

- The classification of biological samples via different expression patterns measured by DNA microarrays is of major interest but is hampered by many pitfalls. The use of high-dimensional data implies challenges which are in general not recognized. Classification of microarrays is prone to overfitting and a derived model might be biologically meaningless and could break down on future data.
- The key to predictive models is regularization. Gene filtering is the most widely used approach. But there are good alternatives. Consider feature selection as part of the classification method.
- Several approaches to microarray data classifier construction have been described in the scientific literature, among the most often used are Linear Discriminant Analysis, Decision Trees, Artificial Neural Networks, shrunken nearest centroids and Support Vector Machines. It is also evident from literature that a universally best method for classifier creation does not exist.
- Applying many classification strategies to a data set of interest and choosing the classifier with the best performance may introduce overfitting and a bias in the estimate of the misclassification rate.
- Classifiers for high-dimensional data are very complex algorithms which have to be handled with care to avoid overfitting to the data given. The process of building the classifier has to be separated into two steps:
 1. preprocessing of data by filtering genes or other strategies and fine tuning of parameters within the algorithm.
 2. testing of the classification performance on an independent data set.

The second step is essential as in the process of selecting genes and fine tuning parameters of the classification method the test set can be learned indirectly and the method will perform much worse on an independent test set. This problem of overfitting becomes more likely the more complex a classification method is and the more parameters are learned, if two different methods achieve the same classification performance the more simple method should always be preferred.

- To handle both levels in a correct way which does not impair the estimate of the misclassification rate, two levels of cross-validation are necessary: An inner cross-validation which helps to find the appropriate set of parameters for the algorithm, and an outer cross-validation which is used to estimate the classification performance.
- In order to obtain a reliable estimate of the performance of a class prediction method for a given data set, either repeated 10-fold cross-validation or the 0.632 bootstrap should be used as an error estimation procedure. It is important to include the entire classifier generation process (including filtering or gene selection steps) in every single iteration (i.e. on every example subset used for training) of the cross-validation or bootstrap procedure. Don't cheat yourself by selecting informative genes globally outside the cross-validation loop.

- Moreover, to assure reproducibility of the classification results achieved and to avoid overfitting or using the test set for learning or choosing the method, another completely separate portion of the available data instances (a test set) should not be used for classifier design and should be set aside for independent classifier validation. The test set should be reasonable large (number of instances $\geq 25\%$ of the complete set) and it is strongly advisable to calculate and discuss the confidence interval of a classifier's performance on the test set.
- The outer evaluation procedure to obtain a reliable estimate of the misclassification error can use either repeated 10-fold cross-validation or the 0.632 bootstrap. The inner cross-validation step may be based on a leave-one-out or a 5 fold cross-validation to make the classification process not to computational demanding.
- General experience shows that classification results are difficult to reproduce even on the same data set. It is not enough to simply publish the data set, a short description of the classification strategy, and the result. It is necessary to make the whole program code accessible in a commonly used language.
- There is a caveat on the interpretation of the results: Genes, which allow high classification accuracy, are not necessarily the ones functionally related to the feature of interest. The main effects one observes on microarrays will often be secondary or tertiary ones. What can be seen is the avalanche, not the little pebble causing it. The induction from discriminative power of genes to biological importance is misleading in the vast majority of cases.

Recommended literature discussing these points:

Ruschhaupt M., Huber W., Poustka A. and Mansmann U.: A Compendium to Ensure Computational Reproducibility in High-Dimensional Classification Tasks. *Statistical Applications in Genetics and Molecular Biology* 3(1): article 37 (2004).

Is cross-validation valid for small-sample microarray classification?

U. M. Braga-Neto and E. R. Dougherty
Bioinformatics, Vol. 20, pp. 374-80, 2004

Rules of evidence for cancer molecular-marker discovery and validation.

D. F. Ransohoff
Nat Rev Cancer, Vol. 4, pp. 309-14, 2004

Commentary: Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification

R. Simon and M. D. Radmacher and K. Dobbin and L. M. McShane
Journal of the National Cancer Institute, Vol. 95, pp. 14-18, 2003

On Comparing Classifiers: Pitfalls To Avoid And A Recommended Approach

S. L. Salzberg
Data Mining and Knowledge Discovery, Vol. 1, pp. 317--327, 1997

