Title: Variable selection in model-based clustering for predicting the functions of orphan genes.

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 proposed variable selection procedure in model-based clustering takes 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.

Cathy MAUGIS
INSA, Département de Génie Mathématique
135, avenue de Rangueil
31077 TOULOUSE Cedex 4, FRANCE

INSA: Bureau 120 Tél: 05 61 55 92 30
IMT: Bureau 220 Tél: 05 61 55 67 71

http://www.math.univ-toulouse.fr/~maugis