Sélection de variables pour la classification par mélanges gaussiens pour prédire la fonction des gènes orphelins

Cathy Maugis[1], Marie-Laure Martin-Magniette[2,3], Jean-Philippe Tamby[3], Jean-Pierre Renou[3], Alain Lecharny[3], Sébastien Aubourg[3], Gilles Celeux[4]

[1] Département de Mathématiques, Université Paris-Sud 11, Orsay, France
E-mail: cathy.maugis@math.u-psud.fr
[2] UMR AgroParisTech/INRA MIA 518, Paris, France
E-mail: marie_laure.martin@agroparistech.fr
[3] URGV UMR INRA 1165, CNRS 8114, UEVE, Evry, France
E-mail: aubourg@evry.inra.fr; lecharny@evry.inra.fr; renou@evry.inra.fr; tamby@versailles.inra.fr
[4] INRIA Saclay Île-de-France, France
E-mail: gilles.celeux@inria.fr

Abstract Biologists are interested in predicting the gene functions of sequenced genome organisms according to microarray transcriptome data. The microarray technology development allows one to study the whole genome in different experimental conditions. The information abundance may seem to be an advantage for the gene clustering. However, the structure of interest can often be contained in a subset of the available variables. The currently available variable selection procedures in model-based clustering assume that the irrelevant clustering variables are all independent or are all linked with the relevant clustering variables. A more versatile variable selection model is proposed, taking into account three possible roles for each variable: The relevant clustering variables, the redundant variables and the independent variables. A model selection criterion and a variable selection algorithm are derived for this new variable role modelling. The interest of this new modelling for discovering the function of orphan genes is highlighted on a transcriptome dataset for the *Arabidopsis thaliana* plant.
Keywords: Variable selection, model-based clustering, transcriptome data, orphan genes.

Résumé Les biologistes s'attachent actuellement à prédire la fonction des gènes d'organismes de génome séquencé à partir de données transcriptomes, issues de l'utilisation des puces à ADN. Le développement de cette technologie permet de tester l'expression de l'ensemble du génome dans de nombreuses conditions expérimentales. Cette quantité d'information peut alors sembler être un atout pour la classification des gènes. Pourtant il est courant que seul un sous-ensemble contienne l'information pertinente pour la classification. Les procédures de sélection des variables en classification non supervisée par mélanges gaussiens supposent généralement que les variables non informatives pour la classification sont soit toutes indépendantes, soit liées à des variables informatives. Nous proposons une nouvelle modélisation du rôle des variables plus polyvalente : les variables sont soit informatives pour la classification, soit redondantes, soit totalement indépendantes. Nous proposons un critère de sélection des variables et un algorithme pour cette nouvelle modélisation. L'intérêt