Le logiciel SpaCEM^3 pour la classification de données complexes

Juliette Blanchet, Florence Forbes, Sophie Chopart, Lamiae Azizi

Equipe Mistis, INRIA Grenoble Rhône-Alpes, Laboratoire Jean Kuntzmann

Résumé Le logiciel SpaCEM^3 (Spatial Clustering with EM and Markov Models) propose une variété d'algorithmes pour la classification, supervisée ou non supervisée, de données uni ou multi-dimensionnelles en interaction, certaines de ces données pouvant être manquantes. Les structures de dépendances prises en compte sont celles pouvant être décrites par un graphe fini quelconque. Elles incluent le cas particulier des grilles régulières utilisées notamment en segmentation d'images. L'approche principale se fonde sur l'utilisation de l'algorithme EM pour une classification *floue* et sur les modèles de champs de Markov pour la modélisation des dépendances. L'estimation est basée sur des développements récents [6, 8, 7] mettant en oeuvre des techniques d'approximations variationnelles de type champ moyen.

Keywords : champs de Markov cachés, modèles de Markov triplets, données manquantes, algorithmes de type EM, champ moyen, sélection de modèles.

1 Introduction

La classification consiste à regrouper des individus en groupes homogènes par rapport aux mesures effectuées sur ces individus. Un individu au sens large peut être un pixel d'une image, un gène, un segment de texte, etc. Les mesures effectuées sur les individus peuvent être de nature variable (réelles, entières, dans l'intervalle [0, 1], etc.), uni- ou multi-dimensionnelles. L'approche probabiliste repose alors sur la donnée d'un modèle pour le couple des observations et des classes, généralement décomposé en un modèle régissant les classes et un modèle (de bruit) régissant la génération des observations lorsque les classes sont connues. Dans la pratique, des hypothèses simplificatrices sont souvent adoptées :

(1) au niveau de la modélisation, on suppose en général que les classes sont indépendantes et que le modèle de bruit se factorise sur les individus (on parle alors de bruit indépendant). Sous ces deux hypothèses, les individus sont alors implicitement supposés indépendants. Enfin, le bruit est supposé être de forme assez simple, gaussien en général, ou au moins unimodal ;

(2) au niveau des cas traités, les observations sont, en général, ou bien de dimension raisonnable, ou bien les composantes de chaque observation sont supposées indépendantes. De plus, les données utilisées sont complètes. Lorsque, pour différentes raisons, certaines observations viennent à manquer, soit ces observations ne sont pas traitées (comme si aucune mesure n'avait été faite sur l'individu correspondant), soit les valeurs manquantes sont remplacées de manière brutale (par des zéros, la moyenne, etc.).

En pratique, il existe beaucoup de cas où ces hypothèses sont mises en défaut et ne donnent pas de résultats satisfaisants. En particulier, les observations effectuées sont souvent dépendantes (les niveaux de gris des pixels d'une image par exemple). De plus, du fait des progrès des appareils de mesure et des capacités de stockage, nombre de données modernes sont en grande dimension. Sans paramétrisation particulière, le bruit doit alors