Feature Selection by Transfer Learning with Linear Regularized Models [1]

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, rows 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

Multivariate Selection with l2-AROM [3]

Constrained zero-norm minimization approximation problem:

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

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.

PS-l2-AROM Algorithm [2]

We proposed 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.

Transfer Learning via PS-l2-AROM

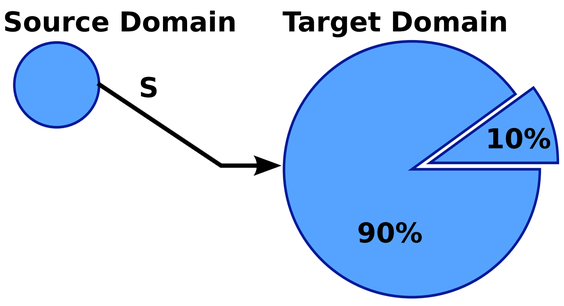
1. Microarray datasets generally offer too few samples. Leads to poor classification generalization performances and unstable feature selection. Getting more samples is too expensive.

2. Expert knowledge may be unavailable: The PS-l2-AROM algorithm may be used to incorporate prior knowledge from field experts, leading to higher stability and improved classification performances. In case no prior knowledge is available, use **similar datasets as information source.**

3. PS-l2-AROM can be used to **transfer information about feature relevance** from **source** dataset to **target** dataset. Selection is biased toward features that are relevant on source domains.

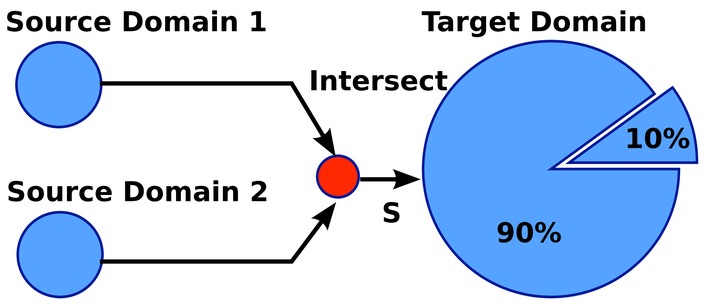
Experimental Setting

Single Transfer



Select 50 features (S) on source domain with t-Test

Multiple Transfer



Select features on source domains with t-Test such that their intersection is of size 50.

200 Splits 90%-10%:
1) Feature Selection with PS-l2-AROM on 90% data with β=10 for transfered features of S and β=1 otherwise
2) Train SVM on 90% data on selected features
3) Test on 10% data
Average BCR and Stability

Classification Performances

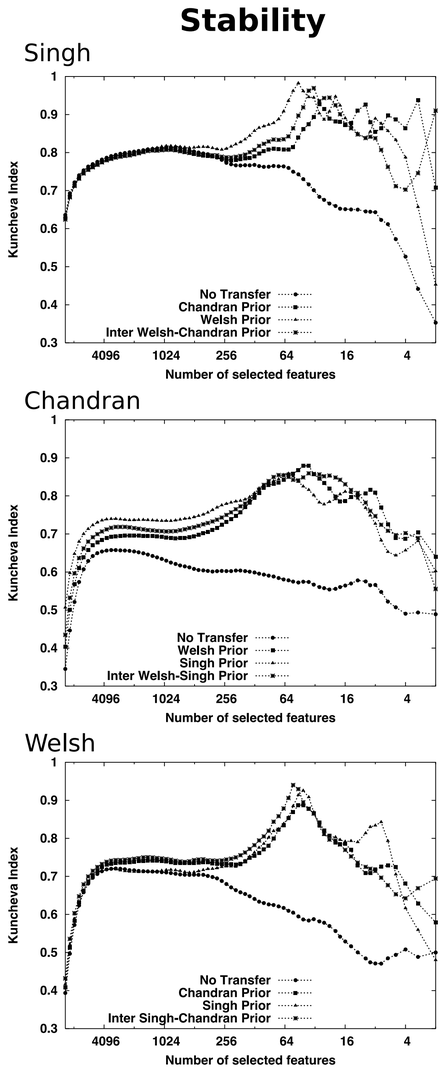
Unbalanced datasets: Balanced Classification Rate (BCR) instead of Accuracy: $BCR = \frac{1}{2}(\frac{TP}{P} + \frac{TN}{N})$

Stability

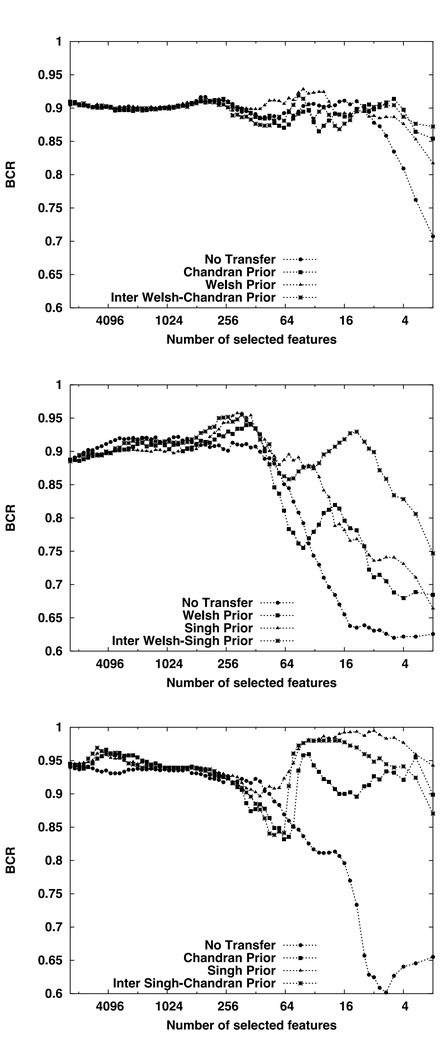
Robustness of the selected dimensions. Kuncheva Index [7]: $KI = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{|S_i \cap S_j| - \frac{s^2}{n}}{s - \frac{s^2}{n}}$

where k is the number of selection rounds ($k = 200$), S_i and S_j are two signatures (sets of selected dimensions), s is the size of the signatures and n is the total number of features.

Stability

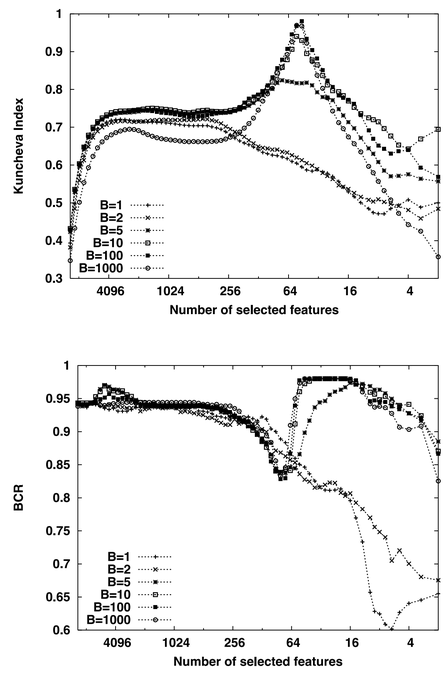


Classification Perfs.



Sensitivity Analysis

Impact of β parameter on stability and classification performances on Welsh



Conclusions

1. **PS-l2-AROM [2] can be used to perform inductive Transfer Learning at the feature level. Microarray Datasets can be obtained from Gene Expression Omnibus at no cost.**
2. **Transfer via PS-l2-AROM improves both stability and classification performances, specially for small signatures.**
3. **Transfer can be performed from one or multiple source domains.**
4. **The method is insensitive to the choice of β for a large range of values.**

Prostate Data sets

Singh [4]

Samples (patients): 102 (52 tumor vs. 50 normal tissues)
Features (genes): 12625
Technology: Affymetrix HG-U95Av2

Chandran [5]

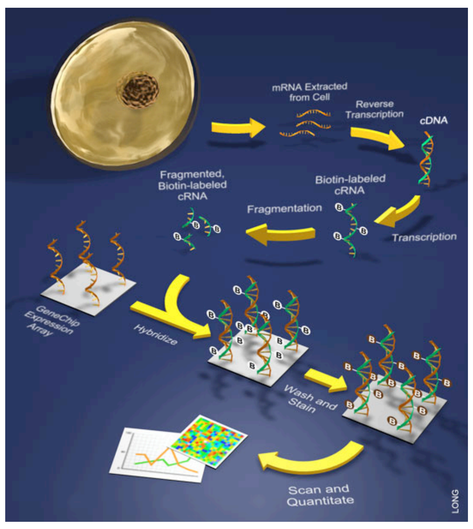
Samples (patients): 104 (86 tumor vs. 18 normal tissues)
Features (genes): 12625
Technology: Affymetrix HG-U95Av2

Welsh [6]

Samples (patients): 34 (25 tumor vs.9 normal tissue)
Features (genes): 12626
Technology: Affymetrix HG-U95A

Common feature set

Datasets are restricted to the 12600 common features they share.



Main References

[1] Helleputte, T. and Dupont, P. **Feature Selection by Transfer Learning with Linear Regularized Models.** Lecture Notes in Artificial Intelligence, No. 5781, Springer, European Conference on Machine Learning (ECML/PKDD), Bled, Slovenia, September 7-11, pp. 533-547, 2009.

[2] Helleputte, T. and Dupont, P. **Partially supervised feature selection with regularized linear models.** Proceedings of the 26th International Conference on Machine Learning. 2009

[3] 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.

[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] Chandran, U., Ma, C., Dhir, R., Bisceglia, M., Lyons-Weiler, M., Liang, W., Michalopoulos, G., Becich, M., Monzon, F. **Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process.** BMC Cancer 7(1) (2007) 64.

[6] Welsh, J.B., Sapinoso, L.M., Su, A.I., Kern, S.G., Wang-Rodriguez, J., Moskaluk, C.A., Frierson, Henry F., J., Hampton, G.M. **Analysis of Gene Expression Identifies Candidate Markers and Pharmacological Targets in Prostate Cancer.** Cancer Res 61(16) (2001) 5974–5978

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

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