Synthèse d'article
On Spectral Clustering : Analysis and an algorithm

Ali Ghezal
Majeur SIR / Master MFA

10 mars 2016

1 Introduction

Le problème de classification est sujet de grand travaux de recherche dans les domaines d'apprentissage statistique et de reconnaissance de formes. Pour le cas de données dans $\mathbb{R}^n$, L'approche standard basée sur les models génératifs cherche à apprendre une densité mixte de la distribution. Cependant, elle peut présenter certains défauts puisqu'elle part d'hypothèses fortes. La possibilité de convergence vers des minimums locaux rend ces algorithmes sensibles aux conditions intiales et nécessite plusieurs exécutions avant de converger vers la bonne solution.
Une alternative de cette approche consiste en l'utilisation de méthodes spectrales. Ces méthodes cherchent à classifier les points en utilisant les vecteurs propres de matrices construites à partir des données.
Cet article étudie un algorithme de classification spectrale en se basant sur la théorie de perturbation des matrices. Il présente les conditions nécessaires assurant des bons résultats et propose une implémentation simple de l'algorithme.

2 Analyse de l'algorithme

2.1 Position du problème

On se donne un ensemble de points $S = \{s_1, ..., s_n\}$ dans $\mathbb{R}^l$. On va cherche à les séparer en k classes. On définit alors les paramètres de notre problème tels qui suit :

La matrice $A \in \mathbb{R}^{n \times n}$ définit par $A_{ij} = \exp(-\|s_i - s_j\|/2\sigma^2)$ si $i \neq j$ et $A_{ii} = 0$.

La matrice diagonale D définit par $D_{ii} = \sum_{j=1}^{n} A_{ij}$

On déduit alors la matrice L définit par $L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$

Soient $x_1, x_2, ..., x_k$ les k plus grands vecteurs propres de L. On construit la matrice $X = [x_1 x_2 ... x_k]$ $\in \mathbb{R}^{n \times k}$

Finalement, la matrice Y est déduite de X après normalisation des lignes de cette dernières.
$Y_{ij} = X_{ij}/(\sum_j X_{ij}^2)^{1/2})$

Dans ce qui suit, on va montrer que l'ensemble de données S peut être séparé en plusieurs clusters dont les vecteurs prototypes sont les lignes de la matrice Y. On va tout d'abord considérer le cas "idéal" où les points qui forment les clusters sont bien distincts avant de présenter le cas général.

2.2 Le cas "idéal

Cette partie ne présente pas une vraie preuve de l'algorithme mais va nous permettre de mieux le comprendre en l'appliquant à un cas "idéal", où les clusters sont bien éloignés.

L'objectif est donc de montrer la proposition suivante :

Proposition 1 : Soit $\widehat{A}$ une matrice par blocks dont les blocks hors diagonale $\widehat{A}^{(ij)}$, $i \neq j$, sont nuls. On suppose aussi que chaque cluster S_i est connecté (i.e $\widehat{A}^{(ii)}_{jk} > 0, j \neq k$).
Alors il existe k vecteurs orthogonaux $r_1, r_2, ..., r_k$ ($r_i^T r_j = 1$ si $i = j$ et 0 sinon) tels que les lignes de $\widehat{Y}$ vérifient :

$$\widehat{y_j}^{(i)} = r_i \tag{1}$$

$\forall i = 1, ..., k, j = 1, ..., n_i$.

Pour ce faire, on prend $k = 3$ et on note S_1, S_2, S_3 les clusters de cardinaux respectifs n_1, n_2, n_3. On suppose aussi que les points s_i sont ordonnées selon leur appartenance aux clusters, cad les n_1 premiers points correspondent au cluster S_1 et ainsi de suite. L'hypothèse que les clusters sont bien éloignés se traduit par l'affectation de la valeur 0 aux éléments A_{ij} où s_i et s_j n'appartiennent pas au même cluster. A partir de ces hypothèses, on définit les matrices $\widehat{A}, \widehat{D}, \widehat{L}, \widehat{X} et \widehat{Y}$ déduites respectivement de A, D, L, X et Y.

On note que dans ce cas, $\widehat{A}$ et $\widehat{L}$ sont diaonales par blocks.

On définit aussi $\widehat{d} \in \mathbb{R}^n$ le vecteur contenant les éléments de la diagonale de la matrice $\widehat{D}$.

La matrice $\widehat{X}$ est construite à partir des trois premiers vecteurs propres de $\widehat{L}$. Or comme cette dernière est diagonale, ses valeurs propres et vecteurs propres sont l'union de ceux des matrices en blocks.
Il parait évident que $\widehat{L}^{(ii)}$ possède une valeur propre égale à 1 avec un vecteur propre associé positif $x_1^{(i)} \in \mathbb{R}^{n_i}$. Comme $\widehat{A}^{(ii)}_{jk} > 0$ ($j \neq k$), la vaeur propre suivante est inférieure à 1. On obtient alors $\widehat{X} \in \mathbb{R}^{n \times 3}$ sous la forme :

$$\begin{bmatrix} x_1^{(1)} & \vec{0} & \vec{0} \\ \vec{0} & x_1^{(2)} & \vec{0} \\ \vec{0} & \vec{0} & x_1^{(3)} \end{bmatrix}$$

Il est important ici de signaler que 1 est valeur propre multiple de $\widehat{L}$ et que dans ce cas, on peut considérer n'importe quels vecteurs orthogonaux recouvrant le même espace que les colonnes de $\widehat{X}$ pour la construction de $\widehat{L}$. Ainsi, $\widehat{X}$ pourrait être remplacée par $\widehat{X}R$ pour n'importe quelle matrice orthogonale $\widehat{R} \in \mathbb{R}^{n \times n}$ ($R^T R = R R^T = I$). D'un autre côté, même si la construction de X à partir des vecteurs propres de L est faite à une rotation prêt et peut être altérée par des perturbation dues à l'implémentation de l'algorithme, on peut espérer que ce denier va garantir une stabilité des sous-espaces couverts par les colonnes de $\widehat{X}$. Ainsi, après normalisation des lignes de $\widehat{X}$, on obtient :

$$\widehat{Y} = \begin{bmatrix} \widehat{Y}^{(1)} \\ \widehat{Y}^{(2)} \\ \widehat{Y}^{(3)} \end{bmatrix} = \begin{bmatrix} \vec{1} & \vec{0} & \vec{0} \\ \vec{0} & \vec{1} & \vec{0} \\ \vec{0} & \vec{0} & \vec{1} \end{bmatrix} R \tag{2}$$

où $\widehat{Y}^{(i)}$ représente le i-ème block de $\widehat{Y}$. Ainsi, on obtient que la j-ème ligne de $\widehat{Y}^{(i)}$ est égale i-ème ligne de la matrice orthogonale R. Ce qui prouve la proposition 1.

2.3 Le cas général

Dans ce cas, les blocks hors diagonale de A sont non nuls. En posant $E = A - \widehat{A}$, on peut alors exprimer A come somme d'une matrice idéale et d'une perturbation $A = \widehat{A} + E$. On cherche alors à exprimer les conditions qui donnent un résultat similaire à celui présenté dans la section précédante, à savoir des vecteurs de L proches de ceux de $\widehat{L}$ idéale.

La théorie de perturbation des matrice (Matrix perturbation theory) indique que la stabilité des vecteurs propres d'une matrice est déterminée par le "eigengap" défini par $\delta = \|\lambda_3 - \lambda_4\|$, la différance entre les 3^{eme} et la 4^{eme} valeurs propres. Ainsi les trois premiers vecteurs propres de $\widehat{L}$ vont être stables par rapport à des petites perturbations si et seulement si le "eigengap" est grand.
Comme signalé dans la section précédante, les vecteurs propres de $\widehat{L}$ sont l'union de ceux de $\widehat{L}^{(11)}$, $\widehat{L}^{(22)}$ et $\widehat{L}^{(33)}$ avec $\lambda_3 = 1$. On posant $\lambda_j^{(i)}$ la j^{me} valeur propre de $\widehat{L}^{(ii)}$, on a alors que $\lambda_4 = max_i \lambda_2^{(i)}$. L'hypothèse précédante revient alors à supposer $max_i \lambda_2^{(i)}$ est loin de 1.

On rappelle que l'objectif de cette partie est de trouver une version de la proposition 1 adapté au cas général. Afin de pouvoir énoncer le théorème, on va tout d'abord présenter les hypothèses considérées.

Hypothèse A1 : Il existe $\delta > 0$ tel que $\forall i = 1, ..., k$, $\lambda_2^{(i)} \leq 1 - \delta$.

Dans le contexte de clustering (classification), cette hypothèse souligne le fait que pour pouvoir distinguer trois clusters S_1, S_2 et S_3 il faut que chacun d'eux soit "compact".
Cette connexion entre le "eigengap" et la cohérance des clusters peut être exprimée de différentes façon, par exemple grâce à la *constante de Cheeger* exprimée par l'hypothèse suivante.

Hypothèse A1.1 : On définit la *constante de Cheeger* pour un cluster S_i par

$$h(S_i) = min_I \frac{\sum_{j \in I, k \notin I} A_{jk}^{(ii)}}{min \sum_{j \in I} \widehat{d}_j^{(i)}, \sum_{k \notin I} \widehat{d}_k^{(i)}} \tag{3}$$

où $I \subseteq 1, ..., n_i$. On suppose qu'il existe $\delta > 0$ tel que $(h(S_i))^2/2 \geq \delta \ \forall i$.

Un résultat standard de la théorie spectrale des graphes montre que l'hypothèse 1.1 implique l'hypothèse 1. $\widehat{d}_j^{(i)} = \sum_k A_{jk}^{(ii)}$ caractérise le degré de similarité entre le point j et les autres points appartenant au même cluster. Le terme à l'intérieur de min_I. donne une indication sur la capacité de $(I, \bar{I})$ à diviser S_i en deux sous-ensembles. L'hypothèse d'avoir une *constante de Cheeger* grande est équivalente à avoir un cluster S_i "compact".

Hypothèse A2 : Il existe un $\epsilon_1 > 0$ fixé tel que $\forall i_1, i_2 \in 1, ..., k, i_1 \neq i_2$, on a

$$\sum_{j \in S_{i_1}} \sum_{k \in S_{i_2}} \frac{A_{jk}^2}{\widehat{d}_j \widehat{d}_k} \leq \epsilon_1 \tag{4}$$

Ici, $\widehat{d}_j$ caractérise la connectivité entre le point j et les autres points du même cluster. Ainsi, tant que les A_{ij} sont petits, la somme sera aussi petite et l'hypothèse sera vérifiée pour un ϵ_1 petit.

Hypothèse A3 : Il existe un $\epsilon_2 > 0$ fixé tel que $\forall i = 1, ..., k, j \in S_i$, on a

$$\frac{\sum_{k:k\notin S_i} A_{jk}}{\widehat{d}_j} \leq \epsilon_2 (\sum_{k,l\in S_i} \frac{A_{jk}^2}{\widehat{d}_k \widehat{d}_l})^{-1/2} \tag{5}$$

Cette hypothèse traduit le fait que chaque point doit être plus connecté aux points appartenant au même cluster qu'à ceux appartenan à d'autres clusters. Le rapport entre les deux quantités doit alors être petit.

Hypothèse A4 : Il existe une constante $C > 0$ telle que $\forall i = 1, ..., k, j = i = 1, ..., n_i$, on a

$$\widehat{d}_j^{(i)} \geq \frac{\sum_{k=1}^{n_i} \widehat{d}_k^{(i)}}{(Cn_i} \tag{6}$$

Cette hypothèse assure qu'aucun point d'un cluster ne soit pas "moins" connecté que les autres points appartenant au même cluster.

Théorème 2 : On suppose les hypothèses A1, A2, A3 et A4 vérifiées. Soit $\epsilon = \sqrt{k(k-1)\epsilon_1 + k\epsilon_2^2}$. Si $\delta > (2+\sqrt{2})\epsilon$, alors il existe k vecteurs orthogonaux $r_1, r_2, ..., r_k$ ($r_i^T r_j = 1$ si $i = j$ et 0 sinon) tels que les lignes de Y vérifient :

$$\frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \|\widehat{y}_j^{(i)} - r_i\|_2^2 \leq 4C(4 + 2\sqrt{k})^2 \frac{\epsilon^2}{(\delta - \sqrt{2}\epsilon)})^2 \tag{7}$$

Cette condition assure que les lignes de Y forment des clusters compacts autours de k points bien distincts sur la surface de la k-sphère selon leur vraie cluster S_i.

3 Implémentation de Algorithme

L'algorithme final est le suivant :

(1) Construire la matrice d'affinité $A \in \mathbb{R}^{n\times n}$ définit par $A_{ij} = \exp(-\|s_i - s_j\|/2\sigma^2)$ si $i \neq j$ et $A_{ii} = 0$.

(2) Construire la matrice diagonale D définit par $D_{ii} = \sum_{j=1}^{n} A_{ij}$ et en déduire la matrice L définie par $L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$

(3) Trouver $x_1, x_2, ..., x_k$ les k plus grands vecteurs propres de L et construire la matrice $X = [x_1 x_2 ... x_k] \in \mathbb{R}^{n\times k}$ en collant les vecteurs propres en colonnes.

(4) Construire la matrice Y déduite de X après normalisation des lignes de cette dernières. $Y_{ij} = X_{ij}/(\sum_j X_{ij}^2)^{1/2})$

(5) Traiter chaque ligne de Y comme un point de $\mathbb{R}^n$. Les classifier en k clusters en utilisant un Kmeans ou tout autre algorithme minimisant la distortion.

(6) Finalement, associer le point original s_i au cluster j si et seulement si la ligne i de la matrice Y a été assignée au cluster j.

4 Résultats et discussion

4.1 Résultats

L'algorithme a été appliqué à différents problèmes de classification et présente des résultats satisfaisants.
Pour le choix du paramètre σ^2, on se base sur le thèorème 2 qui prédit que les lignes de Y forment des clusters "compacts" sur la surface de la k-sphère. On lance l'algorithme avec différentes valeurs et garde celle qui donne les clusters les plus "compacts" possibles.
L'initialisation du Kmeans à l'étape 5 a été faite en sachant que les clusters sont séparés de 90°.
La validation et la discussion des résultats de l'algorithme sera développée d'avantage suite à l'implémentation de ce dernier et l'évaluation de ses performances par rapport à d'autres algorithmes.
Les résultats des tests seront présentés lors de la soutenance avec une démonstration sur machine.

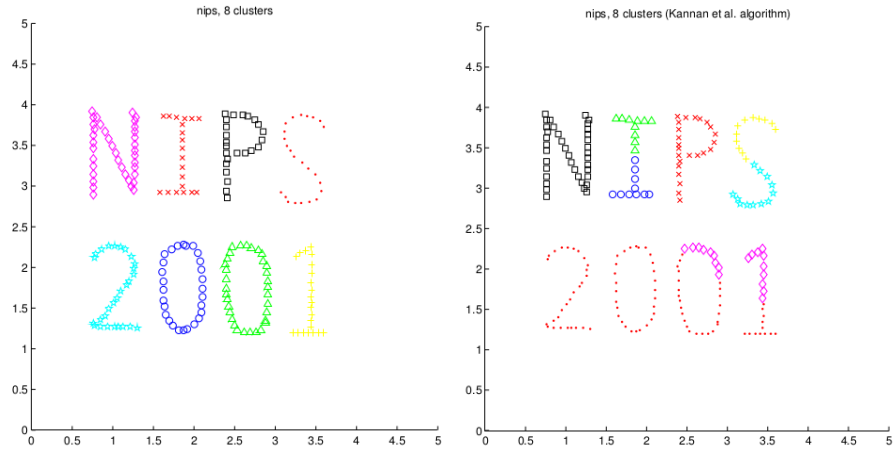

FIGURE 1 – nips (8 clusters) avec notre algorithme à gauche et Kannan et al. algorithme à droite

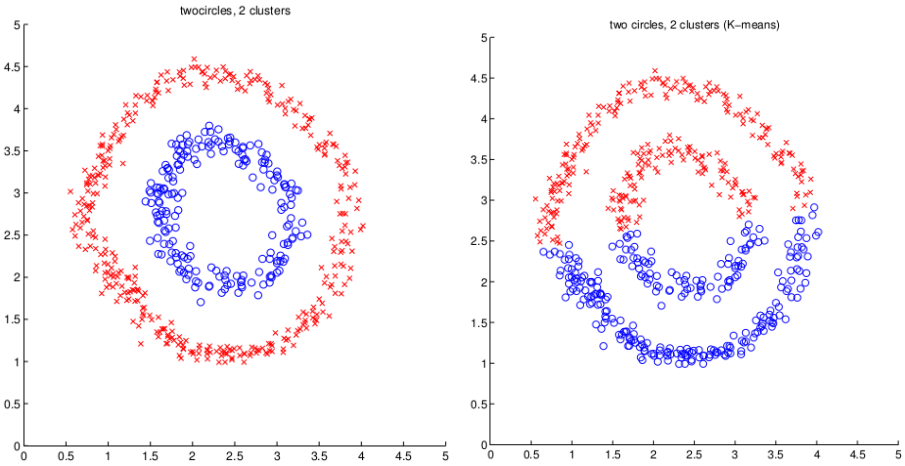

FIGURE 2 – Deux cercles (2 clusters) avec notre algorithme à gauche et Kmeans à droite

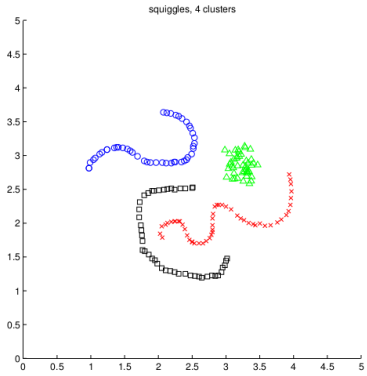

FIGURE 3 – squiggles (4 clusters) avec notre algorithme

4.2 Conclusion

Cette article aborde une approche d'apprentissage non-supervisé basé sur les méthodes spectrales. On a tout d'abord commencé par expliciter le résultat sur un cas "idéal" pour ensuite le généraliser. La généralisation se base sur la téorie de perturbation des matrice ainsi que la théorie spectrale des graphes. Afin d'énoncer le théorème finale, un ensemble d'hypothèses a été émi. Ces hypothèses concernent essentiellement la distribution des données initiales et permettent de s'assurer de l'existance de clusters "compacts", bien disjoints et dont tous les points sont bien connectés. Même si l'article passe rapidement sur certains éléments, l'ensemble reste cohérent et le renvoie à des références dans la littérature permet de mieux appréhender les théories mises en application.