

Proteomic and genomic data classification using decision tree based ensemble methods

Application to the diagnosis of inflammatory diseases

P. Geurts¹ M. Fillet² D. de Seny² M.-A. Meuwis²
M.Malaise² M.-P. Merville² L. Wehenkel¹

¹Department of Electrical Engineering and Computer Science

²Laboratory of Clinical Chemistry and Rheumatology

^{1,2}CBIG - Centre of Biomedical Integrative Genoproteomics
University of Liège

Benelux Bioinformatics Conference, 2005

- 1 Problem and Motivation
- 2 Methods
- 3 Application to the Diagnosis of Inflammatory Diseases

Based on the paper:

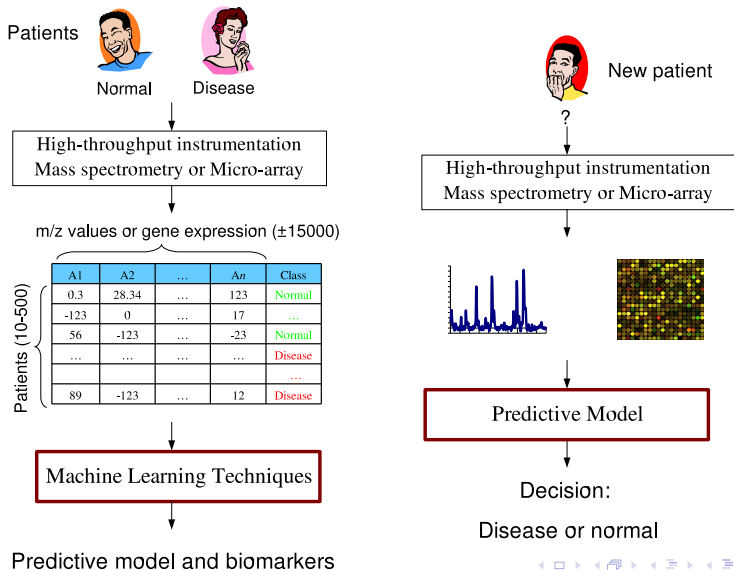


P. Geurts, M. Fillet, D. de Seny, M.-A. Meuwis, M. Malaise, M.-P. Merville, L. Wehenkel.

Proteomic mass spectra classification using decision tree based ensemble methods

To appear in *Bioinformatics* (2005).

Basic methodology



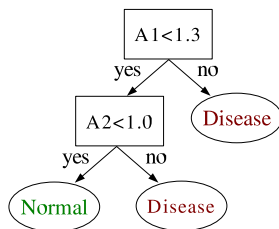
Machine Learning Problem Characteristics

- High number of (potentially irrelevant) variables (several thousand)
 - ⇒ Variable selection, use of robust machine learning methods
 - Small set of samples (several tens of measurements for each class)
 - ⇒ Error estimates by cross-validation, variance reduction techniques
 - High noise due to the data acquisition technology
 - ⇒ Pre-processing specific to the data acquisition technology
- ⇒ Difficult problems from a machine learning point of view

A Sample of Machine Learning Methods

Several machine learning methods used for these problems:

- *k*-NN
 - Simple but not accurate
- Decision trees
 - Simple, interpretable but not accurate
- Neural networks
 - Accurate but complex and not interpretable
- Support vector machines
 - (Very) accurate but complex and not interpretable
- **Tree based ensemble methods**
 - **Simple, accurate but not interpretable**



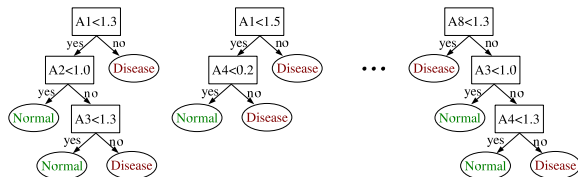
(A1 and A2 are the expressions of some genes or the intensities at some m/z positions)

Advantages: interpretable, easy to use, highly scalable, robust

Drawbacks: not so accurate, not so interpretable (high variance)

Ensemble of Decision Trees

- Idea: Build several (random) trees and aggregate their predictions (by voting)



- Examples of methods: bagging, boosting, random forests, extra-trees

Advantages: accurate, easy to use, fast, robust

Drawback: not interpretable

Extremely Randomized Trees (Extra-Trees)

- Extreme randomization of the tree induction method
- Algorithm's main ideas:
 - Selection of the best among only K variables drawn at random
 - Random selection of the discretization threshold
 - fully developed trees (no pruning) from the full learning set (no bootstrap)
- Main advantages:
 - good accuracy
 - computational efficiency (specially when the number of attributes is important)
 - conceptual simplicity

Variable Ranking with Tree Ensembles

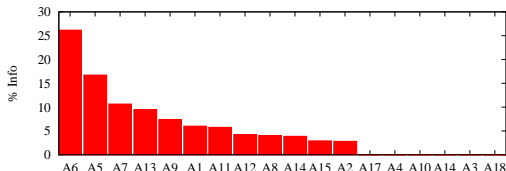
- Compute for each split in a tree the information brought by the split about the classification



$$I(\text{node}) = \#S \cdot H_C(S) - \#S_{\text{yes}} \cdot H_C(S_{\text{yes}}) - \#S_{\text{no}} \cdot H_C(S_{\text{no}}), \text{ with}$$

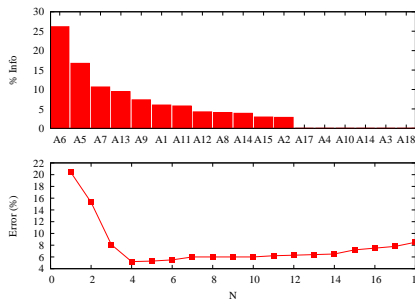
- $S, S_{\text{yes}}, S_{\text{no}}$: the initial, left and right samples
- $H_C(S)$: Shannon entropy of the class frequencies in S .

- For each variable, sum over all splits where this variable intervenes and average over all trees of the ensemble



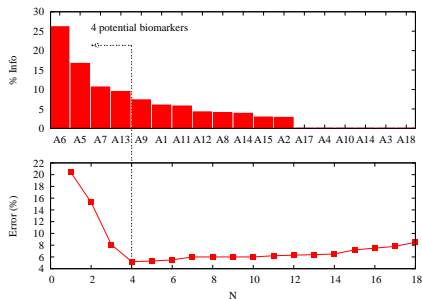
Variable Selection with Tree Ensembles

- Goal: determine how many variables among the top ranked are really necessary
 - 1 For $N=1$ to n :
 - build a model using only the first N ranked variables
 - estimate its error $Error(N)$
 - 2 Select N^* that minimizes $Error(N)$



Variable Selection with Tree Ensembles

- Goal: determine how many variables among the top ranked are really necessary
 - 1 For $N=1$ to n :
 - build a model using only the first N ranked variables
 - estimate its error $Error(N)$
 - 2 Select N^* that minimizes $Error(N)$



Application to the Diagnosis of Inflammatory Diseases

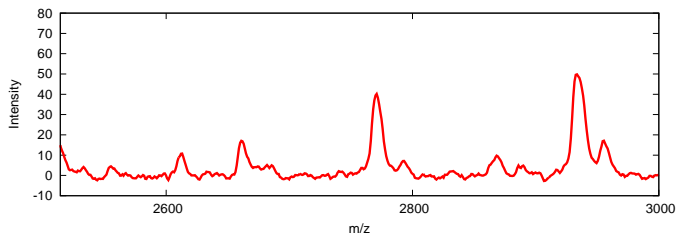
- Goals:
 - early diagnosis of rheumatoid arthritis (RA)
 - rapid and non invasive diagnosis of inflammatory bowel diseases (IBD)
- Each dataset is composed of three groups of patients (from the University Hospital of Liège):

	RA	IBD
Disease patients	34	60
Negative controls	29	30
Inflammatory controls	40	30
Total	103×2	120×4

- Mass spectra were obtained by SELDI-TOF mass spectrometry using several chip arrays
 - hydrophobic (H4),
 - weak cation-exchange (CM10),
 - and strong anion-exchange (Q10) surfaces
- Each spectrum is composed of about 15000 m/z values

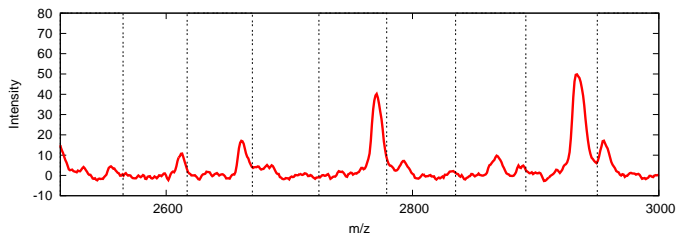
Pre-Processing: Spectrum Smoothing

- Goal: filter out noise on peak intensities and positions
- Proportional m/z values discretization yielded better results than more sophisticated peak alignment and filtering



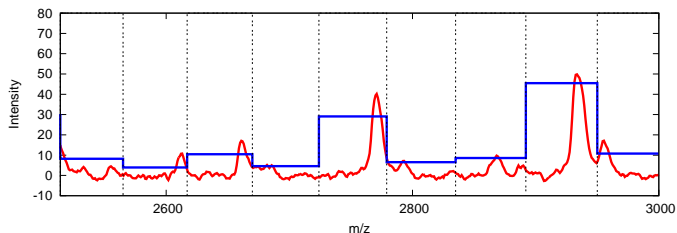
Pre-Processing: Spectrum Smoothing

- Goal: filter out noise on peak intensities and positions
- Proportional m/z values discretization yielded better results than more sophisticated peak alignment and filtering



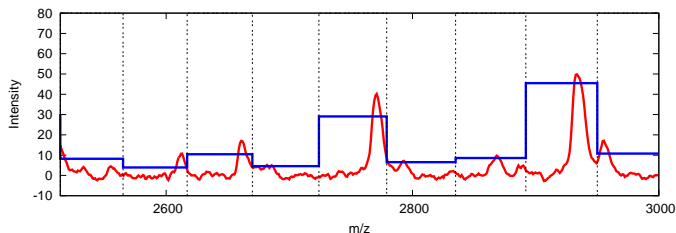
Pre-Processing: Spectrum Smoothing

- Goal: filter out noise on peak intensities and positions
- Proportional m/z values discretization yielded better results than more sophisticated peak alignment and filtering



Pre-Processing: Spectrum Smoothing

- Goal: filter out noise on peak intensities and positions
- Proportional m/z values discretization yielded better results than more sophisticated peak alignment and filtering

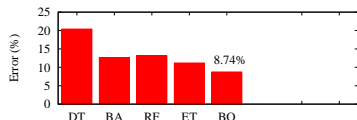
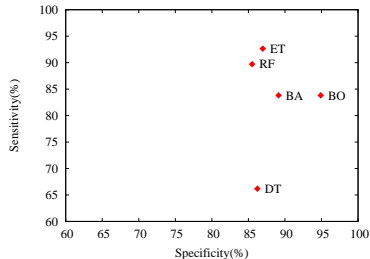


- e.g.: $r = 1\%$ \rightarrow from 15000 to about 300 variables

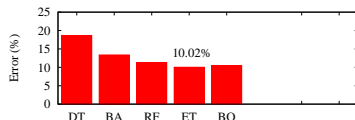
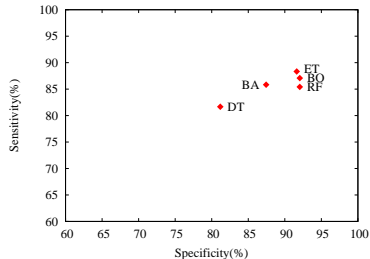
Accuracy Results

- Sensitivity, specificity, and error rates estimated by leave-one-out cross-validation (removing all spectra corresponding to a patient)

RA



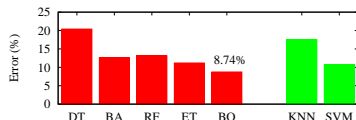
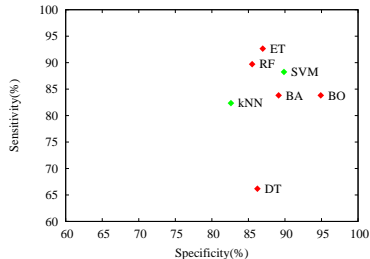
IBD



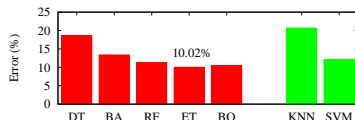
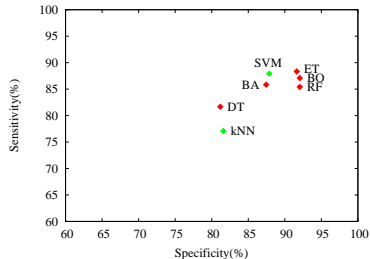
Accuracy Results

- Sensitivity, specificity, and error rates estimated by leave-one-out cross-validation (removing all spectra corresponding to a patient)

RA

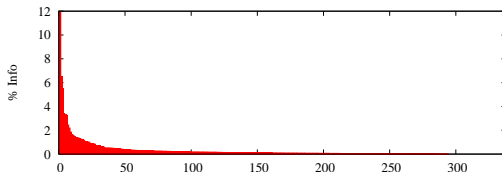


IBD

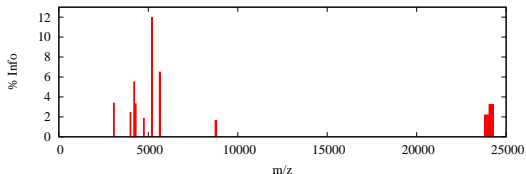


Variable Ranking

- Variable ranking with boosting on the IBD dataset

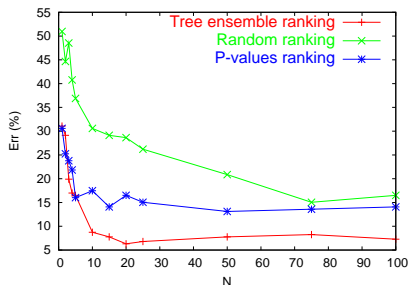


- The 10 most important variables in the spectrum:



- Tree based variable ranking compared with random variable selection and selection by p-values

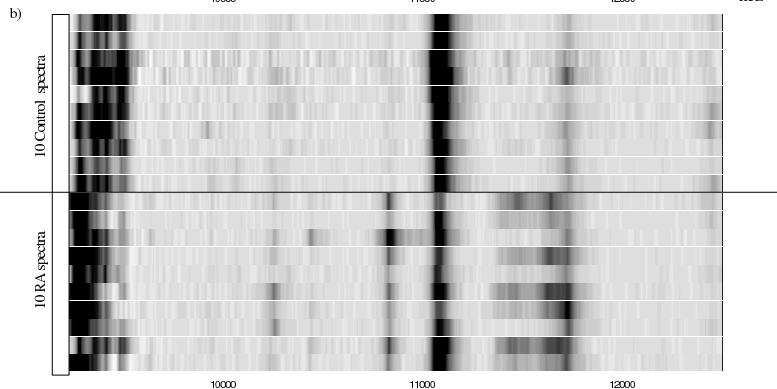
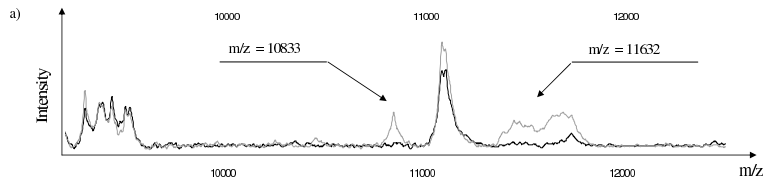
RA



IBD



Variable selection



Conclusion and Future Works

- Ensemble methods are very useful and powerful tools to analyse high throughput data in Biology
- We get competitive results with existing diagnostic tests for RA and new single and non invasive tests for IBD

- Future works
 - Application to genomic data (ongoing research on brain tumors)
 - Interpretable rules (e.g. decision trees on the selected biomarkers)
 - Validation of the model on independent patients (instead of leave-one-out)
 - Identification of the proteins associated to the biomarkers

Acknowledgements

Department of Electrical Engineering and Computer Science

Louis Wehenkel

Laboratory of Clinical Chemistry and Rheumatology

Marianne Fillet

Dominique de Seny

Marie-Alice Meuwis

Michel Malaise

Marie-Paule Merville

Funding

FNRS

Région Wallonne

ULG GIGA bioinformatics and proteomic platforms