A comparative study of normalization and feature selection techniques for breast cancer prognosis from gene expression

Thibault Helleputte and Pierre Dupont

thibault.helleputte@uclouvain.be, pierre.dupont@uclouvain.be
Department of Computing Science & Engineering INGI - Université catholique de Louvain
UCL Machine Learning Group - www.ucl.ac.be/mlg

Microarrays for medical prognosis

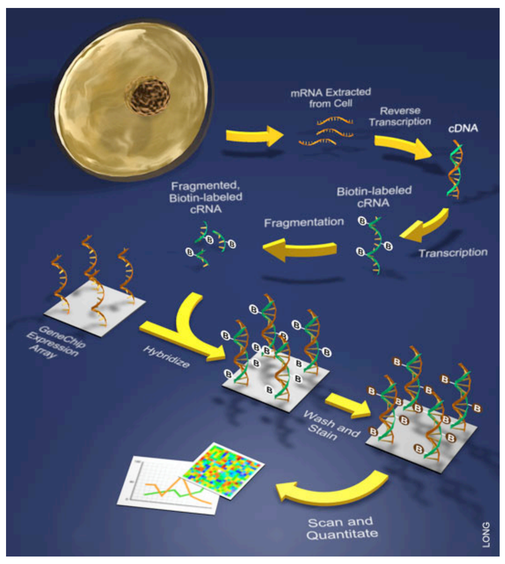
Microarrays measure expression of tens of thousands genes from an individual in a single experiment.

Typical microarray data:

	gene 1	gene 2	...	gene N	Pathology (class label)
sample 1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,N}$	y_1
sample 2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,N}$	y_2
...	...	...	...	...	...
sample M	$x_{M,1}$	$x_{M,2}$	...	$x_{M,N}$	y_M

In our case, samples are ***observations***, genes are ***features***

Given the high cost of this technology, only some tens of experiments may be led so **N>>M**, which is statistically a hard context for knowledge inference.



Prognosis: Feature Selection and Classification

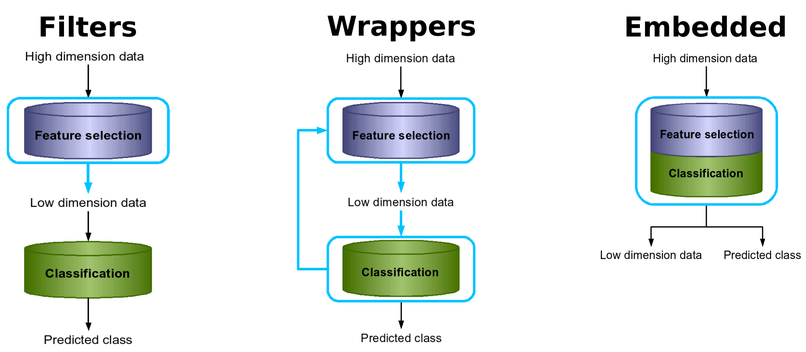
Double aim:

1. Identify a small subset of genes as pathology risks factors to launch a deeper medical research project and to evaluate treatment development feasibility.
2. Train a classifier based on these genes to be able to make a prognosis about the given pathology.

Acknowledgments:

1. UCL Intensive Computing and Mass Storage Institute (CISM)
2. Benjamin Haibe-Kains (ULB) for the Breast Cancer Data Sets

Feature Selection and Classification



A filter approach: t-Test feature ranking & SVM

Use t-Test to score features. P-Value is a measure of the probability to wrongly reject the hypothesis that two means are equal. Genes with lower p-Values are preferred to other genes with a higher p-Value.

Based on the n<<N top ranked features, a classifier may be built. For example, a Support Vector Machine.

SVM builds an hyperplane separating the training samples with the largest possible margin. Data can be mapped into a feature space with potentially higher dimensionality where their separation by a hyperplane may be easier.

Another filter approach: S2N feature ranking & SVM [4]

Use a Signal-to-Noise measure to score features. The S2N value is computed as:

$$S2N_j = \frac{\mu_{j_1} - \mu_{j_2}}{\sigma_{j_1} + \sigma_{j_2}}$$

where X_{j_c} is the X statistic computed on measurements of feature j for samples belonging to class c. The absolute value of S2N is generally taken.

Here also, a Linear SVM may be built based on the n<<N top ranked features.

A wrapper approach: Recursive Feature Elimination [1]

RFE is an algorithm that iteratively drops features reducing the SVM margin the least possible. In this case, a linear SVM is used. RFE is a wrapper approach because it uses the classifier to perform feature selection. It may be seen as an embedded method because it uses the classifier structure. However, the optimization of the classifier and the feature selection are two separate processes.

1. Train SVM: $max_\alpha W(\alpha) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$ sub. to $\alpha_i \geq 0 \; \forall \; i = 1, ..., M$ and $\sum_{i=1}^{M} \alpha_i y_i = 0$
2. Compute the weight vector: $\mathbf{w} = \sum_{i=1}^{M} \alpha_i y_i x_i$
3. Suppress the feature with the smallest $(w_i)^2$
4. Iterate 1-3 with the subset of the remaining features

Breast Cancer Data sets

All data sets are based on Affymetrix technology (HU133 A, B or Plus 2.0)

The Wang data set [1]

Number of samples (patients): **209** (80 distant metastasis vs. 129 metastasis-free, all ER+)
Number of features (genes): **22283**
Proportion of classes: 38.3% distant metastasis and 61.7% metastasis-free.

The Transbig data set [2]

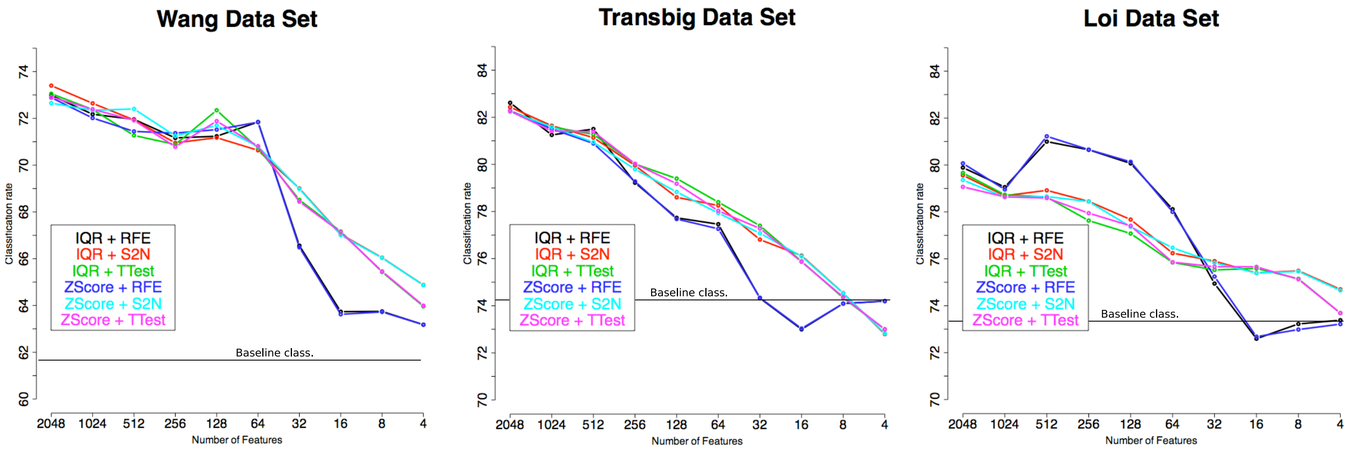
Number of samples (patients): **198** (51 distant metastasis vs 147 metastasis-free)
Number of features (genes): **22283**
Proportion of classes: 25.8% distant metastasis and 74.2% metastasis-free

The Loi data set [3]

Number of samples (patients): **255** (68 distant metastasis vs. 187 metastasis-free, all tamoxifen treated)
Number of features (genes): **44928**
Proportion of classes: 26.7% distant metastasis and 73.3% metastasis-free

Results

Classification rates are evaluated on **30 independant data samplings** using **Bootstrap 632**. For each data set and for each of the 6 methods tested, the **bootstrap samplings are the same**.



Standard Deviations on 30 runs range from 1 to 4 depending on the method and the number of features

Common features proportion among selected feature sets for different feature selection methods computed on Transbig data set with 10 runs:

8	TTest	S2N	RFE
TTest	1	0.662	0.087
S2N	0.662	1	0.112
RFE	0.087	0.112	1

32	TTest	S2N	RFE
TTest	1	0.8	0.197
S2N	0.8	1	0.222
RFE	0.197	0.222	1

128	TTest	S2N	RFE
TTest	1	0.884	0.294
S2N	0.884	1	0.309
RFE	0.294	0.309	1

2048	TTest	S2N	RFE
TTest	1	0.981	0.628
S2N	0.981	1	0.628
RFE	0.628	0.628	1

Prefiltering technique

The initial number of features is generally too high. Use of a prefiltering technique based on Interquartile Range:

Keep only features which have a IQR greater than the 0.75 quantile of all the data.

It generally allows quick selection of only about 10 - 15.000 features among 45.000 or 5 - 10.000 features among 25.000, depending on the data set.

Normalization techniques

ZScore Feature normalization

Each feature is a vector of gene expressions across patients: $f_j = \{x_{1,j}, x_{2,j}, ..., x_{M,j}\}$

Normalization of feature ***j*** is: $f_j^* = \frac{f_j - \mu_j}{\sigma_j}$

Motivation: give the same weight to all features.

IQR Feature normalization

For robustness concerns, median is used instead of the mean and the standard deviation is replaced by the interquartile range.

Conclusions

1. Global coherence across methods (similar trends)
2. TTest and S2N perform similarly, selecting nearly the **same features** (except for very few features).
3. TTest and S2N give globally the **same performances**.
4. The choice of IQR vs. ZScore normalization has nearly **no impact** on performances.
5. Feature Selection is not always beneficial on a strict classification accuracy point of view.
6. RFE could perform better by discarding only one feature at a time.

Perspectives

1. Test RFE by discarding one feature at a time below the threshold of 256 features [5]
2. Compare the tested methods with the one mentioned in [6].

Some References

[1] Wang et al. ***Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.*** Lancet 2005

[2] Desmedt et al. ***Strong Time Dependence of the 76-Gene Prognostic Signature for Node-Negative Breast Cancer Patients in the TRANSBIG Multicenter IndependentValidation Series.*** Clinical Cancer Research 2007.

[3] Loi et al. ***Definition of Clinically Distinct Molecular Subtypes in Estrogen Receptor-Positive Breast Carcinomas Through Genomic Grade.*** Journal of Clinical Onclology 2007.

[4] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield and E.S. Lander. ***Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.*** Science, 286:531–537, 1999.

[5] I. Guyon, J. Weston, S. Barnhill and V. Vapnik. ***Gene selection for cancer classification using support vector machines.*** Machine Learning, 46(1-3):389–422, 2002. Also: personnal communication with I. Guyon.

[6] J. Weston, A. Elisseef, B. Schölkopf and M. Tipping. ***Use of the zero-norm with linear models and kernel methods***. Journal of Machine Learning Research, 3:1439–1461, 2003.