Algorithmes de classification non-supervisée pour le partitionnement de données

Classification spectrale – Présentation et applications

Plan

- Classification spectrale (spectral clustering)
  - Introduction
  - Illustration
  - Formulation
  - Algorithme
  - Applications
- Applications
  - Exemple dans R
  - Désambiguïsation automatique pour améliorer la RI

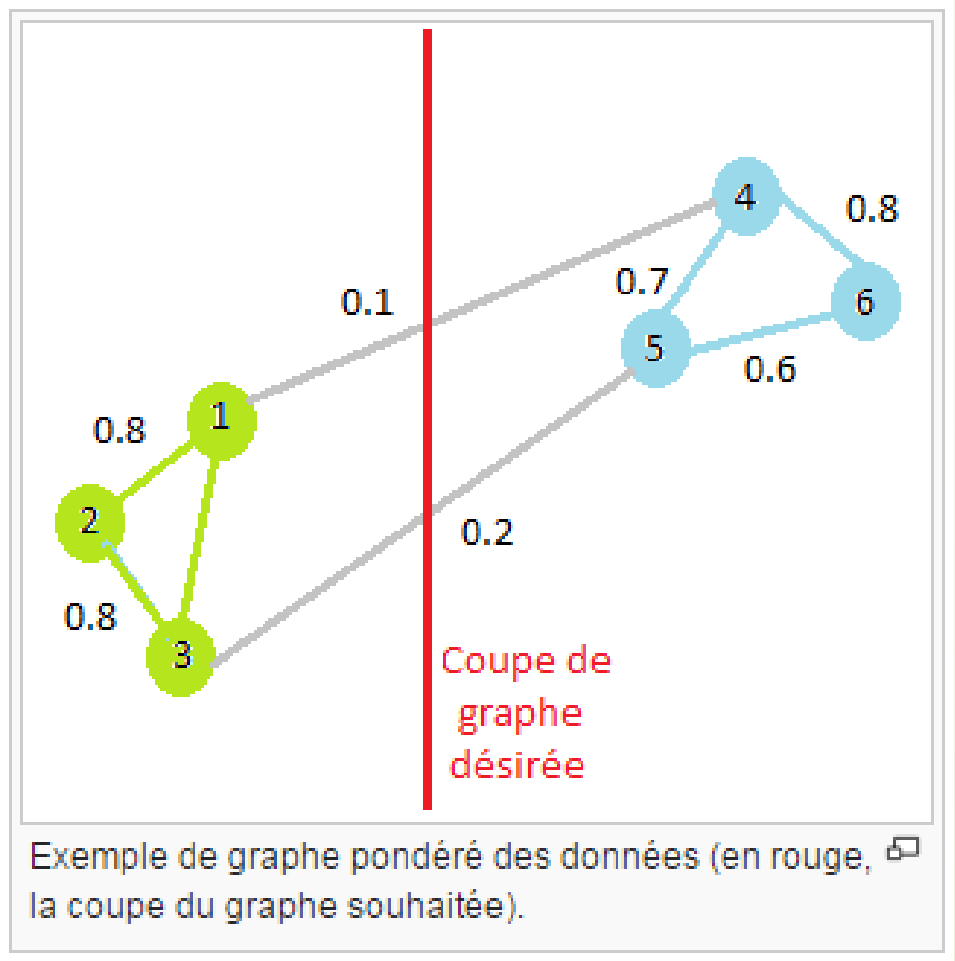
Classification spectrale

Présentation

Classification spectrale – Introduction

- Domaine de l'intelligence artificielle
- Famille d'algorithmes de classification non-supervisée pour le partitionnement de données (*clustering*)
- Avantages:
  - Simplicité relative d'implémentation (extraction de vecteurs propres d'une matrices de similarité)
  - Efficacité

Classification spectrale – Illustration



Exemple de graphe pondéré des données (en rouge, la coupe du graphe souhaitée).

La matrice laplacienne d'un graphe non orienté

- $G = (V, E)$, un graphe non orienté
- La matrice laplacienne $L = D - A$ ou D est la matrice des degrés de G et A est la matrice d'adjacence de G.
- $$L_{i,j} := \begin{cases} \deg(v_i) & \text{si } i = j \\ -1 & \text{si } i \neq j \text{ et si } i \text{ et } j \text{ sont reliés par une arête} \\ 0 & \text{sinon} \end{cases}$$
- Plus généralement: soit $G = (V, E)$, un graphe non orienté et non réflexif avec n sommets, pondéré par la fonction *poids* qui associe son poids $w(v_i, v_j)$ à toute arête (v_i, v_j)
- $$L_{i,j} := \begin{cases} \deg(v_i) = \sum_{k=1}^{n} w(v_i, v_j) & \text{si } i = j \\ -w(v_i, v_j) & \text{si } i \neq j \text{ et } (v_i, v_j) \in E \\ 0 & \text{sinon} \end{cases}$$
- Généralisation possible pour les graphes orientés; la matrice L n'est plus symétrique

Classification spectrale – Formulation

- $(x_1, \dots, x_n)$ un ensemble d'observations
- $s_{ij} \geq 0$ une mesure de similarité entre les paires d'observations (x_i, x_j)
- But: diviser les observations dans plusieurs groupes; les observations du même groupe sont similaires; les observations dans des groupes différents sont dissimilaires
- Le graphe de similarité $G = (V, E)$. Le sommet v_i représente une observation x_i
- Deux sommets sont connectés si la similarité s_{ij} est positive (ou alors supérieure à un seuil)
- Les arêtes sont pondérées par les valeurs s_{ij}
- Problème reformulé comme un problème de partition des graphes

Classification spectrale – Formulation

- Graphe de similarité locale:
  - ε-voisinage (ε-neighborhood)
  - Les k plus proches voisins (k-nearest neighbors)
  - Graphe complet
- Les k plus proches voisins
- Problème de partition NP-difficile (NP-hard)
- La classification spectrale trouve les m vecteurs propres $U_{n\times m}$qui correspondent aux m plus petites valeurs propres de L
- Les lignes de la matrice U sont classifiées en utilisant l'algorithme k-moyennes

Classification spectrale – Algorithme

- ***Entrée: la matrice de similarité*** $S \in \boldsymbol{R}^{n \times n}$***; le nombre*** $\boldsymbol{k}$ ***de classes à construire***

1. **Construire le graphe de similarité**
2. **Calculer la matrice laplacienne** L **non normalisée**
3. **Calculer les premiers** $k-1$ **vecteurs propres de** L **qui correspondent aux** $k-1$ **plus petites valeurs propres de** L
4. **Soit** $U \in \boldsymbol{R}^{n \times (k-1)}$ **la matrice qui contient les vecteurs** $(u_1, \dots, u_{k-1})$ **comme colonnes**
5. **Pour** $i = 1, \dots, n,$ **soit** $y_i \in \boldsymbol{R}^{k-1}$ **le vecteur correspondent à la** i**-ème ligne de** U
6. **Grouper les points** $(y_i)_{i=1,\dots,n}$ **dans** $\boldsymbol{R}^{k-1}$ **avec l'algorithme** k**-moyennes pour obtenir les classes** $C_1, \dots, C_k$

- ***Sortie: Les classes (groupes)*** $A_1, \dots, A_k$ ***avec*** $A_i = \{j | y_j \in C_i\}$

Classification spectrale – Applications

- Indexation et recherche par le contenu
- Recherche de documents Web
- Segmentation d'images
- Analyse de marchés
- Analyse de documents
- Classification non-supervisée

Applications

Exemple dans R; Désambiguïsation automatique en RI

Exemple dans R

- Voir le script `Spect_clust.R`

Désambiguïsation automatique en RI [Chifu et al. 2014]

- « Word sense discrimination in information retrieval: A spectral clustering-based approach »
- WSD (Discrimination des sens des mots) par l'intermède de la classification spectrale: méthode état de l'art
- But: désambiguïsation → réordonnancement des documents → amélioration de la précision de top (P@5, P@10, P@30)
- Tests sur des collections TREC
- Amélioration de 8%
- Focus sur les requêtes avec des mauvaises performances
- http://www.sciencedirect.com/science/article/pii/S0306457314001046/pdfft?md5=d155b743b87e028e73a5552946636f0&pid=1-s2.0-S0306457314001046-main.pdf