Une heuristique pour le paramétrage automatique de l'algorithme de *clustering* spectral

Pierrick Bruneau, Olivier Parisot, Philippe Pinheiro

Centre de Recherche Public - Gabriel Lippmann, 41, rue du Brill, L-4422 Belvaux
(bruneau | parisot | pinheiro)@lippmann.lu

Résumé. Trouver le nombre optimal de groupes dans le contexte d'un algorithme de *clustering* est un problème notoirement difficile. Dans cet article, nous en décrivons et évaluons une solution approchée dans le cas de l'algorithme spectral. Notre méthode présente l'avantage d'être déterministe, et peu coûteuse. Nous montrons qu'elle fonctionne de manière satisfaisante dans beaucoup de cas, même si quelques limites amènent des perspectives à ce travail.

1 Introduction

Le *clustering* d'un ensemble d'objets en un nombre de groupes pré-déterminé est un problème souvent difficile suivant le critère d'optimisation ou le modèle choisi. Le choix optimal du nombre de groupes (identifié de manière univoque par la variable k dans le reste de l'article) l'est probablement davantage. Le principe généralement accepté du rasoir d'Occam, favorisant un nombre minimal de *clusters*, s'oppose à leur exhaustivité, sans qu'un compromis satisfaisant pour tous soit possible *a priori*. En pratique, ce paramètre est donc souvent laissé à la discrétion du praticien par les logiciels d'analyse de données, même récents. Dans le cas d'une approche exploratoire, où k peut être inconnu, une heuristique est souhaitable.

Dans cet article, nous nous limitons à l'algorithme de *clustering* spectral, et proposons une nouvelle manière extrêmement simple, peu coûteuse, et bien fondée, d'estimer k à partir du spectre de laplacien propre à cet algorithme. Le test de Bartlett pour l'égalité des variances est utilisé depuis longtemps pour déterminer le nombre de facteurs à retenir dans le contexte d'une Analyse en Composantes Principales (ACP) (James, 1969). Nous montrons qu'il est possible de l'adapter assez facilement pour estimer k dans le contexte de l'algorithme de *clustering* spectral.

Dans un premier temps, nous rappelons l'état de l'art du *clustering* spectral, ainsi que des méthodes d'estimation automatiques de k existantes. Nous décrivons ensuite notre méthode, *in fine* matérialisée par un algorithme simple. L'efficacité de la méthode est illustrée par des expériences sur des données synthétiques et réelles de la littérature. L'analyse critique de nos résultats nous permet de formuler quelques perspectives, données en conclusion.