LETTER TO THE EDITOR

Independent test set performance in the prediction of early relapse in ovarian cancer with gene expression profiles

Frank De Smet[1], Nathalie L.M.M. Pochet[1], Toon Van Gorp[2], Dirk Timmerman[2], Bart L.R. De Moor[1], Ignace B. Vergote[2]

[1]Department of Electrical Engineering ESAT-SCD, K.U.Leuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium

[2]Department of Obstetrics and Gynecology, Division of Gynecologic Oncology, University Hospitals, K.U.Leuven, Herestraat 49, 3000 Leuven, Belgium

Running Title: Independent test set performance

Key Words: prediction, independent test set, microarrays, ovarian cancer, chemotherapy

Corresponding author:

Frank De Smet, MD, MscEng, PhD

Department of Electrical Engineering ESAT-SCD, K.U.Leuven

Kasteelpark Arenberg 10

3001 Leuven-Heverlee

Belgium

Tel: +32-16-328644; Fax: +32-16-321970; E-mail: frank.desmet@esat.kuleuven.be

---

Grant support: KUL PhD/postdoc grants; KUL-GOA AMBioRICS; FWO: G.0115.01, G.0407.02, G.0388.03; BFSPO-IUAP P5/22; EU-RTD: FP6-NoE Biopattern;

To the Editor:

With great interest we read the article by Hartmann et al. (1) investigating whether it is possible to apply gene expression patterns in order to discriminate between ovarian tumors with early and late relapse after platinum-paclitaxel combination chemotherapy. Among others, the authors claim to have derived a 14-gene predictive model with an independent test set accuracy of 86% and a positive predictive value of 95%.

However, after examination of the data analysis strategy of Hartmann et al., we noticed that the test set has been used to perform prior model selection and therefore cannot be called independent. Summarized and after data preprocessing, the authors constructed 100 support vector machine (SVM) models each based on a set of genes with the highest signal-to-noise ratio derived from a random selection (70% of 51 training samples) of the training set. Subsequently, these 100 models were all tested on the (wrongfully called independent) test set (28 samples) and the top model with the fewest prediction errors was selected and reported.

Unfortunately, this selection implies that information from the test set was used to choose a model that optimally fits this particular test set but might perform worse on another and independently chosen test set. As a consequence the reported performance indices might be overestimated and will probably be impossible to reproduce on new prospective data. In our experience and due to the high-dimensional nature of microarray data, even the slightest use of a so called independent test set (or the use of the left-out samples in cross-validation studies (2)) within the model building process will dramatically overestimate the performance of a classifier based on expression patterns. After model selection and in order to obtain a realistic estimate of the true performance, it is therefore imperative to test a new model on completely independent and prospective data (3).

In order to substantiate our claims, we implemented a similar data analysis scheme in MATLAB (Release 13 – script can be obtained on request) based on 14-gene SVM models from LS-SVMlab (Version 1.5 – http://www.esat.kuleuven.ac.be/sista/lssvmlab/) (4, 5). We subsequently applied our script on 10 randomly generated data sets each subdivided in a training and test set (expression levels uniformly and independently drawn between 0 and 1) with the same dimensions and composition as reported by Hartmann et al. For a true independent test set and since the random data does not contain any information about the process under study, one could expect an accuracy around

50%. However, the 10 test set accuracies returned by our MATLAB script (one for each training + test set) ranged between 71.43% and 82.14% and were significantly (P = 0.002; sign test) different from 50%. Therefore, these results indicate that the procedure described in Hartmann et al. strongly overestimates the accuracy that can be expected on independent data. Also noteworthy was the observation that the accuracies on the (in this case truly independent) test set indeed varied around a mean of about 50% if the model selection step was omitted. In the latter case we considered all 1000 models (100 models for each random data set) and not only the 10 models selected for their optimal performance on the test set.

Finally, we want to mention that Hartmann et al. stated that the reported accuracy of 86% was very unlikely to occur by chance alone. This was – similarly as above – assessed by comparing this result with a series of test set accuracies obtained through random models (in this case generated by randomly permuting the outcome labels of the training set). However, this assessment only indicates that the reported accuracy is relatively better than the test set accuracies of the random models. Since our simulation showed that these values themselves are overestimated, this evaluation does not say anything about the validity of the absolute value of the reported accuracy of 86%. Nevertheless, this assessment seems to indicate that the expression patterns indeed contain information about the time of relapse after chemotherapy.

In summary, we believe that the magnitude of the performance indices of the 14-gene model derived by Hartmann et al. will not be confirmed on a truly independent test set. In our opinion and in the absence of new prospective data to properly assess the current model, we believe that model training should be repeated using a method that refrains from exploiting any information from the test set. Only under these circumstances it is possible to correctly estimate the test set performance. Nowadays the authors have the choice between a wide variety of suitable classification methods that have been specifically developed for expression data and that are publicly available (as an example see Pochet et al. and Tibshirani et al. (3, 6)).

Frank De Smet
Nathalie Pochet
Bart De Moor
Department of Electrical Engineering ESAT-SCD
K.U.Leuven
Leuven-Heverlee, Belgium

Toon Van Gorp
Dirk Timmerman

Ignace Vergote
Department of Obstetrics and Gynecology, Division of Gynecologic Oncology
University Hospitals, K.U.Leuven
Leuven, Belgium

References

1. Hartmann LC, Lu KH, Linette GP, et al. Gene expression profiles predict early relapse in ovarian cancer after platinum-paclitaxel chemotherapy. Clin Cancer Res 2005;11:2149-55.
2. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14-8.
3. Pochet NL, Janssens FA, De Smet F, Marchal K, Suykens JA, De Moor BL. M@CBETH: a microarray classification benchmarking tool. Bioinformatics Advance Access, 12 May 2005; doi:10.1093/bioinformatics/bti495.
4. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least Squares Support Vector Machines. Singapore: World Scientific; 2002.
5. Pochet N, De Smet F, Suykens JA, De Moor BL. Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics 2004;20:3185-95.
6. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002;99:6567-72.