Proposition de recherche doctorale

Probabilistic Omics Data integration and Predictive Analysis for Personalized Medicine

Mots clés :

- Directeur de thèse : Jean-Daniel ZUCKER
- Co-encadrant(s) :
- Unité de recherche : Unité de modélisation mathématique et informatique de systèmes complexes
- Ecole doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal: Divers

Résumé du projet de recherche (Langue 1)

{{PRESENTATION DU SUJET}} {1. Contexte} High-throughput technologies such as whole-genome sequencing revolutionized biological research and transformed it from a relatively data-poor discipline into a domain which is rich in data. Today the challenge is to process, analyze, and to interpret the available data, and to derive fundamental and practical biological knowledge. The data acquisition becomes cheaper, and the increasing amount of omics data provides with unprecedented broad views of living organisms and biological systems. Statistical machine learning is a relatively young but actively developing field on the intersection of mathematics, computer science, and statistics. For already a couple of decades machine learning has been applied to a number of biological challenging: predictive modeling (classification), descriptive modeling (clustering), and dimensionality reduction. Machine learning provides machineries for data processing, typical for biological applications: how to deal with tasks where the number of instances is small but the dimensionality it high. These high dimensional problems are called \( p >> N \) problems [Hastie, T. & Tibshirani, R., 2009. The elements of statistical learning: data mining, inference, and prediction, 2nd Edition, Chapter 18]. {2 Objectif} {2.1} One of the main applications is the prediction of a biological parameter based on "omics" data. The supervised classification can be used to detect the presence of a disease, identify its subtype or predict the effect of a drug in order to assist the therapeutic decision. The classifiers with reject do not classify the examples whose confidence in the predicted class is low. These classifiers are more reliable then classic classifier and depend on the tradeoff between accuracy and rejection rate, they: particularly useful in the medical context where a wrong decision can have dramatic consequence. One of the main characteristic of "omics" data is their huge disproportion between the number of examples (some tens) and numbers of feature (several thousands). A feature selection step is therefore necessary. Currently there is not works on feature selection for classifier with rejection option. The topics of this PhD is to develop new feature selection methods taking in account the rejection area. This approach should improve the greatly the reliability and usefulness of the diagnosis systems based on "omics" data in particular in metagenomics. {2.2} The cardiometabolic diseases (CMD) are progressive metabolic disorders leading to chronic stages of cardiovascular diseases and heart insufficiency. For a long time the genetic diversity has been ignored and the same treatment has been applied to all patients with a similar diagnosis. However, it was not clear how individual patients respond to it. Advances in data processing from large epidemiological and genome-wide studies are expected to contribute to the resolution of the worldwide cardiometabolic epidemics. The identification of responders to therapies is crucial to provide the most appropriate treatment and avoid unnecessary medications. Machine learning possesses powerful frameworks to integrate a vast amount of data from heterogeneous sources, design new models, and test multiple hypotheses and therapeutic products. Probabilistic graphical models (PGM), such as Bayesian networks, are widely used to model structures in various applications, and in particular, in bioinformatics. The PGM techniques have been applied to data with arbitrary structure. The PhD candidate will develop methods of data integration of bioclinical and environmental phenotyping together with personalized omics (metagenomics, metabolomics, transcriptomics) with the objective of developing new strategies for personalized medicine.