FONDAZIONE
BRUNO KESSLER

A RANKING STABILITY INDICATOR IN BIOINFORMATICS

Giuseppe Jurman, Samantha Riccadonna, Roberto Visintainer
Giorgio Guzzetta, Cesare Furlanello

Predictive Models for Biomedicine and Environment
Fondazione Bruno Kessler
Trento
Italy

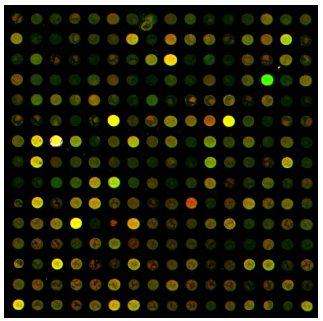
Cancer Bioinformatics Workshop

Cambridge Research Institute, 2nd-4th September 2010

Poster PB15

FONDAZIONE BRUNO KESSLER

THE PROBLEM



$\mathcal{M}(n \times p, \mathbb{R})$

$n \simeq 10^3$ samples

$p \simeq 10^5$ genes / 10^6 SNPs

A result of a classification/ranking procedure is a **list of biomarkers**, ranked according to their relevance for the classifier.

To ensure repeatability and not to incur in overfitting due to information leakage phenomena and such as the selection bias, methodology has to be carefully designed.

THE NEED FOR A BIOMARKERS STABILITY MEASURE...

1. ...for the resampling scheme of an experiment;
2. ...for the comparison of different schemes;
3. ...to guarantee reproducibility of the study.

Unlike the classifier case, no consolidated theory exists in literature for feature selection stability.

FONDAZIONE
BRUNO KESSLER

CITING BOULESTEIX & SLAWSKI...

Ranked gene lists are highly instable in the sense that similar measures of differential gene expression may yield very different rankings, and that a small change of the data set usually affects the obtained gene list considerably. Stability issues have long been under-considered in the literature, but they have grown to a hot topic in the last few years, perhaps as a consequence of the increasing skepticism on the reproducibility and clinical applicability of molecular research findings.

Stability and aggregation of ranked gene lists, Briefings in Bioinformatics 2009

Solution: the algebraic theory of the metric methods for permutation groups.

Using quotient groups, **module** awareness and extensions to **partial lists** can be automatically inherited.

FONDAZIONE BRUNO KESSLER

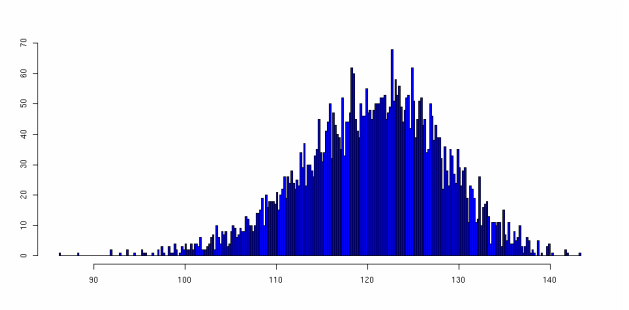
An algebraic approach

A key fact in biomarker lists is that variations in the **upper** part of the lists are much more relevant than differences in the **lower** part.

The Canberra distance

$$\mathrm{Ca}(\tau, \sigma) = \sum_{i=1}^{n} \frac{|\tau(i) - \sigma(i)|}{\tau(i) + \sigma(i)}$$

- Given a set of lists $\mathcal{L} = \{L_t\}_{t=1}^{b}$ of p features, the computation of all mutual distances leads to the construction of a symmetric **distance matrix** $M \in \mathcal{M}(b \times b, \mathbb{R}^+)$.
- Then the corresponding histogram can be built.
- The histogram can be approximated by a gaussian distribution (asymptotically proved).



$n = 100$, $p = 500$, $b = 100$, SVM-RFE, $D = \mathrm{Ca}$

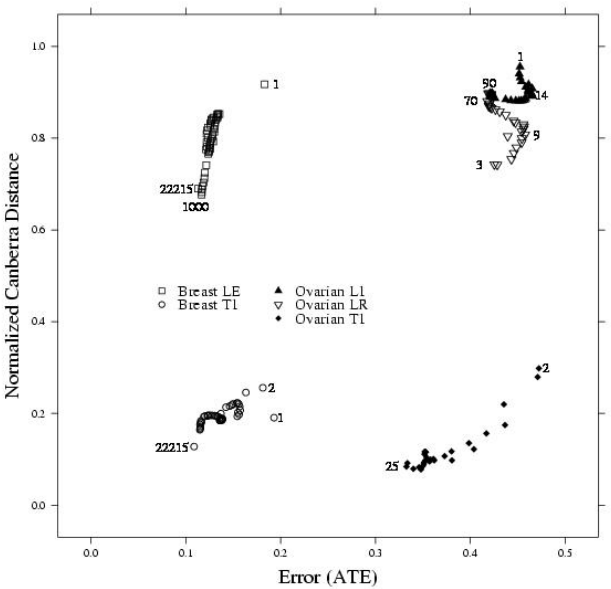
Indicators

It is thus possible to use the **mean** of the matrix M as measures of the mutual list distance in the set $\mathcal{L}$.

Accuracy and stability

Intuitively, the more mutually different the lists, the more unstable the problem (either because of the data or because of the employed classification procedure).
It is worthwhile to study a problem in the stability/accuracy space, to derive a Pareto-like front.

- Thus we can analyze a dataset in the accuracy vs. stability space.
- Left-down direction indicates better performance.
- This diagnostic plot allows the comparison of different datasets, different profiling methods (classifiers/feature ranking algorithm) and different models.



Error vs. stability indicator for different profiling methods (**L**SVM/**E**RFE,**L**SVM/**1**RFE,**T**RSVM/**1**RFE) and cancer datasets (Breast, Ovarian); each point corresponds to a feature subset size, indicated for extremal models

A **correlation between stability and predictivity** in the US-FDA led initiative MAQC-II has been detected both in training and validation sets: the more similar the signatures, the better the average predictions.