Accélération de EM pour données qualitatives : étude comparative de différentes versions

Mohamed Nadif, François-Xavier Jollois

LITA - IUT de METZ, Université de Metz,
Ile du Saulcy, 57045 METZ Cedex, France
{jollois,nadif}@iut.univ-metz.fr,

Résumé. L'algorithme EM est très populaire et très efficace pour l'estimation de paramètres d'un modèle de mélange. L'inconvénient majeur de cet algorithme est la lenteur de sa convergence. Son application sur des tableaux de grande taille pourrait ainsi prendre énormément de temps. Afin de remédier à ce problème, nous étudions ici le comportement de plusieurs variantes connus de EM, ainsi qu'une nouvelle méthode. Celles-ci permettent d'accélérer la convergence de l'algorithme, tout en obtenant des résultats similaires à celui-ci. Dans ce travail, nous nous concentrons sur l'aspect classification. Nous réalisons une étude comparative entre les différentes variantes sur des données simulées et réelles et proposons une stratégie d'utilisation de notre méthode qui s'avère très efficace.

1 Introduction

L'utilisation des modèles de mélange dans la classification est devenue une approche classique et très puissante (voir par exemple [Banfield et Raftery, 1993], et [Celeux et Govaert, 1995]). En traitant la classification sous cette approche, l'algorithme EM [Dempster *et al.*, 1977] composé de deux étapes : Estimation et Maximisation, est devenu quasiment incontournable. Celui-ci est très populaire pour l'estimation de paramètres. Ainsi, de nombreux logiciels sont basés sur cette approche, comme Mclust-EMclust [Fraley et Raftery, 1999], EMmix [McLachlan et Peel, 1998] ou MIXMOD [Biernacki *et al.*, 2001]. Ce succès tient à sa simplicité, à ses propriétés théoriques et à son bon comportement pratique. De plus, un intérêt grandissant se fait ressentir actuellement pour les données qualitatives. On peut citer le logiciel AutoClass [Cheeseman et Stutz, 1996], très utilisé dans la communauté Fouille des Données.

Malheureusement, son principal inconvénient réside dans sa lenteur due au nombre élevé d'itérations parfois nécessaire pour la convergence, ce qui rend son utilisation inappropriée pour les données de grande taille. Plusieurs versions ont été faites pour accélérer cet algorithme et beaucoup d'entres elles agissent sur l'étape maximisation. Ici, comme nous nous intéressons au modèle des classes latentes, l'étape de maximisation ne présente aucune difficulté pour le calcul des paramètres. Nous avons donc choisi d'étudier des versions particulièrement adaptée aux données de grande taille et qui utilisent une étape partielle d'estimation au lieu d'une étape Estimation complète. Cette version semble très efficace pour des mélanges Gaussiens, nous proposons ici de l'appliquer sur le modèle de mélange des classes latentes et de discuter son comportement.