Observations sur les distributions latentes aux matrices laplaciennes de graphes

Pierrick Bruneau
Benoît Otjacques

Luxembourg Institute of Science and Technology
5, avenue des Hauts-Fourneaux
L-4362 Esch-sur-Alzette
prénom.nom@list.lu

Résumé. L'algorithme de clustering spectral permet en principe d'extraire des clusters de formes arbitraires à partir de données numériques. Cette propriété a contribué à sa popularité, et même si ses bases théoriques sont établies depuis plus d'une décennie, des variantes en ont été proposées jusqu'à récemment. Son fonctionnement repose sur une transformation vers un espace latent dans lequel des formes de clusters arbitraires sont converties en structures faciles à traiter par un algorithme tel que k-means. Toutefois, les distributions dans cet espace latent n'ont été que peu discutées, beaucoup d'auteurs supposant que les propriétés prédites par la théorie sont vérifiées. Cet article propose alternativement une approche qualitative pour vérifier si cette structure idéale est effectivement obtenue en pratique. Le travail consiste également à identifier les paramètres de variabilité commandant à la transformation vers l'espace latent, via un état de l'art synthétique de la théorie sous-jacente au clustering spectral. Les observations tirées de nos expériences permettent d'identifier les combinaisons de paramètres efficaces, et les conditions de cette efficacité.

1 Introduction

L'algorithme de clustering spectral (Shi et Malik, 2000; Ng et al., 2002) se distingue des algorithmes à base de centroides (k-means, EM (Bishop, 2006)) principalement par sa capacité à extraire des clusters non convexes et sphériques.

Tous les algorithmes de clustering spectral de la littérature se basent sur l'analyse d'une matrice de similarité entre éléments d'un jeu de données, pouvant être vue comme un graphe pondéré. Les variantes diffèrent ensuite selon qu'elles étudient l'évolution de cette matrice selon un processus stochastique alternant expansion et inflation (van Dongen, 2008; Liu et al., 2013), ou qu'elles analysent les vecteurs propres du laplacien du graphe (Ng et al., 2002; Lin et Cohen, 2010). Dans ce dernier cas, sur lequel nous nous concentrons dans cet article, le reste de l'algorithme se résume essentiellement à appliquer l'algorithme k-means à une représentation extraite des vecteurs propres. La distribution des données dans cette représentation est donc essentielle au bon fonctionnement de l'algorithme de clustering spectral.