Évaluation des algorithmes LEM et *e*LEM pour données continues

F.-X. Jollois *, M. Nadif **

*CRIP5, Université de Paris 5,
45 rue des Saint-Pères,
75270 Paris Cedex 06, France
francois-xavier.jollois@univ-paris5.fr
**LITA - UFR MIM, Université de Metz,
Ile du Saulcy,
57045 METZ Cedex 1, France
nadif@iut.univ-metz.fr

Résumé. Très populaire et très efficace pour l'estimation de paramètres d'un modèle de mélange, l'algorithme EM présente l'inconvénient majeur de converger parfois lentement. Son application sur des tableaux de grande taille devient ainsi irréalisable. Afin de remédier à ce problème, plusieurs méthodes ont été proposées. Nous présentons ici le comportement d'une méthode connue, LEM, et d'une variante que nous avons proposée récemment *e*LEM. 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, et nous illustrons le bon comportement de notre variante sur des données continues simulées et réelles.

1 Introduction

Plusieurs méthodes de classification utilisées sont basées sur une distance ou une mesure dissimilarité. Or, 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), Mixmod (Biernacki et al., 2001) ou AutoClass (Cheeseman et Stutz, 1996).

Malheureusement, le principal inconvénient de EM 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. Ayant testé plusieurs méthodes (Nadif et Jollois, 2004), nous avons retenu l'algorithme LEM (Thiesson et al, 2001) qui utilise une étape partielle d'*Estimation* au lieu d'une étape complète. A partir de cet algorithme, nous avons cherché à améliorer sa performance et avons proposé une variante plus efficace, *e*LEM. Sur des données qualitatives simulées et réelles, les performances de cette nouvelle version ont été très encourageantes. Le principal objectif de