

Improving supervised classification for high dimensional data by adding external information. Application to microarray data.

Maela Kloareg and David Causeur

Supervised classification is a common problem in high dimensional and low sample size situations as, for example, microarray data. Accuracy of the standard techniques is compromised in the case of highly correlated variables. We compare the results of a few classification methods on simulated data under various correlation schemes by focusing on the misclassification rate and suggest improvements based on external information. Finally we apply this method to a real life microarray data set.

Améliorer les procédures de classification supervisée en intégrant de l'information externe. Application à des données génomiques.

La construction de règles d'affectation d'un individu à un groupe prédéfini est une problématique récurrente dans le cas de données en grande dimension, telles que les données post-génomique par exemple, en particulier dans le cas d'une structure de dépendance entre les variables. Nous comparons ici les résultats obtenus par plusieurs méthodes sur des données simulées en étudiant les taux de mauvais classement et proposons d'améliorer ces procédures par l'intégration d'information externe. Enfin, nous mettons en application ces méthodes sur un jeu de données réel obtenu à partir de biopuces.