

Microarray Data Analysis

Feature Selection by Transfer Learning 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

September 10, 2009

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

Microarrays measure genes expression

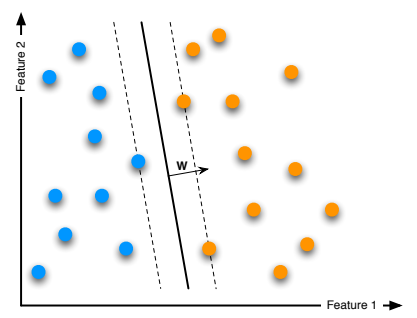
	gene 1	gene 2	...	gene n	class label
sample 1	$X_{1,1}$	$X_{1,2}$...	$X_{1,n}$	y_1
sample 2	$X_{2,1}$	$X_{2,2}$...	$X_{2,n}$	y_2
...
sample m	$X_{m,1}$	$X_{m,2}$...	$X_{m,n}$	y_m

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

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

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

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

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

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

AROM Methods [Weston et al., 2003]

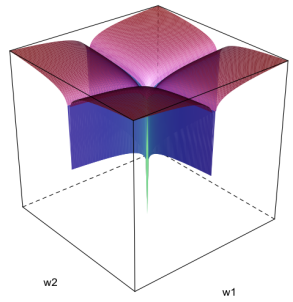
Previous problem solved with the following approximation:

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 $0 < \varepsilon \ll 1$



ℓ_2 -AROM Method

The previous problem leads to a nice algorithm:

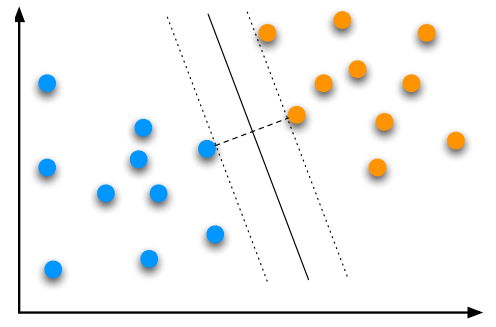
ℓ_2 -Approximation to zero-norm Minimization

- At step $k = 0$, initialize $\mathbf{w}_k = (1, \dots, 1)$
- 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

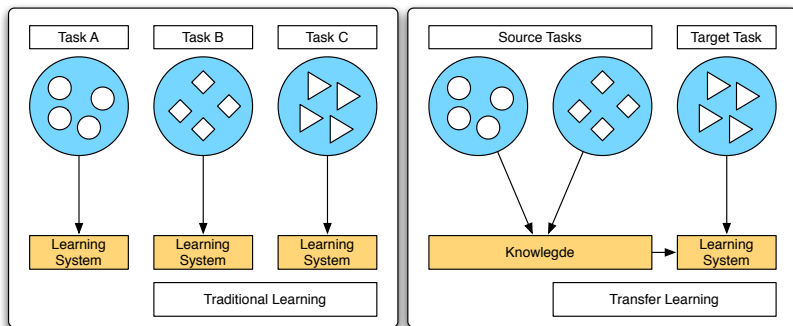
When $m \ll n$: undetermined system, even with linear models!



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

Partial Supervision
Stronger inductive bias

- Need for stronger regularization / inductive bias
- Problem: where to find extra-information?
- Transfer knowledge about feature relevance from similar datasets.



PS-AROM
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-Supervised Approximation to z_{RO} -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 $0 < \varepsilon \ll 1$.

Partial Supervision
Partially Supervised Feature Selection

Partially Supervised Feature Selection

- Helleputte & Dupont, ICML'09.
- PSFS = Incorporation of expert 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.

PS- l_2 -AROM Method
PS- l_2 -AROM Method

Partially-Supervised l_2 -Approximation to z_{RO} -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 $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k * \bar{\mathbf{w}} * \beta$

Datasets

3 Microarray Datasets

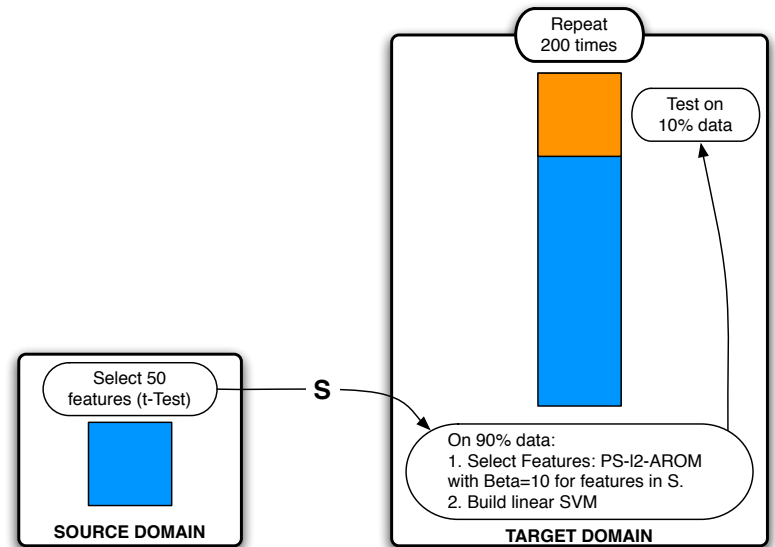
Use of 3 Prostate Cancer datasets.

Data Set	Normal/Tumor	Feat.	Technology	Ref.
Singh	50/52	12625	HGU95Av2	Singh '02
Chandran	18/86	12625	HGU95Av2	Chandran '99
Welsh	9/25	12626	HGU95A	Welsh '01

12600 probesets in common.

Evaluation

Protocol 1: Single Transfer



Evaluation Metrics

Robustness: Stability Index [Kuncheva, 2007]

- Shared features among k signatures \mathbf{S} of size s .

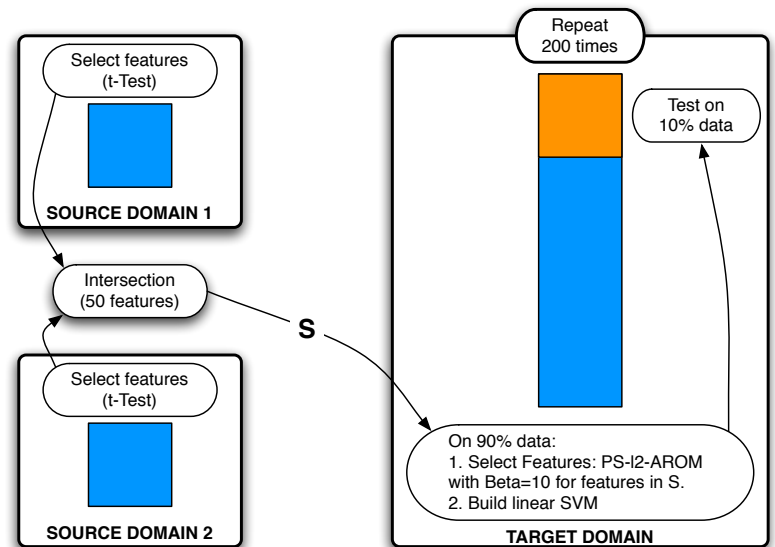
Kuncheva Index:
$$K(\{\mathbf{S}_1, \dots, \mathbf{S}_k\}) = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^k \frac{|\mathbf{S}_i \cap \mathbf{S}_j| - \frac{s^2}{n}}{s - \frac{s^2}{n}}$$
 $-1 < K \leq 1$, n is the total number of features and $\mathbf{S}_i, \mathbf{S}_j$ are two signatures.

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*.

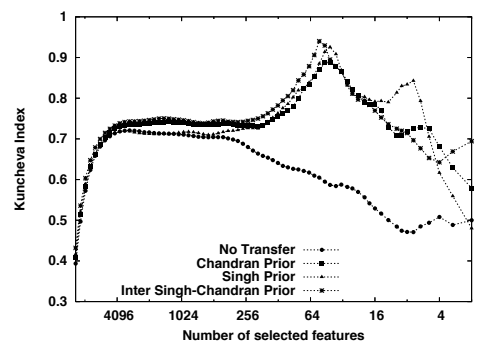
Evaluation

Protocol 2: Multiple Transfer

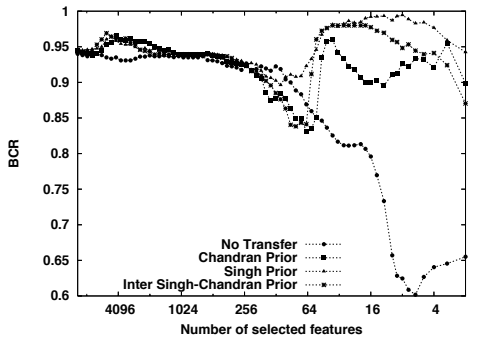


Results

Stability



Classification Performances

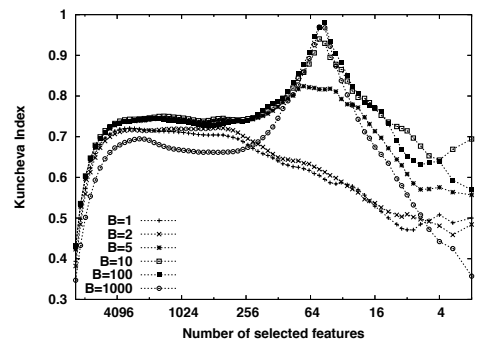


Take Home Messages

- Stability should be considered for feature selection evaluation (but not alone).
- PS- ℓ_2 -AROM can be used for inductive **transfer learning at the feature level**.
- Microarray datasets can be obtained from Gene Expression Omnibus at no cost.
- Transfer Learning via PS- ℓ_2 -AROM **improves classification performances** and **stability** of selected features with respect to sampling variations.
- Strong effect for **small signatures**.
- **Single** or **Multiple** sources domains.
- Method **insensitive** to the choice of B.

Sensitivity Analysis

Stability



Classification Performances

