Sur l'apprentissage d'une matrice d'affinité bistochastique en clustering

Julien Ah-Pine

Sur l'apprentissage d'une matrice d'affinité bistochastique en clustering*

Julien Ah-Pine[1,2,3]

[1] Université de Lyon, Lyon 2, ERIC UR 3083
[2] Université Clermont Auvergne, LMBP CNRS, UMR 6620
[3] Université Clermont Auvergne, CERDI CNRS IRD, UMR 6587

julien.ah-pine@univ-lyon2.fr

Résumé

Nous nous intéressons à la tâche de clustering du point de vue graphe à l'instar du partionnement spectral (spectral clustering). Dans ce cas, la matrice d'affinité qui mesure l'intensité du lien (arête du graphe) pour chaque paire d'éléments (sommets du graphe) joue un rôle crucial. Plusieurs travaux antécédents ont montré l'intérêt de transformer une matrice d'affinité initiale de sorte à satisfaire certaines propriétés. La bistochasticité est une condition pertinente à cet égard. Dans ce travail, nous mettons en avant une autre condition : l'idempotence. Par la suite, En utilisant les propriétés existantes entre les matrices bistochastiques et idempotentes d'une part, et leurs matrices Laplaciennes associées d'autre part, nous proposons une nouvelle méthode d'apprentissage non-supervisé de matrice d'affinité. Notre procédure d'optimisation repose sur la méthode des multiplicateurs de Lagrange avec directions alternées (ADMM). Des résultats expérimentaux montrent l'intérêt pratique de notre approche.

Mots-clés

Clustering, Matrice d'affinité, Bistochasticité, Idempotence, ADMM.

Abstract

We are interested in graph based clustering such as spectral clustering. In this context, the affinity matrix that provides the strenght of the similarity between each pair of elements plays a crucial role. Several previous works have showed that transforming a given affinity matrix so that it becomes double stochastic was beneficial. In this work, we highlight another property : idempotency. By leveraging the relationships between double stochastic and idempotent matrices on the one hand, and their related Laplacian matrices on the other hand, we introduce a new unsupervised learning method for affinity matrices. Our learning algorithm is based on ADMM. Some experimental results are provided in order to demonstrate the interest of our proposal.

Keywords

Clustering, Affinity matrix, doubly stochasticity, Idempotence, ADMM.

*Cette communication est issue de l'article suivant : [1].

1 Contexte et travaux antérieurs

La tâche de clustering consiste à partitionner un ensemble d'éléments en des sous-ensembles homogènes appelés clusters. Soit un ensemble de n vecteurs $\{\mathbf{x}_i\}_{i=1,\ldots,n}$ appartenant à $\mathbb{R}^p$, que nous cherchons à analyser. Nous nous intéressons à la partition en k clusters $C = \{C_1, \ldots, C_k\}$ qui minimise le critère SSE (Sum of Squared Errors) suivant :

$$\sum_{j=1}^{k} \sum_{\mathbf{x}_i \in C_j} \|\phi(\mathbf{x}_i) - \mathbf{c}_j\|^2 \tag{1}$$

où $\phi : \mathbb{R}^p \to \mathbb{F}$ est une projection des $\{\mathbf{x}_i\}$ dans un espace de grande dimension $\mathbb{F}$, $\mathbf{c}_j = \sum_{\mathbf{x}_i \in C_j} \phi(\mathbf{x}_i)/n_j$ est le vecteur moyen du cluster C_j qui est de cardinal n_j et $\|.\|$ est la norme Euclidienne dans $\mathbb{F}$.
Le critère SSE peut être formalisé à l'aide de la matrice de noyau $\mathbf{K}$ de terme général $\mathbf{K}_{ii'} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_{i'}) \rangle = \kappa(\mathbf{x}_i, \mathbf{x}_{i'})$ où $\kappa : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ est une fonction noyau. Il s'agit dans ce cas de déterminer $\mathbf{X}$, la matrice de $\mathbb{R}^{n \times n}$ qui minimise la fonction objectif suivante :

$$\mathrm{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{X})) \tag{2}$$

où Tr est l'application trace dans $\mathbb{R}^{n \times n}$, $\mathbf{I}_n$ est la matrice identité d'ordre n, et $\mathbf{X}$ est de terme général :

$$\mathbf{X}_{ii'} = \begin{cases} 1/n_j & \text{si } \mathbf{x}_i \text{ et } \mathbf{x}_{i'} \text{ sont dans } C_j, \\ 0 & \text{sinon.} \end{cases} \tag{3}$$

La matrice $\mathbf{X}$ ainsi définie possède plusieurs propriétés. Plus précisément [6] montre que la minimisation du SSE peut s'exprimer de façon équivalente comme suit :

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times n}} \mathrm{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{X})) \tag{4}$$
$$\text{s.t. } \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n, \mathbf{X} = \mathbf{X}^2, \mathrm{Tr}(\mathbf{X}) = k.$$

où $\mathbf{0}_n$ est la matrice nulle d'ordre n, $(\mathbf{X} \geq \mathbf{0}_n) \Leftrightarrow (\mathbf{X}_{ii'} \geq 0, \forall i, i' = 1, \ldots, n)$, $\mathbf{X}^\top$ est la matrice transposée de $\mathbf{X}$ et $\mathbf{e}_n$ est le vecteur rempli de 1 de dimension n.

La matrice $\mathbf{X}$ recherchée est ainsi non-négative, symétrique, bistochastique (les sommes de chaque ligne et de chaque colonne valent 1), idempotente et de trace égale à k le nombre de clusters désiré. En fait, il existe une bijection entre l'ensemble des partitions d'un ensemble de n éléments et l'ensemble des classes d'équivalence des matrices bistochastiques et idempotentes pour la relation $\mathbf{X} \sim \mathbf{Y}$ si et seulement si il existe une matrice de permutation $\mathbf{P}$ telle que $\mathbf{X} = \mathbf{PYP}^\top$ [7].
Par exemple la partition $\{(a, e), (b, c, d)\}$ peut être représentée par les matrices bistochastiques et idempotentes suivantes appartenant à la même classe d'équivalence :

$$\begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \end{pmatrix} \end{array} \sim \begin{array}{c} \\ a \\ e \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{ccccc} a & e & b & c & d \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \end{pmatrix} \end{array}$$

Le problème ainsi formulé, donne un point de vue graphe à la tâche de clustering : $\mathbf{K}$ (à condition d'être non-négative -comme pour le noyau Gaussien par exemple-) peut être vue telle une matrice d'adjacence pondérée d'un graphe sans structure particulière et $\mathbf{X}$ peut être interprétée comme la matrice d'adjacence pondérée d'un graphe représentant une partition des sommets. Il s'agit alors d'approximer $\mathbf{K}$ par $\mathbf{X}$ au sens de la norme de Frobenius. En effet, soit $\mathrm{Tr}(\mathbf{X}^\top \mathbf{Y}) = \langle \mathbf{X}, \mathbf{Y} \rangle_F$ le produit scalaire de Frobenius dans $\mathbb{R}^{n \times n}$. Si $\mathbf{X}$ vérifie les contraintes stipulées dans (4), alors il est facile de montrer que :

$$\min_{\mathbf{X}} \mathrm{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{X})) \Leftrightarrow \min_{\mathbf{X}} \|\mathbf{K} - \mathbf{X}\|_F^2 \qquad (5)$$

Le modèle d'optimisation 4, appelé 0-1 Semi-Definite Program dans [6], est NP-difficile en raison de la nature discrète et de l'idempotence de la matrice $\mathbf{X}$. En pratique, une démarche heuristique permettant de résoudre de façon approchée 4 consiste à (i) définir un problème relaxé solutionnable en temps polynomial, (ii) discrétiser la solution optimale du problème relaxé afin d'obtenir une solution réalisable du problème initial.
De nombreuses méthodes d'apprentissage de matrice d'affinité et de clustering découlent de cette démarche. Dans ce travail nous nous penchons plus particulièrement, sur les approches présentées dans [10] et [11]. Ces travaux reviennent à remplacer $\mathbf{K}$ par une matrice $\mathbf{X}$ non-négative, symétrique et bistochastique (étape (i)). Autrement dit, la contrainte d'idempotence est abandonnée, le nombre de cluster k n'est alors plus associé à la trace et les contraintes restantes sont toutes linéaires. Des procédures efficaces sont proposées pour déterminer $\mathbf{X}$: dans [10] il s'agit d'une version symétrique de l'algorithme de Sinkhorn-Knoop [7] dénoté SSK, alors que dans [11] est introduit l'algorithme DSN (Double Stochastic Normalization). Plus précisément, l'algorithme de Sinkhorn-Knoop vise à minimiser la divergence de Kullback-Leibler entre $\mathbf{K}$ et $\mathbf{X}$ alors que l'approche DSN résulte de la minimisation de la distance de Frobenius entre $\mathbf{K}$ et $\mathbf{X}$. Une fois $\mathbf{X}$ déterminée une méthode de discrétisation est utilisée. Le spectral clustering qui, en bref, applique l'algorithme des k-means sur les k premiers vecteurs propres de $\mathbf{X}$ est une approche classique à cet égard (étape (ii)).

2 Approche proposée

Contrairement aux deux approches précédentes, nous cherchons à tenir compte de la contrainte d'idempotence tout en évitant de se ramener à un problème NP-difficile. Néanmoins, nous ne considérons pas le nombre de clusters k comme paramètre de notre modèle et n'imposons donc pas $\mathrm{Tr}(\mathbf{X}) = k$. L'approximation basée sur la distance de Frobenius reste centrale dans notre approche qui vise, en somme, à définir un problème relaxé du modèle suivant :

$$\begin{aligned} &\min_{\mathbf{X} \in \mathbb{R}^{n \times n}} \|\mathbf{K} - \mathbf{X}\|_F^2 \qquad (6) \\ &\text{s.t. } \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n, \mathbf{X} = \mathbf{X}^2. \end{aligned}$$

$\mathbf{X}$ étant bistochastique, la matrice des degrés vaut $\mathbf{I}_n$ et la matrice Laplacienne associée à $\mathbf{X}$ est donnée par $\mathbf{L}_\mathbf{X} = \mathbf{I}_n - \mathbf{X}$. Clairement, le problème (6) peut être reformulé de façon équivalente en fonction de $\mathbf{L}_\mathbf{X}$ comme suit :

$$\begin{aligned} &\min_{\mathbf{L}_\mathbf{X} \in \mathbb{R}^{n \times n}} \|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_\mathbf{X}\|_F^2 \qquad (7) \\ &\text{s.t. } \mathbf{L}_\mathbf{X} \leq \mathbf{I}_n, \mathbf{L}_\mathbf{X} = \mathbf{L}_\mathbf{X}^\top, \mathbf{L}_\mathbf{X}\mathbf{e}_n = \mathbf{n}_n, \mathbf{L}_\mathbf{X} = \mathbf{L}_\mathbf{X}^2. \end{aligned}$$

où $\mathbf{n}_n$ est le vecteur nul de dimension n.
Nous exploitons à présent les relations algébriques existantes entre $\mathbf{X}$ et $\mathbf{L}_\mathbf{X}$. En effet, ces deux matrices étant symétriques et idempotentes, elles représentent des projections orthogonales. De façon plus singulière, l'une est l'unique projecteur orthogonale complémentaire de l'autre et *vice-versa* : l'image de $\mathbf{X}$ est le noyau de $\mathbf{L}_\mathbf{X}$, l'image de $\mathbf{L}_\mathbf{X}$ est le noyau de $\mathbf{X}$ et nous avons la relation suivante, centrale dans notre travail :

$$\mathbf{X}\mathbf{L}_\mathbf{X} = \mathbf{0}_n \qquad (8)$$

Nous proposons de relaxer (6) et (7) en considérant le modèle suivant :

$$\begin{aligned} &\min_{\mathbf{X}, \mathbf{L}_\mathbf{X} \in \mathbb{R}^{n \times n}} \frac{1}{2}\|\mathbf{K} - \mathbf{X}\|_F^2 + \frac{1}{2}\|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_\mathbf{X}\|_F^2 + \frac{\mu}{2}\|\mathbf{X}\mathbf{L}_\mathbf{X}\|_F^2 \\ &\text{s.t. } \begin{cases} \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n, \\ \mathbf{L}_\mathbf{X} \leq \mathbf{I}_n, \mathbf{L}_\mathbf{X} = \mathbf{L}_\mathbf{X}^\top, \mathbf{L}_\mathbf{X}\mathbf{e}_n = \mathbf{n}_n, \\ \mathbf{X} + \mathbf{L}_\mathbf{X} = \mathbf{I}_n. \end{cases} \qquad (9) \end{aligned}$$

où $\mu > 0$ est un paramètre de pénalité.
Notre modèle intitulé DSNI (Doubly Stochastic and Nearly Idempotent), consiste en un apprentissage joint de $\mathbf{X}$ et de sa matrice Laplacienne associée $\mathbf{L}_\mathbf{X}$. Il est facile de montrer que sous la condition $\mathbf{X} + \mathbf{L}_\mathbf{X} = \mathbf{I}_n$, les trois propriétés qui suivent sont équivalentes : $\mathbf{X} = \mathbf{X}^2$, $\mathbf{L}_\mathbf{X} = \mathbf{L}_\mathbf{X}^2$, $\mathbf{X}\mathbf{L}_\mathbf{X} = \mathbf{0}_n$. Cependant, ces propriétés étant la source première de la complexité des problèmes (6) et (7), nous ne les intégrons pas dans les contraintes de notre modèle. Pour pallier à ce manque, nous ajoutons, en revanche, un terme de pénalité dans la fonction objectif, $\|\mathbf{X}\mathbf{L}_\mathbf{X}\|_F^2$, afin d'encourager les solutions obtenues à être ainsi quasi-idempotentes.

Le problème DSNI (9) étant bi-convexe, nous pouvons utiliser la méthode ADMM (voir par exemple [3]) comme procédure d'optimisation. Les différentes étapes sont alors les suivantes :

0. Initialisation : $\mathbf{X}^0 \leftarrow \mathbf{K}$ (en ayant au préalable annuler les valeurs négatives de $\mathbf{K}$ le cas échéant).
1. Déterminer $\mathbf{L}_{\mathbf{X}}^{t+1}$ avec $\mathbf{X}^t$ fixé :

$$\begin{aligned}\mathbf{L}_{\mathbf{X}}^{t+1} \leftarrow \underset{\mathbf{L}_{\mathbf{X}} \in \mathbb{R}^{n\times n}}{\arg\min}\ & \frac{1}{2}\|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_{\mathbf{X}}\|_F^2 \\ & + \frac{\mu}{2}\|\mathbf{X}^t\mathbf{L}_{\mathbf{X}}\|_F^2 \\ & + \frac{\rho}{2}\|\mathbf{X}^t + \mathbf{L}_{\mathbf{X}} - \mathbf{I}_n + \mathbf{U}^t\|_F^2 \\ & \text{s.t. } \mathbf{L}_{\mathbf{X}} \leq \mathbf{I}_n, \mathbf{L}_{\mathbf{X}} = \mathbf{L}_{\mathbf{X}}^\top, \mathbf{L}_{\mathbf{X}}\mathbf{e}_n = \mathbf{n}_n.\end{aligned} \tag{10}$$

2. Déterminer $\mathbf{X}^{t+1}$ avec $\mathbf{L}_{\mathbf{X}}^{t+1}$ fixé :

$$\begin{aligned}\mathbf{X}^{t+1} \leftarrow \underset{\mathbf{X}\in \mathbb{R}^{n\times n}}{\arg\min}\ & \frac{1}{2}\|\mathbf{K} - \mathbf{X}\|_F^2 \\ & + \frac{\mu}{2}\|\mathbf{X}\mathbf{L}_{\mathbf{X}}^{t+1}\|_F^2 \\ & + \frac{\rho}{2}\|\mathbf{X} + \mathbf{L}_{\mathbf{X}}^{t+1} - \mathbf{I}_n + \mathbf{U}^t\|_F^2 \\ & \text{s.t. } \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n.\end{aligned} \tag{11}$$

3. Déterminer $\mathbf{U}^{t+1}$:

$$\mathbf{U}^{t+1} \leftarrow \mathbf{U}^t + \mathbf{X}^{t+1} + \mathbf{L}_{\mathbf{X}}^{t+1} - \mathbf{I}_n \tag{12}$$

4. Répéter 1., 2., 3. tant qu'une condition d'arrêt n'est pas satisfaite.

Les sous-problèmes (10) et (11) peuvent être résolus efficacement par projections successives sur des ensembles convexes (voir par exemple [2]).

3 Validation empirique de l'approche proposée

Nous avons testé notre approche sur plusieurs jeux de données réels classiques disponibles en ligne [1]. Le protocole expérimental est le suivant : calcul de $\mathbf{K}$ en utilisant un noyau Gaussien ; approximation de $\mathbf{K}$ par une matrice d'affinité bistochastique $\mathbf{X}$ obtenue par SSK ou DSN ou DSNI (sauf pour la baseline) ; application du spectral clustering [4] sur $\mathbf{X}$ en fixant k au nombre correct de clusters ; comparaison de la partition obtenue et de la vérité terrain en utilisant la mesure NMI (Normalized Mutual Information). Pour le noyau Gaussien, l'hyperparamètre σ^2 est fixé à p et pour DSNI le paramètre de pénalité μ est fixé à $\sqrt{n}$.
Les résultats obtenus sont donnés dans la Table 1. La colonne SC représente la baseline et utilise le spectral clustering directement sur $\mathbf{K}$ la matrice de noyau Gaussien. Sur l'ensemble des jeux de données, nous constatons que DSNI donne de meilleurs résultats que SSK et DSN ce qui valide l'intérêt de notre modèle.

1. https://archive.ics.uci.edu/ml/index.php

Dataset	n	p	k	SC	SSK	DSN	DSNI
Glass	214	9	6	0.253	0.276	0.243	**0.297**
Ionosphere	351	34	2	0.038	0.066	0.076	**0.131**
Breast cancer	569	30	2	0.010	0.010	0.010	**0.670**
Yeast	1484	8	10	0.070	0.258	0.256	**0.263**
Digits	1797	64	10	0.015	0.044	0.743	**0.767**

TABLE 1 – Statistiques des jeux de données et mesures NMI des différentes méthodes.

Références

[1] J. Ah-Pine. Learning doubly stochastic and nearly idempotent affinity matrix for graph-based clustering. *European Journal of Operational Research*, 299(3) :1069–1078, 2022.

[2] H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM review*, 38(3) :367–426, 1996.

[3] S. Boyd, N. Parikh, and E. Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[4] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering : Analysis and an algorithm. *Advances in neural information processing systems*, 2 :849–856, 2002.

[5] J. Peng and Y. Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM journal on optimization*, 18(1) :186–205, 2007.

[6] J. Peng and Y. Xia. A new theoretical framework for k-means-type clustering. In *Foundations and advances in data mining*, pages 79–96. Springer, 2005.

[7] R. Sinkhorn. Two results concerning doubly stochastic matrices. *The American Mathematical Monthly*, 75(6) :632–634, 1968.

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

[9] F. Wang, P. Li, and A. C. Konig. Learning a bistochastic data similarity matrix. In *2010 IEEE International Conference on Data Mining*, pages 551–560. IEEE, 2010.

[10] R. Zass and A. Shashua. A unifying approach to hard and probabilistic clustering. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 294–301. IEEE, 2005.

[11] R. Zass and A. Shashua. Doubly stochastic normalization for spectral clustering. In *Advances in neural information processing systems*, pages 1569–1576, 2007.