Nouvelle méthode de classification adaptée aux données de grande dimension : application aux données de biopuces

Doulaye Dembélé

IGBMC, CNRS-IMSERM-ULP, 1 rue Laurent Fries, BP 10142
Parc d'Innovation, 67404 Illkirch Cedex, France
Doulaye.Dembele@igbmc.u-strasbg.fr,
http://www-microarrays.u-strasbg.fr

Résumé. Nous proposons une nouvelle méthode de classification adaptée aux données de grande dimension. Pour ces données la distance de Chebyshev semble intéressante, car elle nécessite moins de temps de calcul comparée à la distance Euclidienne, plus utilisée en raison de ses bonnes propriétés géométriques. La méthode proposée combine les méthodes de regroupement hiérarchique et par partition pour obtenir le nombre de classes dans les données. Des données issues d'expériences de biopuces sont utilisées pour illustrer les performances de la méthode proposée.

1 Introduction

La classification permet de représenter des données sous une forme plus aisée à interpréter, à visualiser ou à manipuler. Nous proposons une nouvelle méthode de classification adaptée aux données de grande dimension. Pour ces données le problème d'espace vide est connu (Dohono, 2000). D'autres faiblesses sont liées à l'utilisation de la distance Euclidienne : sous certaines hypothèses sur la distribution des échantillons, les distances entre toutes les paires de points des données sont identiques quand la dimension augmente (Beyer et al., 1998). Dans ces conditions il est impossible de discriminer les classes, s'il y en a, dans les données. Il est aussi montré dans (Herault et al., 2002) qu'en augmentant l'ordre de la métrique de Minkowski, il est possible d'augmenter le rang de la matrice des distances des données prises deux à deux. Le maximum de dimension pour la matrice des distances sera obtenu pour un ordre infini, c'est-à-dire en utilisant la distance de Chebyshev. Notons que le rang de la matrice des distances définit le degré de contraste permettant d'obtenir des classes dans les données, c'est-à-dire le degré de redondance dans les données. La distance de Chebyshev semble alors intéressante. De plus, elle nécessite moins de temps de calcul comparé à la distance Euclidienne généralement utilisée en raison de ses bonnes propriétés géométriques. Le gain en temps de calcul n'est pas négligeable pour les données de grande dimension comme celles générées par les biologistes dans le cadre de l'étude de l'expression des gènes à l'aide de la technologie des biopuces.

Nous nous intéressons ici aux méthodes de classification heuristiques automatiques et non supervisées également appelées clustering. Ces méthodes se divisent en deux grandes familles : les méthodes hiérarchiques et les méthodes par partition. Les méthodes hiérarchiques ne nécessitent pas la connaissance *a priori* du nombre de classes dans les données. Leur résultat