Yet another Feature Selection Study for Microarrays

Matteo Pardo[(1)], Giorgio Sberveglieri[(1)], and Barbara Wold[(2)]

[(1)] INFM- National Institute for Matter Physics & University of Brescia, Chemistry and Physics Dept.,
Via Valotti 9, 25133 Brescia, Italy
pardo@ing.unibs.it
[(2)] Division of Biology, California Institute of Technology, 1200 E California Blvd MC 127-72, Pasadena CA 91125

Keywords. Feature selection, Gene selection, Microarrays, Classification

Introduction

Two histopathologically different kinds of rhabdomyosarcoma (RMS) -alveolar and embryonal RMS- are associated with distinct clinical characteristics and different cytogenetic properties. Affymetrix microarrays (U133A/B) were used to characterize the 74 tumoral tissues of both kinds. For consistency with previous work, 8801 genes have been considered in our analysis. Also, the train/test division had been fixed to 56 training and 18 test data.
Feature Selection (FS) is both useful for enhancing the classification performance and, more importantly, to discover biologically relevant genes. Therefore, FS is a hot topic in the application of machine learning to the analysis of microarray data [1,2].

Results

Two kinds of methods have generally been studied for FS, filter and wrapper methods. The essential difference between these approaches is that a wrapper method makes use of the algorithm that will be used to build the final classifier, while a filter method does not [1]. Therefore, filters are less computationally intensive, while wrappers produce better classifications.
In this paper we first filter individual genes with a Fisher-like Index (FI). The FI ranks genes by measuring the separability between the two tumor classes due to each individual gene. Few genes (~100) have high FI.
In figure 1 we show the Principal Component Analysis (PCA) plot from the top ten scoring genes. We see a good discrimination except for three samples. These are three Alveolar samples whose gene expression puts them inside the Embrional cluster (training data).
Classification with different multilayer perceptron (MLP), trained using from 100 to 4 genes (with highest FI), results in one and the same *test* mistake. Reducing to 3 genes there is a second error.
The fact is that genes ranked strictly according to FI can be, and probably are *correlated*. To find uncorrelated genes we simply turned to the loading plots. In figure 2 we show the loading plots for the first three PC. We see that the 1st PC has rather uniform loadings, except for six genes. Gene 53 (HBNF), in particular, is known to be differentially expressed in the two tumors. Genes 1,12,52 (red circled in figure 2) eventually gave test error zero with a small 3-5-2 MLP.
We then tried to find more subsets with optimal performance (zero test set error). Starting from the best (highest FI) 60 genes, we made an exhaustive search over all subsets of 1,2,3 genes (i.e. we used a wrapper approach on the filtered genes). For speed reasons, kNN (k=3) was adopted instead of MLP. We thereby obtained e.g. 677 3-genes subsets and 34 2-genes subsets with zero test set error.
We then devised a simple way to obtain a ranking of single genes from the rankings of the genes subsets. The new single gene index is a weighted mean of the performances of each gene over the top scoring subsets (those with classification performance=1). The new ranking of the sensors is very interesting, having a neat peak for genes with FI 53, 5 for the 2-genes set and 53, 5, 52 for 3-genes set (see figure 3). Genes with FI ranking 52,53 had already been singled out from the previous loading plots, while gene 5 is a new finding. In fact gene 5 looks like gene 1 from the loading plots. Obviously the new ranking is very different to the one produced by the FI.
A further question we answered is: what genes affect the tissue's appearance and make three samples seem alike the true Alveolar tissues (under the microscope), while the gene chip analysis put them in the Embrional cluster, as seen in figure 1? We again used the Fisher index ranking but the two classes are now: 1) the three irregular samples 'Embrional-like Alveolar' and 2) the 'true' embrional samples. With a loading plot analysis we found

four genes, which permit to clearly distinguish the two classes in a PCA plot. These genes probably cause histological differences, while they are not important for the tumor discrimination.

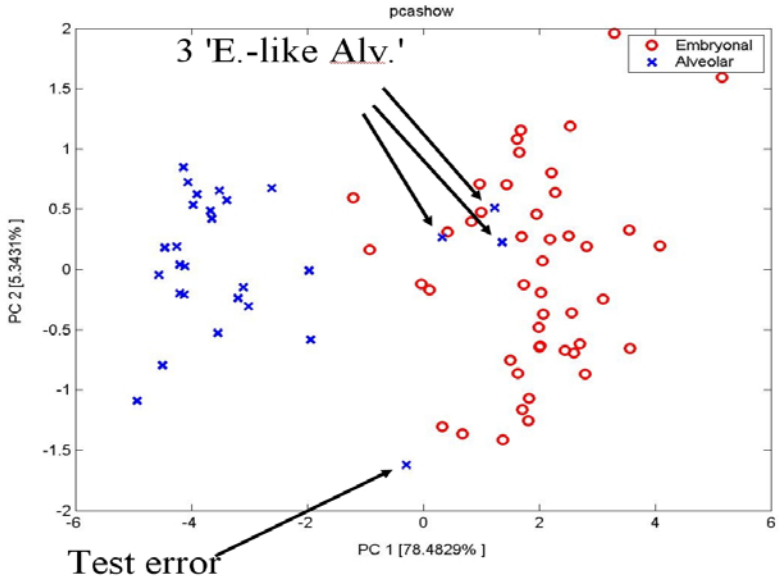

Fig. 1 PCA plot from 10 genes with highest FI score

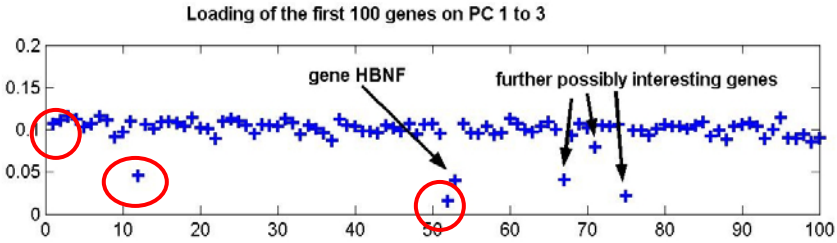

Fig. 2 Loading plot for PC1, genes are ordered according to decreasing FI. The three genes with highest FI and different loadings are circled.

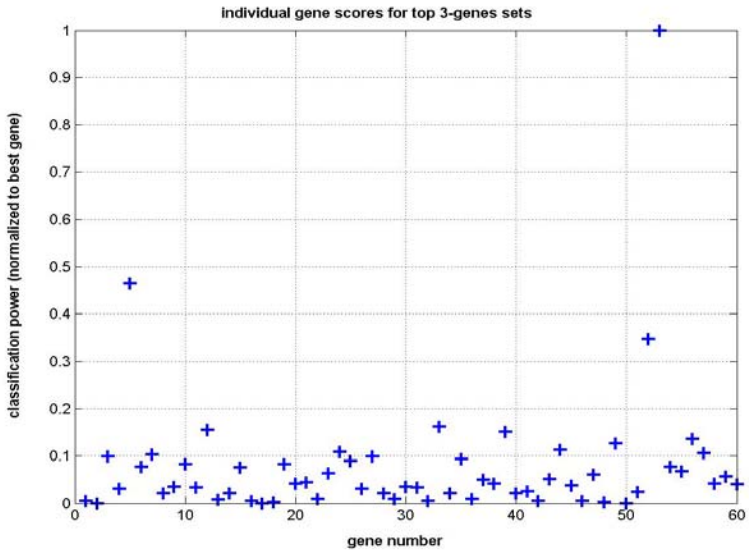

Fig. 3 Individual genes scores derived from 3-genes subsets. Three genes are clearly dominant, and gene 53 has highest ranking.

References

[1] E. P. Xing, M. I. Jordan, and R. M. Karp, Feature selection for high-dimensional genomic microarray data. Machine Learning: Proceedings of the Eighteenth International Conference, San Mateo, CA: Morgan Kaufmann, 2001.

[2] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. Machine Learning, Vol. 46, pp. 389—422, 2002