Partially Supervised Feature Selection with Linear Regularized Models

Thibault Helleputte & Pierre Dupont
thibault.helleputte@uclouvain.be
http://www.ucl.ac.be/mlg

Machine Learning Group, Université catholique de Louvain

June 10, 2009

Microarray Data

Microarrays measure genes expression

	gene 1	gene 2	...	gene n	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

- Class labels come from external annotation.

Microarray Data

Microarrays measure genes expression

	gene 1	gene 2	...	gene n	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

- Class labels come from external annotation.
- With recent technology, $n \approx 55000$
- Very expensive technology, so $m \leq 300$

Microarray Data Analysis

Microarray data classification

- Diagnosis, Prognosis
- Clinical, Pharmaceutical applications

Microarray Data Analysis

Microarray data classification

- Diagnosis, Prognosis
- Clinical, Pharmaceutical applications

Feature Selection on Microarray data

Signature discovery:

- Explanatory concerns (no feature extraction)
- Diagnosis/Prognosis Kits
- May improve classification performances

SVM RFE

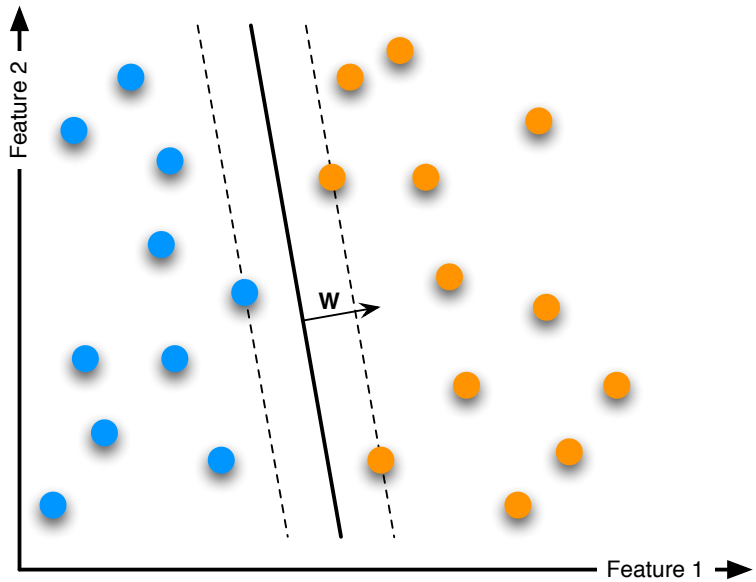
SVM generally show good classification performances and extensions for feature selection exist.

SVM RFE

SVM generally show good classification performances and extensions for feature selection exist.

RFE [Guyon et al., 2002]

- RFE iteratively trains a linear SVM and drops the features decreasing the less the margin.
- Embedded technique, using classifier structure



Zero-Norm Minimization

$$\min_{\mathbf{w}} ||\mathbf{w}||_0^0$$

$$\text{subject to: } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$$

$$\text{where } ||\mathbf{w}||_0^0 = card\{w_i | w_i \neq 0\}$$

- Elegant embedded formulation
- This problem has been shown to be NP-Hard
- Relaxations have been proposed...

AROM Methods [Weston et al., 2003]

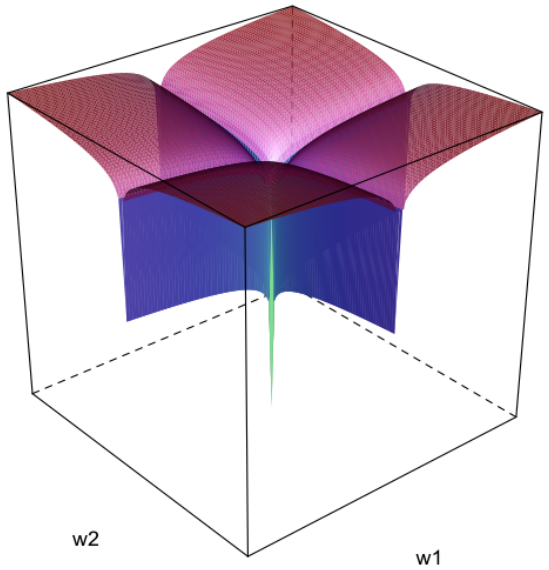
Previous problem solved with the following approximation:

Approximation to zeRO-norm Minimization

$\min_{\mathbf{w}} \sum_{j=1}^{N} ln(\varepsilon + |w_j|)$

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

where $0 < \varepsilon \ll 1$



l2-AROM Method

The previous problem leads to a nice algorithm:

l2-Approximation to zeRO-norm Minimization

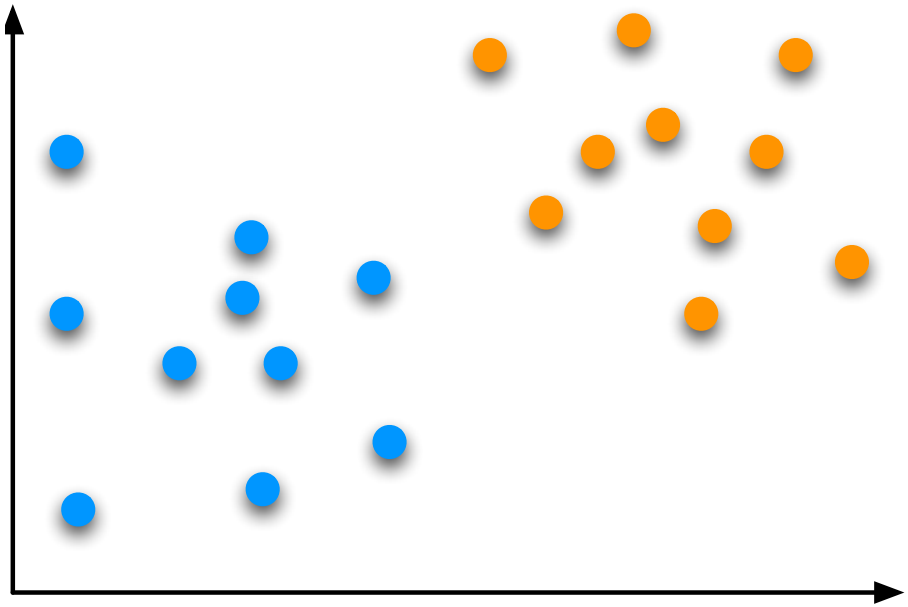
- At step $k = 0$, initialize $\mathbf{w}_k = (1, ..., 1)$
- Iterate until convergence:
  1. $\min_{\mathbf{w}} ||\mathbf{w}||_2^2$

     subject to: $y_i(\mathbf{w} \cdot (\mathbf{x}_i * \mathbf{w}_k) + b) \geq 1$
  2. Let $(\bar{\mathbf{w}})$ be the solution, set $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k * \bar{\mathbf{w}}$

Note: $*$ denotes component-wise product.

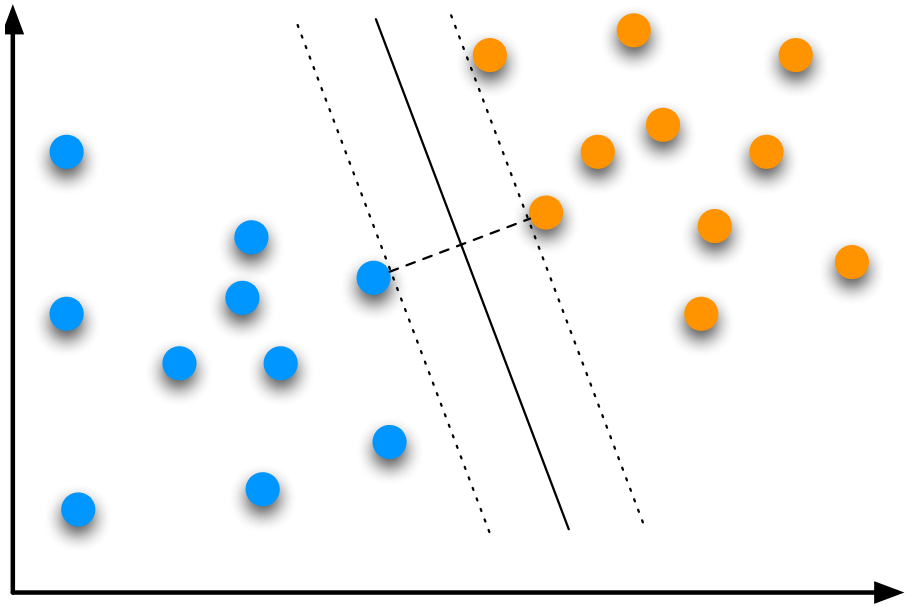
Problems with HD-Data analysis

When $m \ll n$: undetermined system, even with linear models!



Problems with HD-Data analysis

When $m \ll n$: undetermined system, even with linear models!



Regularization needed (Ex: max margin).
Still: Overfitting, lack of robustness.

Stronger inductive bias

- Need for stronger regularization / inductive bias
- Problem: where to find extra-information?

Stronger inductive bias

- Need for stronger regularization / inductive bias
- Problem: where to find extra-information?
- Ask the field experts.

Stronger inductive bias

- Need for stronger regularization / inductive bias
- Problem: where to find extra-information?
- Ask the field experts.

Prior Knowledge About Feature Relevance

- Field experts may know or guess that *some* features are likely to be more relevant
- Even if partial/insufficient for a complete model,...
- Even if imprecise,...
- ... it *is* extra knowledge

Partially Supervised Feature Selection

Partially Supervised Feature Selection

- PSFS = use of prior knowledge on feature relevance to bias feature selection.
- Full supervision on class labels

Partially Supervised Feature Selection

Partially Supervised Feature Selection

- PSFS $=$ use of prior knowledge on feature relevance to bias feature selection.
- Full supervision on class labels

Partially Supervised Selection vs. Semi-Supervised Classification

- Semi-Supervised Classification uses both labeled and unlabeled samples to build a classification model.
- PSFS $\neq$ Feature Selection techniques for Semi-Supervised Classification.

Partially Supervised AROM

- Relevance vector β
- Prior relevance of feature j encoded in β_j.
- The more (a priori) relevant feature j, the higher β_j.
- If no information on j, $\beta_j = 1$.

Partially-Supervized Approximation to zeRO-norm Minimization

$\min_{\mathbf{w}} \sum_{j=1}^{N} \frac{1}{\beta_j} ln(\varepsilon + |w_j|)$

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

where $0 < \varepsilon \ll 1$.

PS-*l*2-AROM Method

Partially-Supervized *l*2-Approximation to zeRO-norm Minimization

- At step $k = 0$, initialize $\mathbf{w}_k = \beta$
- Iterate until convergence:
  1. $\min_{\mathbf{w}} ||\mathbf{w}||_2^2$

     subject to: $y_i(\mathbf{w} \cdot (\mathbf{x}_i * \mathbf{w}_k) + b) \geq 1$
  2. Let $(\bar{\mathbf{w}})$ be the solution, set $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k * \bar{\mathbf{w}} * \beta$

4 Microarray Datasets

Data Set	Samples	Features	Priors	Ref.
DLBCL	77	7129	75%/25%	[Shipp et al. '02]
Leukemia	72	7129	65%/35%	[Golub et al. '99]
Prostate	102	6033	51%/49%	[Singh et al. '02]
Colon	62	2000	65%/35%	[Alon et al. '99]

Evaluation Metrics

Robustness: Stability Index [Kuncheva, 2007]

- Shared features among k signatures $\mathbf{S}$ of size s.
- Kuncheva Index: $K(\{\mathbf{S}_1, \ldots, \mathbf{S}_k\}) = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{|\mathbf{S}_i \cap \mathbf{S}_j| - \frac{s^2}{n}}{s - \frac{s^2}{n}}$

  $-1 < K \leq 1$, n is the total number of features and $\mathbf{S}_i$, $\mathbf{S}_j$ are two signatures.

Classification Performances: BCR

- Stability alone cannot characterize a signature quality.
- Balanced Classification Rate: $BCR = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right)$
- Unbalanced data: BCR preferred to accuracy.
- Average between *specificity* and *sensitivity*.

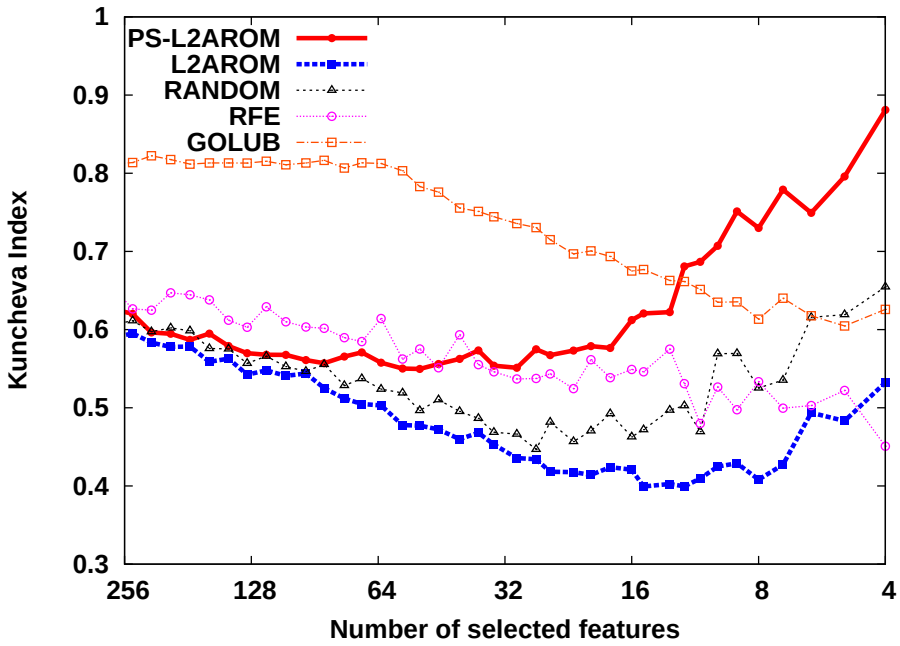
Protocol 1: Real Prior Knowledge

For DLBCL and Leukemia, 2-3 genes are used as clinical markers

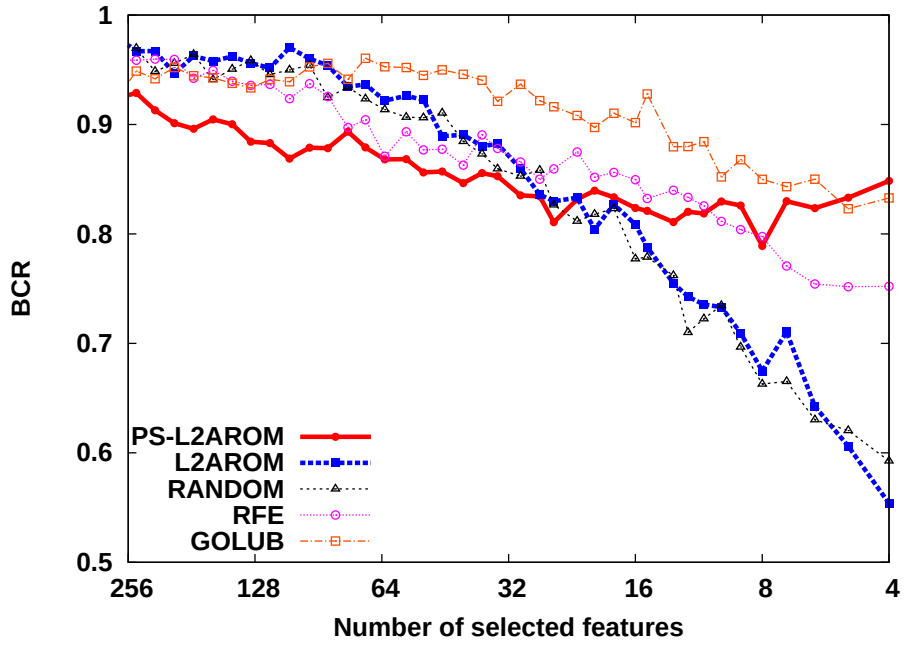
1. Set all β_j to 1, except those corresponding to used markers: $\beta_{markers} = 10$
2. Repeat 200 times:
   1. Split data into 90% train - 10% test
   2. Normalize - Select Feature - Build model on training part
   3. Evaluate BCR on test part
3. Average the BCRs and compute Stability (Kuncheva Index) on the 200 selected sets of features

DLBCL with 2 favored genes

Stability

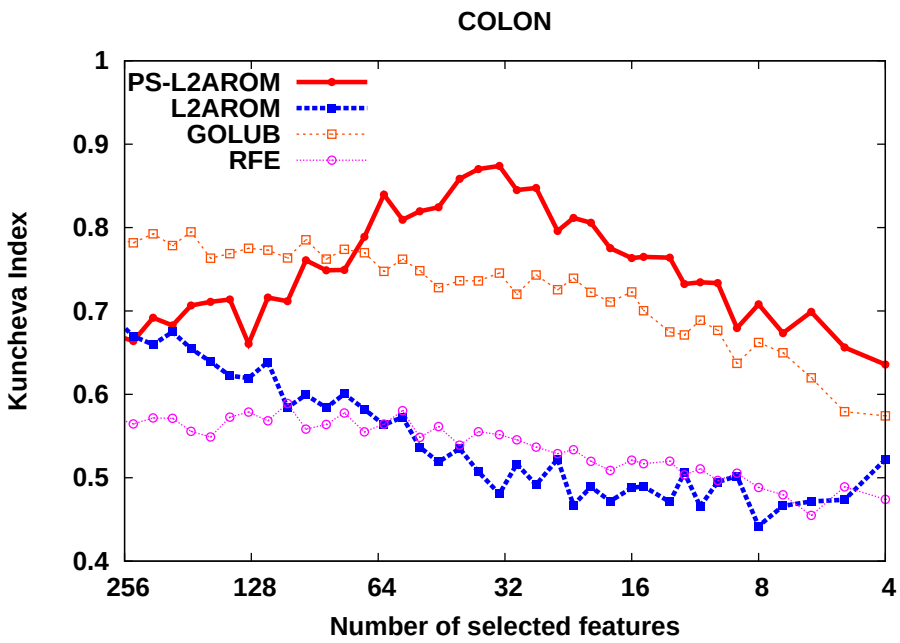


Classification Performances

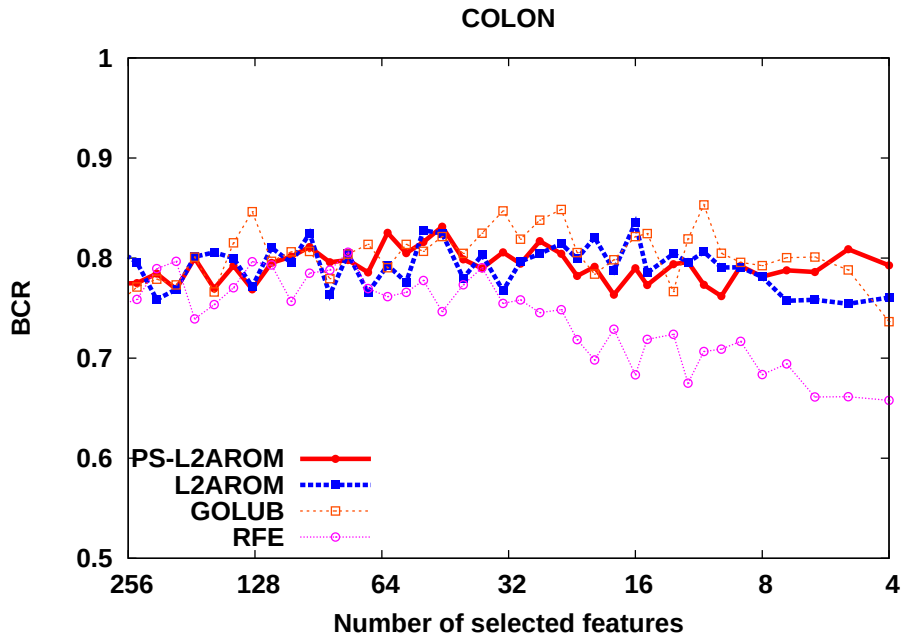


Colon with simulated knowledge

Stability



Classification Performances



Take Home Messages

- Stability should be considered for feature selection evaluation (but not alone).
- PSFS allows to include prior knowledge on a priori important dimensions while letting the feature selection procedure depart from it.
- PSFS naturally extends AROM methods.
- PSFS increases stability of selected features with respect to sampling variations.
- Partial Supervision also improves classification performances in most cases.
- Multivariate method: supervision of few dimensions influence the selection of other ones.

Thank you

RFE / AROM / PS-AROM

RFE, $l2$-AROM and PS-$l2$-AROM can be rephrased in a unified framework with different update rules for $\mathbf{w}$:

RFE

- $w_{i,k+1} \leftarrow w_{i,k} \ \forall i \mid \bar{w}_i \neq \min \bar{\mathbf{w}}$
- $w_{i,k+1} \leftarrow 0$ if $\bar{w}_i == \min \bar{\mathbf{w}}$

$l2$-AROM

- $w_{i,k+1} \leftarrow w_{i,k} \times \bar{w}_{i,k} \ \forall i$

PS-$l2$-AROM

- $w_{i,k+1} \leftarrow w_{i,k} \times \bar{w}_{i,k} \times \beta_i \ \forall i$

Transfer Learning