 2 les concepts et la procédure de base en partitionnement spectral. En particulier, les matrices Laplaciennes du graphe jouent un rôle fondamental et représentent le sujet principal de cette contribution. En effet, nous présentons en section 3 une famille de matrices Laplaciennes basée sur une généralisation de la normalisation par division symétrique. Nous comparons ces nouvelles normalisations avec la méthode de base sur plusieurs jeux de données et nous présentons les résultats obtenus en section 4.

2 Partitionnement spectral

Soit $G = (\mathbb{V}, \mathbb{E})$ un graphe non-orienté, sans boucle, où $\mathbb{V} = \{v_1, \ldots, v_n\}$ est l'ensemble des sommets, et $\mathbb{E}$ l'ensemble des arêtes. $\mathbb{V}$ représente les éléments que nous souhaitons partitionner et $\mathbb{E} \subset \mathbb{V} \times \mathbb{V}$ est l'ensemble des paires de ces éléments qui sont similaires. Ces relations de similarité sont pondérées et si $(v_i, v_j) \in \mathbb{E}$, alors $l(v_i, v_j) > 0$ est la valuation de

la similarité [1] pour (v_i, v_j). Nous représentons G par une matrice d'adjacence pondérée notée $\mathbf{W} = (w_{ij})_{i,j=1,\ldots,n}$ et définie comme suit :

$$w_{ij} = \begin{cases} l(v_i, v_j) & \text{si } (v_i, v_j) \in \mathbb{E} \\ 0 & \text{sinon} \end{cases} \tag{1}$$

$\mathbf{W}$ possède des termes positifs ou nuls, elle est symétrique et contient des 0 sur sa diagonale. Pour chaque $v_i \in \mathbb{V}$, nous calculons son degré $d_i = \sum_{j=1}^{n} w_{ij}$. La matrice des degrés est alors la matrice diagonale notée $\mathbf{D} = (d_{ij})_{i,j=1,\ldots,n}$ et définie par :

$$d_{ij} = \begin{cases} d_i & \text{si } i = j \\ 0 & \text{sinon} \end{cases} \tag{2}$$

La matrice Laplacienne non-normalisée de G est une matrice carrée d'ordre n notée $\mathbf{L}$ avec :

$$\mathbf{L} = \mathbf{D} - \mathbf{W} \tag{3}$$

Plusieurs normalisation de $\mathbf{L}$ existent. La première est notée $\mathbf{L}_{rw}$ et est définie par :

$$\mathbf{L}_{rw} = \mathbf{D}^{-1}(\mathbf{D} - \mathbf{W}) = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W} \tag{4}$$

La normalisation par division symétrique est le second type de matrice Laplacienne normalisée. Elle est notée $\mathbf{L}_{sym}$ et est définie par :

$$\mathbf{L}_{sym} = \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} \tag{5}$$

Les matrices Laplaciennes ont la propriété d'être semi-définies positives et elles interviennent dans la résolution approchée de plusieurs problèmes de coupes minimales de graphe (Chung (1997); Von Luxburg (2007)). En partitionnement spectral, l'obtention d'une partition en k classes de $\mathbb{V}$, étant donné G, se fait selon la procédure suivante : (i) on représente les données dans l'espace engendré par les k premiers vecteurs propres de la matrice Laplacienne ; (ii) on applique une méthode classique de classification automatique, la méthode des k moyennes typiquement. L'approche qui utilise $\mathbf{L}_{rw}$ fût proposée par Shi et Malik (2000); Meila et Shi (2000) et celle qui emploie $\mathbf{L}_{sym}$ [2] par Ng et al. (2001).

3 Une famille de matrice Laplacienne normalisée

Les matrices Laplaciennes normalisées ont de meilleures propriétés que $\mathbf{L}$ (Von Luxburg et al. (2008)). Par ailleurs, les vecteur propres de $\mathbf{L}_{rw}$ sont liés à ceux de $\mathbf{L}_{sym}$. Par conséquent, dans la suite, nous étudions exclusivement $\mathbf{L}_{sym}$.

Nous proposons une généralisation de la normalisation décrite en (5). Dans cette perspective, nous introduisons $\mathbf{N} = (n_{ij})_{i,j=1,\ldots,n}$ la matrice carrée d'ordre n de terme général :

$$n_{ij} = \frac{1}{\sqrt{d_i d_j}} \tag{6}$$

1. En partitionnement spectral, le noyau Gaussien est typiquement utilisé.
2. Dans ce cas, les vecteurs propres sont multipliés par la racine carrée de leur valeur propre associée.

$\mathbf{N}$ possède des termes positifs ou nuls, elle est symétrique et les termes de sa diagonale sont $n_{ii} = 1/d_i, \forall i = 1, \ldots, n$.

Notre généralisation passe au préalable par une formulation équivalente de (5) faisant intervenir $\mathbf{N}$ et le produit de Hadamard [3], noté $\circ$, qui est le produit terme à terme entre deux matrices de même taille. Nous avons en effet la relation suivante :

$$\mathbf{L}_{sym} = \mathbf{L} \circ \mathbf{N} \tag{7}$$

Nous généralisons la matrice $\mathbf{N}$ par une famille paramétrique de matrices notées $\mathbf{N}^t = (n_{ij}^t)_{i,j=1,\ldots,n}$ qui sont définies pour tout réel t non nul comme suit :

$$n_{ij}^t = \frac{1}{\left(\frac{1}{2}(d_i^t + d_j^t)\right)^{\frac{1}{t}}} \tag{8}$$

n_{ij}^t est l'inverse de la moyenne généralisée [4] d'ordre t des degrés de v_i et de v_j. L'équation (6) est alors un cas particulier de (8) puisque $\mathbf{N} = \lim_{t \to 0} \mathbf{N}^t$.

La famille paramétrique de matrice Laplacienne normalisée que nous proposons est notée $\mathbf{L}_{sym}^t$ et est définie comme suit :

$$\mathbf{L}_{sym}^t = \mathbf{L} \circ \mathbf{N}^t \tag{9}$$

Nous avons le résultat suivant.

Théorème 1. *Pour tout $t > 0$, $\mathbf{L}_{sym}^t$ est semi-définie positive.*

Afin d'appréhender l'impact du paramètre t, nous utiliserons la forme quadratique associée à $\mathbf{L}_{sym}^t$. Soit $\mathbf{f} \in \mathbb{R}^n$, nous avons alors :

$$\mathbf{f}^\top \mathbf{L}_{sym}^t \mathbf{f} = \sum_{i=1}^{n} f_i^2 - \sum_{i,j=1}^{n} \frac{w_{ij}}{\left(\frac{1}{2}(d_i^t + d_j^t)\right)^{\frac{1}{t}}} f_i f_j \tag{10}$$

Comme il s'agit de minimiser $\mathbf{f}^\top \mathbf{L}_{sym}^t \mathbf{f}$, v_i et v_j auront une d'autant plus forte probabilité d'être dans la même classe que $w_{ij} / \left(\frac{1}{2}(d_i^t + d_j^t)\right)^{\frac{1}{t}}$ est grand. Comme pour tout $t > 0$, $\sqrt{d_i d_j} \leq \left(\frac{1}{2}(d_i^t + d_j^t)\right)^{\frac{1}{t}}$, et que l'égalité est atteinte si $d_i = d_j$, nous voyons que la moyenne généralisée d'ordre $t > 0$ pénalise la différence entre d_i et d_j (toute chose étant égale par ailleurs). Autrement dit, pour que v_i et v_j soient dans la même classe, non seulement w_{ij} doit être grand mais d_i et d_j doivent être proches également.

4 Validation expérimentale

Nous avons testé le partitionnement spectral avec différentes matrices Laplaciennes normalisées sur des données artificielles et réelles. Notre résultat de référence est celui obtenu avec $\mathbf{L}_{sym}$. Nous avons comparé celui-ci avec les performances des matrices $\mathbf{L}_{sym}^t$ pour $t = 1, 2, 5, 10$. Pour cela, nous avons utilisé un critère de validation externe : l'indice de Rand corrigé. Les jeux de données utilisés sont librement accessibles en ligne [5].

3. Également connu sous le nom de produit de Schur.

4. Également appelée, moyenne puissance ou moyenne de Hölder d'ordre t.

5. jain et flame à partir de `https://cs.joensuu.fi/sipu/datasets/`; yeast et pima à partir de `http://archive.ics.uci.edu/ml/datasets.html`; et smart (texte) à partir de `http://www.`

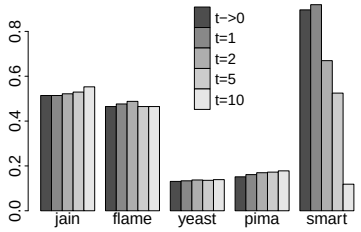

FIG. 1 – *Résultats sur 5 jeux de données et 5 normalisation différentes.*

Les résultats sont illustrés dans la figure 1. Nous constatons que la normalisation avec $t > 0$ permet d'obtenir des meilleurs résultats que la méthode de référence.

Références

Chung, F. R. (1997). *Spectral graph theory*, Volume 92. American Mathematical Soc.

Meila, M. et J. Shi (2000). Learning segmentation by random walks. In *NIPS*, Volume 14.

Ng, A. Y., M. I. Jordan, Y. Weiss, et al. (2001). On spectral clustering : Analysis and an algorithm. In *NIPS*, Volume 14, pp. 849–856.

Shi, J. et J. Malik (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence 22*(8), 888–905.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing 17*(4), 395–416.

Von Luxburg, U., M. Belkin, et O. Bousquet (2008). Consistency of spectral clustering. *The Annals of Statistics*, 555–586.

Summary

Spectral clustering has been a popular clustering technique for the last years. It better integrates the intrinsic geometry of the data as compared to the conventional k-means algorithm. In that respect, a core ingredient in spectral clustering is the graph Laplacian derived from the similarity matrix. However, there are different ways to employ the graph Laplacian: either it remains unnormalized or it can be normalized from different manners. In this paper, we introduce a generalization of the so-called symmetric normalization. We expose the motivations and properties of our approach and we show its performances on several benchmarks.

---

dataminingresearch.com/index.php/2010/09/classic3-classic4-datasets/