Variance Estimators for t-Test Ranking Influence the Stability and Predictive Performance of Microarray Gene Signatures

Nizar Thouleimat, Daniel Hernández-Lobato and Pierre Dupont

Machine Learning Group, ICTEAM Institute, Université catholique de Louvain

Université catholique de Louvain UCL

Introduction

A Student t-test is a standard statistical approach for ranking differentially expressed genes from microarray data. However, the variance estimator used in this test is unreliable when the number of data instances is very small. Shrinkage t-test and Window t-test have been suggested as practical alternatives to improve the estimation of the variance under these circumstances. The choice of the variance estimator is known to influence the gene ranking. We study here its effect on the classification performance of predictive models built from the resulting gene signatures. We further assess the stability of these gene selection methods with respect to different subsamplings of the data.

Datasets description

Dataset	Classification Task	Patients per Class	Ref.
Colon cancer 1	Normal/Tumor	22/40	[3]
Colon cancer 2	Normal/Tumor	22/25	[4]
Breast cancer	Normal/Tumor	43/43	[5]
Prostate cancer	Normal/Tumor	52/50	[6]
Leukemia 1	Subtype 1/Subtype 2	25/47	[7]
Leukemia 2	Subtype 1/Subtype 2	37/42	[8]
Lymphoma	Subtype 1/Subtype 2	22/23	[9]

Student t-test, Shrinkage t-test and Window t-test

- General formulation of the t-test statistic for gene g:

$$t_g = \frac{m_1 - m_2}{\sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}} \sim \text{Student Distribution}\,,$$

where m_1 and s_1 are the mean and the variance of the first group of patients, of size n_1, and m_2 and s_2 are the mean and the variance of the second group of patients, of size n_2. The degrees of freedom are computed by using the typical Welch-Satterthwaite equation.

- Estimators of the Variance for each group of patients $j \in \{1, 2\}$:
  - Student t-test:

$$s_j = \hat{\sigma}_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x(g)_i - m_j)^2\,.$$

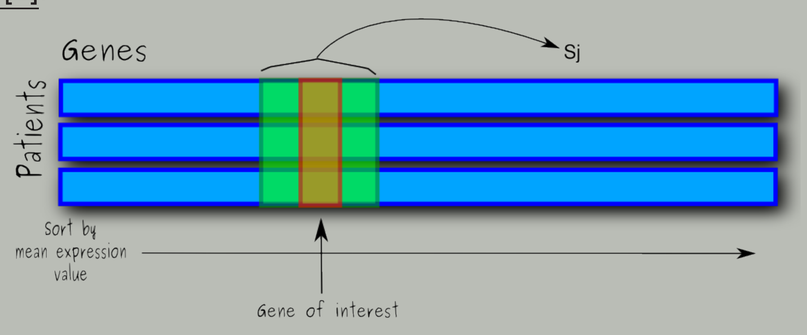
  - Shrinkage t-test [1]:

$$s_j = \lambda \overline{\sigma}_j^2 + (1 - \lambda)\hat{\sigma}_j^2\,,$$

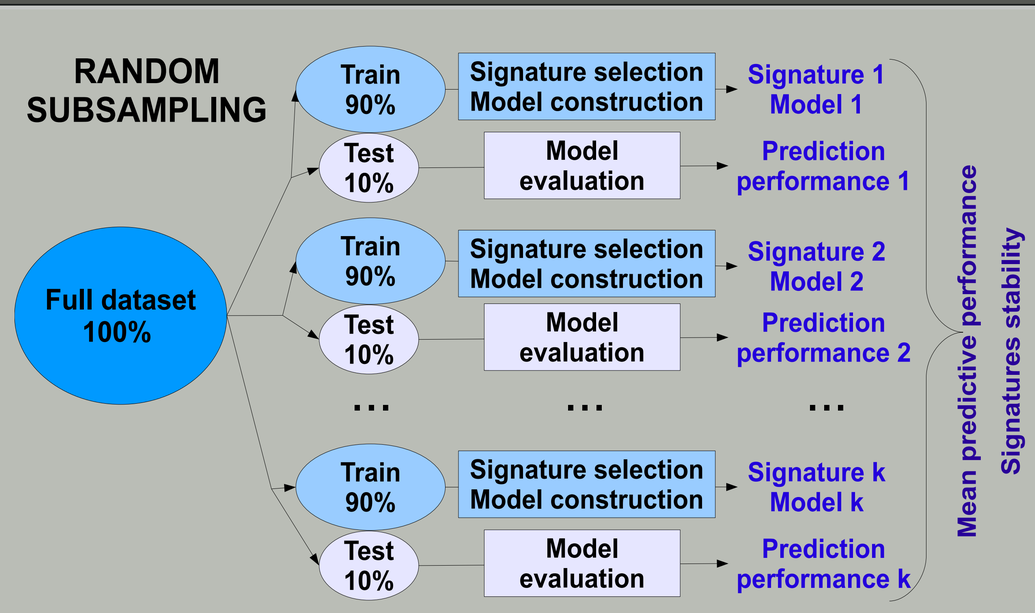
where $\overline{\sigma}_j^2$ is the median of $\hat{\sigma}_j^2$ across the different genes g and

$$\lambda = \min\left(1, \frac{\sum_{g=1}^{p} \text{Var}(\hat{\sigma}_j^2)}{\sum_{g=1}^{p} \left(\hat{\sigma}_j^2 - \overline{\sigma}_j^2\right)^2}\right)\,.$$

  - Window t-test [2]:



Experimental Protocol



Signature selection and model construction:

1. Gene ranking.
2. Select s top ranked genes.
3. Build model.

Model evaluation:

1. Predict class for each test sample (**nearest centroid** classifier with Pearson correlation metric).
2. Calculate prediction performance.

Prediction Performance

Balanced Classification Rate: $BCR = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right)$ (avg. between specificity and sensitivity)

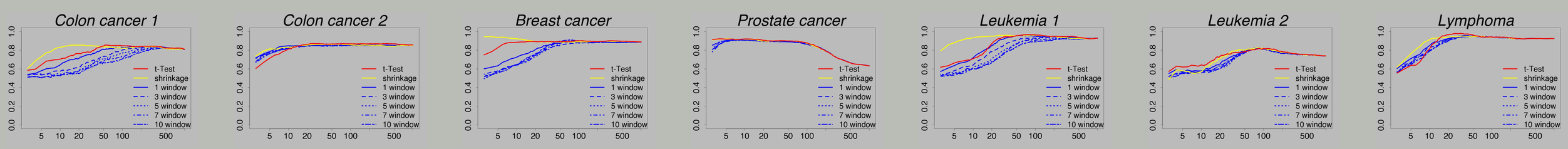
Signature stability

Kuncheva Index [10]: $\text{Stab} = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \left(|\text{Sig}_i \cap \text{Sig}_j| - \frac{s^2}{n}\right) / \left(s - \frac{s^2}{n}\right)$

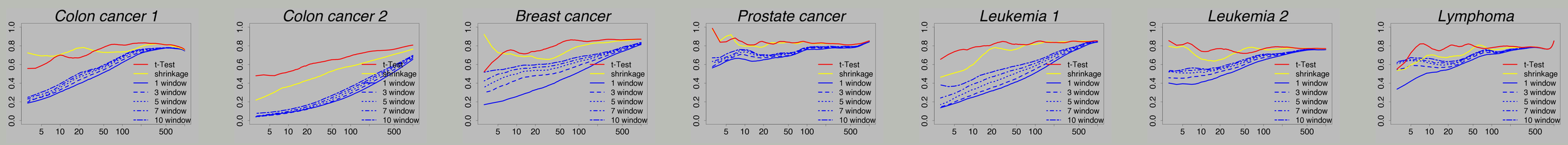
where K is the number of data partitions (**200**), Sig_i is the signature for partition i, s is the size of the signatures and n is the total number of genes.

Experimental Results

Predictive performances (BCR) as a function of gene signature size:



Signature stability (Kuncheva Index) as a function of gene signature size:



Conclusions

The correction of the variance estimate in the t-test may be beneficial in terms of the predictive performance of the models built on the corresponding gene signatures. For signature sizes with less than 50 genes our results show that the Shrinkage t-test provides similar or better balanced classification rates (BCRs) than the standard t-test or the Window t-test. For signatures of 10 genes, a Wilcoxon test rejects the hypothesis of all methods having equal BCRs (p-values $< \mathbf{2.5 \cdot 10^{-8}}$) for the datasets *Breast Cancer*, *Colon Cancer* and *Leukemia 1*. We also find that the optimal window size for the Window t-test should be as small as possible, thus questioning the interest of this variant to obtain good predictive models.

References

[1] Opgen-Rhein R., Strimmer K. **Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach**. *Statistical Applications in Genetics and Molecular Biology, 2007, 6(9).*
[2] F. Berger, *et al.* **The "Window t test": a simple and powerful approach to detect differentially expressed genes in microarray datasets**. *Central European Journal of Biology, 2008, 3(3).*
[3] U. Alon, *et al.* **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays**. *PNAS, 1999, 96(12).*
[4] N. Ancona, *et al.* **On the statistical assessment of classifiers using DNA microarray data**. *BMC Bioinformatics, 2006, 7.*
[5] I.B. Pau Ni, *et al.* **Gene expression patterns distinguish breast carcinomas from normal breast tissues: the Malaysian context**. *Pathology - Research and Practice, 2010, 206(4).*
[6] D. Singh, *et al.* **Gene expression correlates of clinical prostate cancer behavior**. *Cancer cell, 2002, 1(2).*
[7] T.R. Golub, *et al.* **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science, 1999, 286(5439).*
[8] S. Chiaretti, *et al.* **Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival**. Blood, 2004, 103(7).
[9] A. Alizadeh, *et al.* **Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling**. *Nature, 2000, 403(3), 503-511.*
[10] Kuncheva, L. I., **A stability index for feature selection**. Proceedings of the 25th IASTED International Multi-Conference on Artificial Intelligence and Applications, ACTA Press, 2007, 390-395.

http://www.ucl.ac.be/mlg/ {mohamed.touleimat,daniel.hernandez-lobato, pierre.dupont}@uclouvain.be