An Artificial Model for Validating Gene Selection Methods

Marco Muselli [(1)], Francesca Ruffino [(2)], and Giorgio Valentini [(2)]

[(1)] IEIIT - Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni
Consiglio Nazionale delle Ricerche- via De Marini, 6 - 16149 Genova - Italy
marco.muselli@ieiit.cnr.it

[(2)] DSI - Dipartimento di Scienze dell'Informazione Università di Milano
via Comelico, 39/41, 20135 Milano - Italy
{ruffino,valentini}@dsi.unimi.it

Keywords. Gene selection, Artificial model, Validation, Microarray.

Introduction

Every DNA microarray experiment provides thousands of real values that correspond to the gene expression levels of a tissue. This technology can offer a new valuable tool for medical diagnosis, since it can yield a reliable way to determine the state of a patient (e.g. healthy or ill) by measuring the gene expression level of its cells.

The dataset obtained through several microarray experiments can be represented by a table with m rows and n columns: each of its rows is associated with an examined tissues and each column corresponds to one of the considered genes. To specify a particular state for each tissue, a final column must be added to the table. Typically $m \sim 100$, while $n \sim 10000$.

When analyzing this table to retrieve a model for diagnosis, we have two different targets: besides finding a method that recognizes the state pertaining to a specific tissue (*discrimination*), we wish to determine the genes involved in this prediction (*gene selection*). The quality of the discrimination task can be simply estimated through a measure of accuracy, obtained by proper methods (hold-out, cross validation, etc.). On the contrary, it is very difficult to evaluate the results of the gene selection process, since the genes really involved in the onset of a state are actually unknown.

A possible way of validating gene selection could be to analyze the performance of the considered method on a diagnosis problem where significant genes are known. Unfortunately, at the present no problem of this kind is available. An alternative approach consists in building an artificial model, starting from proper biological motivations, that generates data having the same statistical characteristics of gene expression levels produced by microarray experiments.

As proposed in [1], the behavior of a biological system can be described through regulatory networks that represent the interaction between different genes. The nodes and the edges of these networks are ruled by dynamic equations that involve the concentration of products encoded by genes and consequently the gene expression levels. Each concentration is expressed through a real variable that changes with time and can determine the transition of the system from a state to another.

When the organism is in a particular state some concentrations are lower than a given *threshold* (specific for each gene), while others exceed a proper value. Thus, if we select a definite state, we can say that a gene is in the *active* state, if its expression level has a value consistent (lower or greater than a specific threshold) with that state. With this definition each gene can be described by a binary variable, assuming value 1 if the gene is active and 0 otherwise.

Also the presence of the considered state can be expressed through a Boolean variable, which takes the value 1, if the tissue is in that state, and 0 otherwise. Consequently, the whole biological system can be described by a Boolean function f with n inputs. Each of the m available microarray experiments corresponds to a particular entry of the truth table for the function f; it is formed by an input-output pair $(\boldsymbol{x},y)$, where $\boldsymbol{x}$ is a vector of n binary values associated with the examined genes and y is a binary value asserting if the corresponding tissue is in the considered state or not.

According to this setting, a technique to generate artificial data for validating gene selection methods consists in building a proper Boolean function f, whose truth table entries share the same statistical characteristics of gene expression levels produced by microarray experiments. Then, the quality of the gene selection method is measured by the percentage of significant genes retrieved.

Although each Boolean function can be described by a logical expression containing only AND, OR and NOT operations, in our case it is more convenient to obtain *f* in a different way. In fact, it can be observed that in biological systems genes can be assembled into groups of expression signatures, i.e. subsets of coordinately expressed genes related to specific biological functions [4]. These groups of genes are, in some sense, equivalent with respect to the state determination.
Thus, the Boolean function *f* can be viewed as a combination of several groups of genes. Each group is considered *active* if a sufficiently large number of its genes is active. Then, the function *f* assumes value 1 if the number of active groups exceeds a given threshold.
A proper algorithm for constructing Boolean functions with these characteristics has been implemented. It is able to generate data resembling those produced by several microarray experiments for diagnostic purpose. In these cases two or more different states are analyzed and the algorithm constructs a specific Boolean function (adopting the above approach) for each state. Then, to allow the application of the gene selection method, a set of input-output pairs is produced for each Boolean function built.
The algorithm includes several parameters that can be tuned to achieve a good agreement between the resulting collection of input-output pairs and the dataset produced by microarray experiments for a specific problem. An evaluation of this agreement can be obtained by looking at the accuracy values scored by a discriminant method for different numbers of considered genes.
In this contribution, the Leukemia dataset [2] has been considered and a proper artificial model has been generated by constructing a specific Boolean function for each of the two variants of leukemia examined. Figure 1 shows the accuracy values obtained through the leave-one-out approach by applying the SVM-RFE method described in [3] and the technique proposed in [2]. As one can note, the agreement between the success rate curves is excellent in both situations.

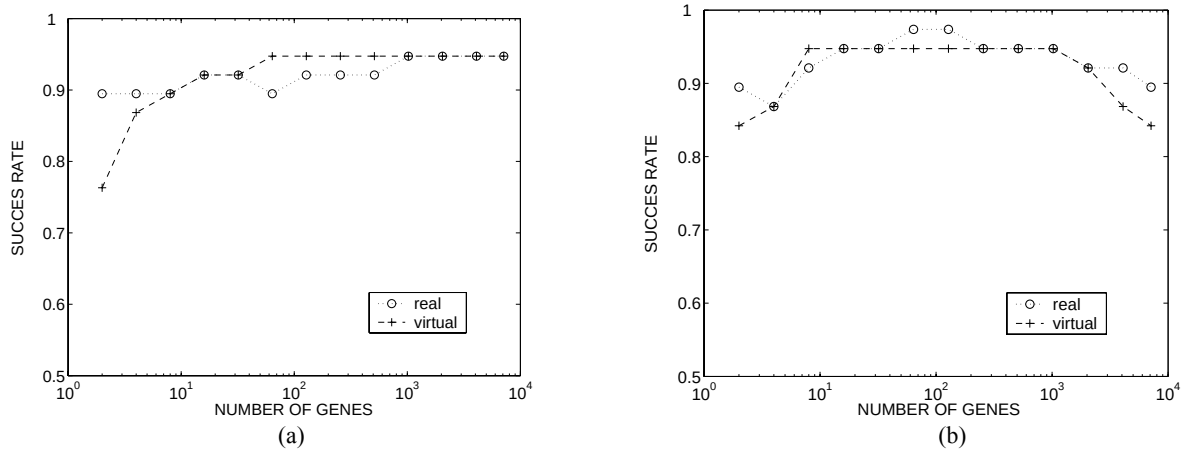

Fig. 1 Accuracy values obtained through the leave-one-out approach for the Leukemia dataset and for the dataset produced by the proposed artificial model, when varying the number of considered genes. (a) SVM-RFE method, (b) Golub's method.

References

[1] D. Repsilber and J.T. Kim, Developing and testing methods for microarray data analysis using an artificial life framework, *Advances in Artificial Life (ECAL 2003),* pages 686-695, 2003.
[2] T. Golub et al., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, vol. 286, pages 531-537, 1999.
[3] I. Guyon et al., Gene selection for cancer classification using support vectors machines, *Machine Learning*, vol. 46, pages 389-422, 2002.
[4] A. Alizadeh et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, vol. 403, pages 503-511, 2000.