A new performance criterion for classification methods for microarray gene expression data

Cristina Botella*, Joan Ferré, Ricard Boqué

*Chemometrics, Qualimetrics and Nanosensors Group. Department of Analytical Chemistry and Organic Chemistry. Rovira i Virgili University. Tarragona, Catalonia, Spain. *e-mail address: cristina.botella@urv.cat*

Microarray gene expression data and multivariate classification have been applied for tumour classification[1,2] and to differentiate pharmacological mechanisms[3] among other applications. Because of the high relevance of a classification result, it is of primary importance to accurately evaluate the expected performance of the classification rule. The evaluation is also needed in order to decide among different classification rules, and also to optimize (fine-tune) them. The performance of a classification rule is usually evaluated with global measures such as: sensitivity[4,5,6], the number of samples misclassified[7] or the Area Under the ROC curve (AUC)[8], among others[9]. These measures are derived either from classifying a validation set (samples of known class) or by cross-validation. Hence, they have a *global* character, since they inform about the expected performance of the classification rule when it is used to classify a large number of future samples. However, they do not take into account the expected loss in taking a particular classification decision (i.e., classifying a new sample into a certain class ω_j). Such expected loss is measured by the *conditional risk* (eq 1 below for two-class classification problem):

$$R(\alpha_j|\, \hat{y}_i) = \lambda(\alpha_j|\, \omega_j) \cdot P(\omega_j|\, \hat{y}_i) + \lambda(\alpha_j|\, \omega_k) \cdot P(\omega_k|\, \hat{y}_i) \qquad \forall k \neq j \qquad [1]$$

where $P(\omega_j|\, \hat{y}_i)$, $P(\omega_k|\, \hat{y}_i)$ are the *a posteriori* probabilities calculated by the Bayes theorem[10] and $\lambda(\alpha_j|\, \omega_k)$, $\lambda(\alpha_j|\, \omega_j)$ are, respectively, the costs associated to deciding that the sample *i*th belongs to "class *j*" when it actually belongs to "class *k*" and of classifying the *i*th sample correctly into its "class *j*". A similar expression exist for the risk of assigning the sample to class *k*.

The *conditional risk* in eq 1 is often used as criterion for classification: a new sample is classified into the class for which the risk of such a decision is the lowest[10]. However, although the risk is used to guide the classification, it is not used for evaluating the performance of a classification rule using a validation set, since it does not include the already known information about the class of each sample. Hence, new measures to evaluate the quality of a classification rule from a validation set (or cross-validation) are needed. They should take into account the cost of classification (through the Bayes risk) and also the global performance (through the sensitivity and specificity).

In this communication we present a new criterion for evaluating the quality of a classification model. The criterion is shown here for probabilistic Discriminant Partial Least Squares (DPLS) models although it might be suitable to evaluate other classification rules. DPLS has recently received much attention in the microarrays field[11,12,13,14].Recently, a new probabilistic version of DPLS has been developed[15] that enables the calculation of the probability density functions for each class, and hence, the calculation of the associated Bayesian risk. The performance of these models depends on the number of factors (latent variables) that are used to describe the data. This number is decided by comparing the performance of DPLS models calculated using different number of factors. Hence, an adequate measure of performance is needed.

The performance criterion proposed in this communication, R^*, is evaluated either for a validation set or by cross-validation, and combines the sensitivity, that is a global measure of performance of the model, and the conditional risk, which is the expected performance for individual samples. The expression is derived for the classification of a sample in either two classes (class "0" or class "1"), with the possibility of rejecting to classify if the risk of classification is above a certain threshold[16].

The criterion is given by:

$$R^* = \frac{\sum_{i=1}^{I}(R_{i,true} - R_{i,assigned})}{I} + \frac{\sum_{i=1}^{I}(R_{i,true})}{I} + \frac{i_{01}}{(i_{01}+i_{11}+i_{R1})}\cdot\lambda_{01} + \frac{i_{10}}{(i_{00}+i_{10}+i_{R0})}\cdot\lambda_{10} \qquad [2]$$

where $R_{i,true}$ is the risk associated with classifying the ith validation sample in its known class, $R_{i,assigned}$ is the risk we are actually taking when we classify the ith validation sample in the class of minimum risk, I is the number of samples in the validation set, $i_{01,}$ $i_{11,}$ i_{R1} are the number of validation samples of class 1 classified into class 0, class 1 (i.e., correctly classified) and rejected, respectively and, λ_{01} the cost of being wrong when classifying a sample of class 1 into class 0. The terms $i_{10,}$ $i_{00,}$ i_{R0} and λ_{10} are interpreted similarly for samples of class 0. Note that the term $i_{01}/(i_{01}+ i_{11}+ i_{R1})$ is (1-especificity) and the term $i_{10}/(i_{00}+ i_{R0}+ i_{10})$ is (1-sensitivity) of the classification rule.

This new criterion was used to decide the optimal complexity of a probabilistic DPLS model applied to discriminate healthy and tumour samples of the Prostate cancer dataset[17]. This dataset consists of 12600 gene expressions from 102 samples. Figure 1 shows R^* for DPLS models calculated with 1 to 6 factors using 82 calibration samples and validated using cross-validation. A minimum is reached for 5 factors, which indicates that is the optimal model complexity that gives a compromise between sensitivity and minimal global risk. Table 1 compares the optimal number of factors obtained by other evaluation criteria. Note that, for a test set ($\hat{y}_i$ values predicted by the PLSD model, presented on figure 2), the models with 4 factors selected by other criteria have large classification errors than the model of 5 factors.

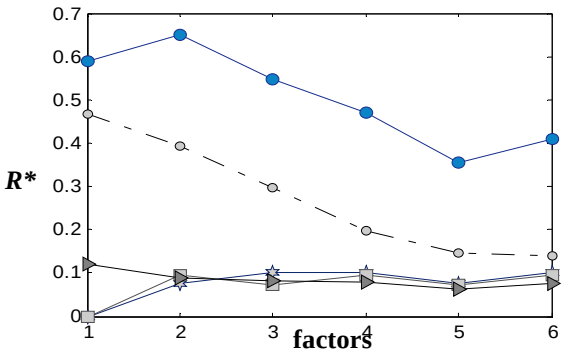

Fig 1. R^* evaluated for the different number of DPLS factors. Contributions of each term of eq 2: first term (●), second term (▶), and (1-especifity) (✦) and (1- sensitivity) (■), the third and the fourth term respectively.

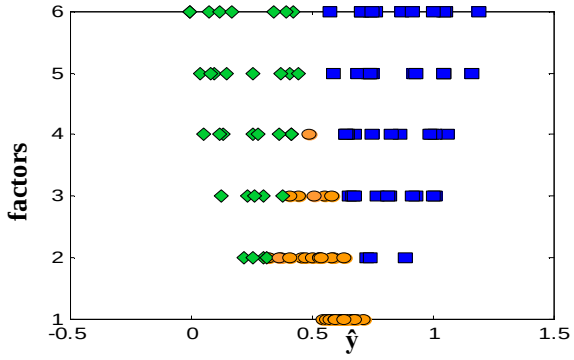

Fig 2 $\hat{y}_i$ values predicted by DPLS models for the test samples. Samples correctly classified as class 0 (◆) and class 1(■), respectively, and (●) samples rejected.

Table 1. Optimal number of factors following the different criteria

	RMSECV	AUC	Q^2	Total Error rate
Optimum number of factors	5	5	5	4
Number of samples misclassified	0	0	0	1

1. Brown M.P.S, et.al. (2000) *Proceedings of the National Academy of Sciences*, **97**, 262-267.
2. Furey, T.S, et. al. (2000) *Bioinformatics*, **16**, 906-914.
3. Gunther E.C., et.al. (2003) *Proceedings of the National Academy of Sciences*, **100**, 9608-9613.
4. Baker, S.G., et. al. (2002) *BMC Medical Research Methodology*, **2**:11.
5. Pepe, M.S., et.al., *Biometrics*, **59**, 133–142.
6. Hastie, T., Tibshirani, R. and Friedman, J. (2001) ISBN: 0–387–95284–5.
7. Nguyen, D.V. and Rocke, D.M. (2002) *Bioinformatics*, **18**, 39-50.
8. Azuaje, F., Dopazo, J. and Díaz-Uriarte, R. (2005) ISBN 978-0-470-09439-6. Chapter 12.
9. Lee, J.W., et al., (2005) *Computational Statistics and Data Analysis*, **48**, 869-885.
10. Duda, R.O., Hart, P.E. and Store, D.G. (2001) ISBN: 0-471-05669-3.
11. Boulesteix, A.-L. (2004) *Statistical Applications in Genetics and Molecular Biology*, **3**:(1), article 33.
12. Man, M.Z., et. al. (2004) *Journal of Biopharmaceutical Statistics*, **14**, 1065-1084.
13. Bylesjö, M., et. al., (2005) *BMC Bioinformatics*, **6**, 250.
14. Pérez-Enciso, M. and Tenenhaus, M. (2003) *Human Genetics*, **112**, 581-592.
15. Pérez, N.F., Boquè, R. and Ferré, J. (2006) *Conference on Chemometrics in Analytical Chemistry*, *Personal Communication.*
16. Webb, A. (2002) ISBN: 0-470-84514
17. Singh, D., et al., (2002) Gene expression correlates of clinical prostate cancer behaviour, *Cancer Cell*, **1**, 203-209.