# Biomarker Selection by Transfer Learning with Linear Regularized Models

Thibault Helleputte[1,2] and Pierre Dupont[1,2]

University of Louvain,
Computing Science and Engineering Dept. & Machine Learning Group,
Place Sainte Barbe 2, B-1348 Louvain-la-Neuve, Belgium
{Thibault.Helleputte, Pierre.Dupont}@uclouvain.be

This work presents a novel feature selection method for classification of high dimensional data, such as those produced by microarrays. Classification of such data is challenging, as it typically relies on a few tens of samples but several thousand dimensions (genes). The number of microarray chips needed to obtain robust models is generally orders of magnitude higher than most datasets offer. The number of available datasets is however continuously rising, for example in databases like the NCBI's Gene Expression Omnibus (GEO). Building a large microarray dataset consisting of the simple juxtaposition of independent smaller datasets is difficult or irrelevant due to differences either in terms of biological topics, technical constraints or experimental protocols.

*Biomarker selection* specifically refers to the identification of a small set of genes, a *signature*, related to a pathology or an observed treatment outcome. The lack of robustness of biomarker selection has been outlined. In the context of biomarker selection from microarray data, a high stability means that different subsets of patients lead to very similar signatures and is a desirable property. The biological process explaining the outcome is indeed assumed to be mostly common among different patients.

Our feature selection technique includes a partial supervision (PS) to smoothly favor the selection of some dimensions (genes) on a new *target* dataset to be classified. The dimensions to be favored are previously selected with a simple univariate technique, like a $t$-test, from similar *source* datasets, for example from GEO, hence performing inductive transfer learning at the feature level. We rely here on our recently proposed PS-$l2$-AROM method, a feature selection approach embedded in a regularized linear model. This algorithm reduces to linear SVM learning with iterative rescaling of the input features. The scaling factors depend here on the selected dimensions on the source domains. The proposed optimization procedure smoothly favors the pre-selected features but the finally selected dimensions may depart from those to optimize the classification objective under rescaled margin constraints.

Practical experiments on several microarray datasets illustrate that the proposed approach not only increases classification performances, as usual with sound transfer learning scheme, but also the stability of the selected dimensions with respect to sampling variation. It is also shown that multiple transfer from various source datasets can bring further improvements.