FEATURE SELECTION METHODS FOR EARLY PREDICTIVE BIOMARKER DISCOVERY USING UNTARGETED METABOLOMIC DATA

Dhouha Grissa[1], Mélanie Pétéra[2], Marion Brandolini[2], Amedeo Napoli[3], Blandine Comte[1] and Estelle Pujos-Guillot[1,2]*

[1]INRA, UMR1019, UNH-MAPPING, 63000 Clermont-Ferrand, France
[2]INRA, UMR1019, Plateforme d'Exploration du Métabolisme, 63000, Clermont-Ferrand, France
[3]LORIA, B.P. 239, 54506 Vandoeuvre-lès-Nancy, France
*E-mail: estelle.pujos@clermont.inra.fr

Untargeted metabolomics is a powerful phenotyping tool for better understanding biological mechanisms involved in human pathology development and identifying early predictive biomarkers. This approach, based on powerful analytical platforms, such as mass spectrometry, chemometrics and bioinformatics, generates massive and complex data that need appropriate analyses to extract biologically meaningful information [1]. In this context, this work consists in designing a workflow describing the general feature selection process, using knowledge discovery and data mining methodologies to propose advanced solutions for predictive biomarker discovery.

Data were collected from a mass spectrometry-based untargeted metabolomic approach performed on subjects from a case/control study within the GAZEL French population-based cohort. Different feature selection approaches were applied either on the original metabolomic dataset or on reduced subsets. The strategy was focused on evaluating a combination of numeric-symbolic approaches for feature selection with the objective of obtaining the best combination of metabolites, producing an effective and accurate predictive model. Relying first on numerical approaches, and especially on machine learning methods (SVM and RF-based methods) and on univariate statistical analyses (ANOVA), a comparative study was performed on the original metabolomic dataset and reduced subsets. As resampling method, LOOCV was applied to minimize the risk of overfitting. The best k-features obtained with different scores of importance from the combination of these different approaches were compared and allowed determining the variable stabilities using Formal Concept Analysis.

The results revealed the interest of RF-Gini combined with ANOVA for feature selection as these two complementary methods allowed selecting the 48 best candidates for prediction. Using linear logistic regression strategy on this reduced dataset enabled us to obtain the best performances in terms of prediction accuracy and number of false positive with a model including 5 top variables. Therefore, these results highlighted the interest of feature selection methods and the importance of working on reduced datasets for the identification of predictive biomarkers issued from untargeted metabolomics data. Even if they are still not usually applied, these data mining methods are essential tools to deal with massive datasets and contribute to elucidate complex phenomena associated with chronic disease development. With this help, the experts of the scientific field will go deeper in interpretation, attesting the success of the knowledge discovery process.

Project funded by the INRA DID'IT Metaprogramme.

KEYWORDS: feature selection, untargeted metabolomics, biomarker discovery, prediction

REFERENCES:

1. Xi B, Gu H, Baniasadi H, Raftery D. *Methods Mol Biol* (2014) **1198**(1):333-53.2.