Partially Supervised Feature Selection with linear Regularized Models

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 - http://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 ***samples***, 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 for further medical research or to evaluate treatment efficacy.
2. Train a classifier based on these genes to design a prognosis kit for the given pathology.

Partially Supervised Feature Selection

Motivations

1. Expert knowledge: The PSFS algorithm may be used to incorporate prior knowledge from field experts. For example, in a microarray experiment, a biologist may know/guess that some genes are likely more relevant. Those will get a higher prior during optimization.

2. This technique increases **stability** in feature selection. Stability is a desired property since the choice of the relevant dimensions in a problem should not be influenced (too much) by varying the data sampling.

3. Increasing **classification performances:** In some cases, this technique also increases classification performances.

PSFS Problem

Constrained zero-norm minimization approximation problem [1]:

$$\min_w \sum_{j=1}^{n} \ln(\epsilon + |w_j|)$$

Subject to: $y_i(\mathbf{w}.\mathbf{x}_i + b) \geq 1$

Elegantly solved by an iterative algorithm called L2AROM. Weight values corresponding to less useful features are rapidly vanishing. The selection process can be stopped at convergence, or when sufficient sparsity is obtained.

We propose to add a prior relevance β to each feature. The more a priori relevant the feature *j*, the higher β_j :

$$\min_w \sum_{j=1}^{n} \frac{1}{\beta_j} \ln(\epsilon + |w_j|)$$

Subject to: $y_i(\mathbf{w}.\mathbf{x}_i + b) \geq 1$

A constrained gradient descent technique is applied:

1. At step $k = 0$, initialize β $s.t.\ \beta_j \geq 1,\ \forall j \in \{1, n\}$ according to prior information. Set $w_{k,j} = \beta_j,\ \forall j \in \{1, n\}$.

2. Solve $\min_w ||w_j||_2^2$ subject to rescaled margin constraints:

$$y_i(\mathbf{w}.(\mathbf{x}_i * \mathbf{w}_k) + b) \geq 1$$

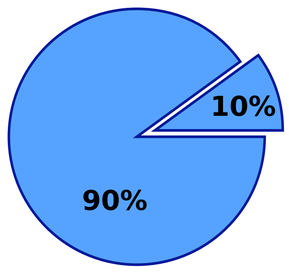
3. Let $\bar{\mathbf{w}}$ be the solution of step 2, set $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k * \bar{\mathbf{w}} * \beta$

4. Go to step 2 until convergence

Note: * denotes the component-wise product.

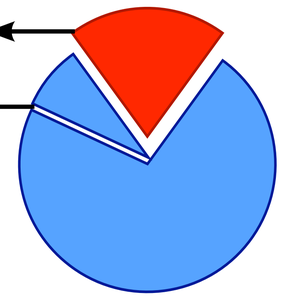
Experimental Design

Protocol 1: Real Knowledge



200 Splits 90%-10%:
1) Feature Selection on 90% with β=10 for a priori favored features and β=1 otherwise
2) Train SVM on 90% on selected features
3) Test on 10%
Average BCR and Stability

Protocol 2: Simulated Knowledge



10 Splits 80%-20%:
1) Select 50 features on 20%
2) On 80%, repeat 20 times:
- Select Features on 90% with β=10 for the 50 selected features and β=1 otherwise.
- Train SVM on 90% on selected features
- Test on 10%
Average BCR and Stability

Comparison of PS-I2-AROM, I2-AROM [1] , RFE [8], Golub Index [3], and Random Selection.

Classification Performances

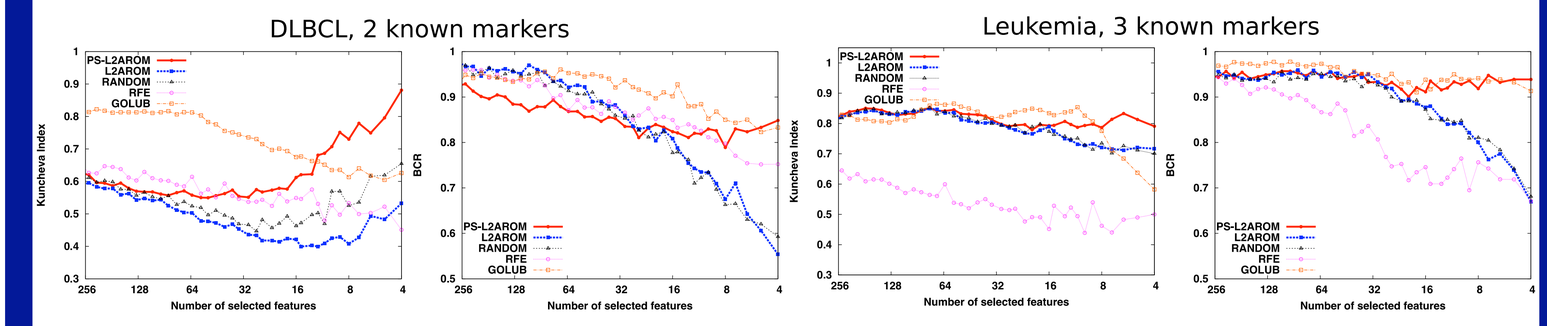
Unbalanced datasets: Balanced Classification Rate (BCR) instead of Accuracy: $BCR = \frac{1}{2}(\frac{TP}{P} + \frac{TN}{N})$

Stability

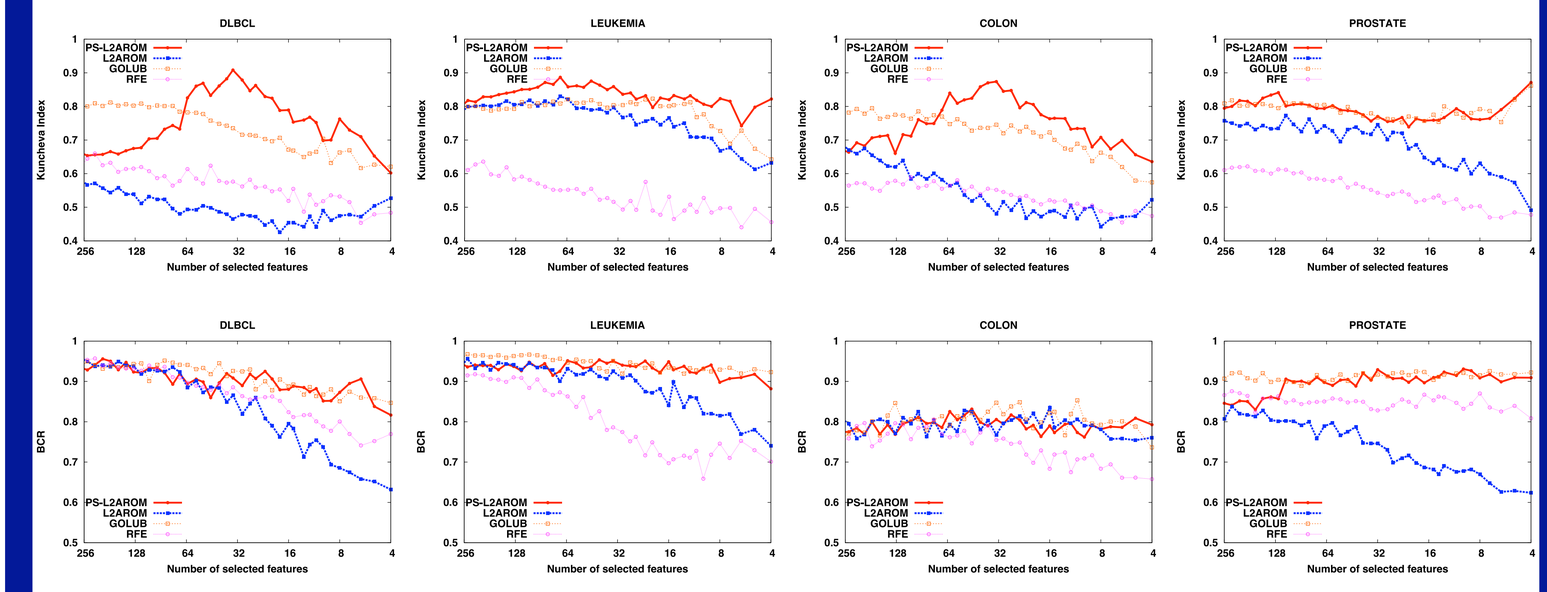
Robustness of the selected dimensions. Kuncheva Index [7]: $Stab = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \frac{I(Sig_i, Sig_j) - \frac{s^2}{n}}{s - \frac{s^2}{n}}$

where K is the number of selection rounds (here, $K = 200$ or 20), Sig is a signature (set of selected dimensions), I is the size of the intersection of two signatures, s is the size of the signatures and n is the total number of features.

Real Prior Knowledge



Simulated Knowledge



Conclusions

1. **PSFS allows to include prior knowledge on a priori important dimensions while letting the feature selection procedure depart from it.**
2. **PSFS naturally extends AROM methods [1].**
2. **Partial Supervision increases stability of selected features with respect to sampling variations.**
3. **Partial Supervision also improves classification performances in most cases.**
4. **Multivariate method: supervision of few dimensions influence the selection of other ones.**

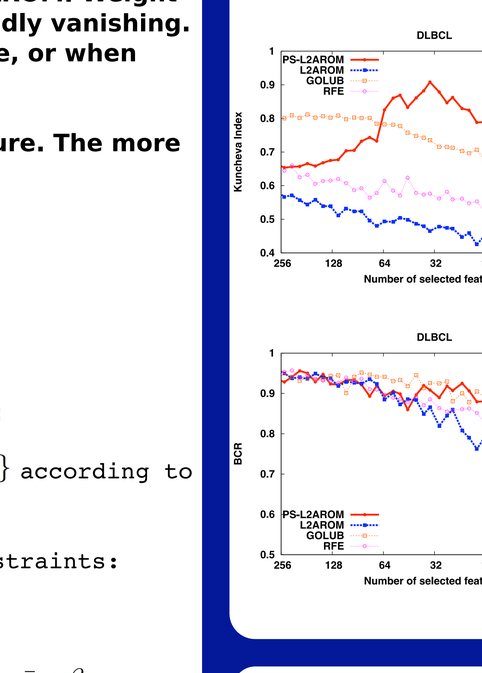
Data sets

Colon Cancer [2]
Samples (patients): 62 (40 tumor vs. 22 normal tissues)
Features (genes): 2000 (after non-specific filtering)
2 Classes: 64.5% tumor and 35.5% normal tissues

Leukemia [3]
Samples (patients): 72 (25 AML vs. 47 ALL tissues)
Features (genes): 7129
2 Classes: 34.7% AML and 65.3% ALL tissues

Prostate Cancer [4]
Samples (patients): 102 (52 tumor vs.50 normal tissue)
Features (genes): 6033
2 Classes: 51% tumor and 49% normal tissue

DLBCL [5]
Samples (patients): 77 (58 tissue 1 vs. 19 tissue 2
Features (genes): 7129
2 Classes: 75% tumor and 25% normal tissue



Main References

[1] 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.

[2] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack and A.J. Levine. ***Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays***. PNAS, 96:6745–6750, 1999.

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

[4] Singh D., Febbo P., Ross K., Jackson D., Manola J., Ladd C., Tamayo P., Renshaw A., D'Amico A., Richie J., Lander E., Loda M., Kantoff P., Golub T., and Sellers W. **Gene expression correlates of clinical prostate cancer behavior**. Cancer Cel l, 1(2) :203–209, March 2002.

[5] Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutok, J., Aguiar, R., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G., Ray, T., Koval, M., Last, K., Norton, A., Lister, T., Mesirov, J., Neuberg, D., Lander, E., Aster, J., & Golub, T. **Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** Nature Medicine, 8, 68–74, 2002.

[6] I. Kuncheva. **A stability index for feature selection**. International Multi-Conference on Artificial Intelligence and Applications. 2007.

[7] Y. Saeys, I. Inza and P. Larranaga. **A review of feature selection techniques in bioinformatics**. Bioinformatics 23:2507-2517, 2007

[8] 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.