Modeling gene expression data via positive Boolean functions

Francesca Ruffino[1], Marco Muselli[2], Giorgio Valentini[1]

[1] DSI, Dipartimento di Scienze dell'Informazione,
Università degli Studi di Milano,
Via Comelico 39, Milano, Italy
[2] IEIIT, Istituto di Elettronica e di Ingegneria dell'Informazione
e delle Telecomunicazioni, Consiglio Nazionale delle Ricerche, via De Marini, 6
Genova, Italy

Abstract. In this work we propose an artificial model for the generation of biologically plausible gene expression data to be used in the evaluation of the performance of gene selection and clustering methods.

The model allows to fix in advance the set of relevant genes and the functional classes involved in the problem; the input-output relationship is constructed by synthesizing a positive Boolean function. Despite its simplicity, it is sufficiently rich to take account of the specific peculiarities of gene expression data, including biological variability.

A Java code had been developed to allow the user choose the model parameters according to the characteristics of the experiment he want to simulate. This permits to insert the artificial model into a distributed system for microarray analysis, in particular one based on a Grid infrastructure.

1 Introduction

A unique possibility of understanding mechanisms regulating biological processes, such as the onset of a disease or the effects of a drug [2], is offered by DNA microarrays, which provide the expression level for thousands of genes pertaining to a given tissue.

Supervised and unsupervised machine learning and statistical methods have been largely applied to the analysis of gene expression data [6, 7, 9, 10] and, in some situations, the quality of the solution offered by a given technique can be easily evaluated. In other problems, instead, the performance of a statistic or of a learning method cannot be assessed since the correct solution is not available, even in a subset of cases.

For instance, *gene selection* methods, where the subset of genes involved in a biological process of interest is to be determined from a collection of microarray experiments, cannot be evaluated since the entire set of genes involved in a specific biological process is usually unknown or only partially known.

Other important problems, such as the discovery of new subclasses of diseases detected at bio-molecular level may be formalized as unsupervised clustering problems [1, 8]. However, besides the fact that unsupervised clustering is in general an ill-posed problem, in this case no a priori solutions are known in advance, as the “real” bio-molecular classes are usually unknown.

To provide some kind of performance evaluation, several models have been proposed to produce synthetic gene expression data for classification, clustering and gene selection problems [3, 11]. Even if in principle they may be helpful to test gene selection methods, their main limitation consists in a drastic simplification of the model, which is not sufficiently rich to take into account the peculiarities of gene expression data.

Our model describes the relationship between the expression levels of the genes of a virtual tissue and its functional state. In this way it is possible to design an artificial system for a genome-wide synthesis of gene expression data. In particular, the randomness due to biological variability and measurement errors is gathered in a specific term, whereas the deterministic part of the model has been implemented by a positive Boolean function acting on relevant genes.

Furthermore, a convenient manner of writing this kind of functions consists in employing *m-of-n expressions*, which are able to capture the main biological characteristics of gene expression, while maintaining a sufficient simplicity. The numerical experiments included in the following section show how to apply the proposed model to the analysis of the performances of largely used statistical and machine learning gene selection methods.

2 Example of gene selection method evaluation

We can associate to each DNA microarray experiment a pair $(\boldsymbol{x}, y)$, where $\boldsymbol{x}$ is a real-valued input vector whose components represent the gene expression levels for the corresponding tissue, whereas the output y can vary into a set of 2 different values (1 and -1), each denoting the class which the associated tissue belongs to; a generalization of the analysis to more than two classes is straightforward.

Our mathematical model can be adopted to describe each of the two functional states defining the classes above. Two subsequent phases have been devised: in the first one the two functions f_1 and f_2, describing the relationship between the expression level of genes and the two possible functional state of a tissue, are built, whereas in the second one the gene expression levels of n virtual tissues are generated.

Randomness inherent the determination of the functional state can be collected into a real parameter e, so that with probability $1 - e$ each virtual tissue belonging to the output class 1 (resp. -1) has gene expression levels forming a vector $\boldsymbol{x}$ verifying $f_1(\boldsymbol{x}) = 1$ (resp. $f_2(\boldsymbol{x}) = 1$). If classes are mutually exclusive (as it is usually the case), it should be guaranteed that each tissue belongs to only one functional state, i.e. if $\boldsymbol{x}$ is the associated input vector only one model provides the output 1.

The collection of virtual tissues generated by the model can be collected into a matrix X, where each row corresponds to a tissue and each column to a gene. Then, a final column Y representing the class of each tissue is added.

Feature selection and clustering methods can be applied to $Z = [X, Y]$ and X respectively. However, since both the rule determining the membership of a tissue to a class and the relationship among the virtual genes are completely known, these methods can be directly tested and their performances can be easily evaluated.

As an example, we compare two feature selection methods, the technique proposed by Golub et al. in [4] (a simple variation of the classic t-test) and the SVM-RFE procedure [5], on two different collections of examples built by adopting the proposed model. The evaluation of the performances of the two methods has been performed by counting how many relevant genes, actually belonging to the expression profile, are found.

The first dataset X_1 is composed by 100 artificial tissues, 60 belonging to the first class and 40 in the second class, with 6000 virtual genes. The expression profiles of the two functional states, represented by the functions f_1 and f_2, contain 144 genes in total. For both the functional states the parameter e has been fixed to 0.1.

Both the Golub's method and SVM-RFE have been applied to the complete dataset $Z_1 = [X_1, Y_1]$, being Y_1 the vector containing the labels y of the class of each tissue $\boldsymbol{x}$ ($y = 1$ if $f_1(\boldsymbol{x}) = 1$ or $y = -1$ if $f_2(\boldsymbol{x}) = 1$). Each gene selection method assigns a rank value to each of the 6000 genes: the higher is the rank the more relevant is the corresponding gene. The first 144 genes with greater rank values are then compared with the 144 genes actually belonging to the two expression profiles.

If we denote with G_{144} and S_{144} the set of the 144 most relevant genes selected by Golub's method and by SVM-RFE, respectively, we can evaluate the intersections between G_{144} or S_{144} and the set M_{144} of the genes included in the two expression profiles. The greater is the size of the intersection, the better is the performance of the gene selection method. A relative measure of this term is given by the fraction P_G (resp. P_S) of relevant genes contained in G_{144} (resp. R_{144}).

The results show that

$$P_G = \frac{|G_{144} \cap M_{144}|}{|M_{144}|} = \frac{132}{144} = 0.92$$

and

$$P_S = \frac{|S_{144} \cap M_{144}|}{|M_{144}|} = \frac{24}{144} = 0.17$$

having denoted with $|A|$ the cardinality (number of elements) of the set A. The comparison between the values of P_G and P_S shows that in this artificial dataset the behavior of the Golub's method is significantly better than that of SVM-RFE. In particular, the former is able to retrieve most (92%) of the relevant genes.

The application of the same approach to a second artificial dataset may help to understand if this result has a more general validity. To this aim a new data matrix $Z_2 = [X_2, Y_2]$ has been generated, where X_2 contains 80 virtual tissues

(50 belonging to the first class and 30 to the second class) and 2500 virtual genes. The value of the parameter e has been fixed to 0.05.

Since, in this case, the total number of genes belonging to the two expression profiles is 133, we consider the sets G_{133} and S_{133} obtained by applying the Golub's method and SVM-RFE, respectively, to the dataset Z_2 and by taking the 133 genes with highest rank for both methods. In this way, we can again compute the quantities P_G and P_S, given by the fraction of relevant genes included in G_{133} and S_{133}:

$$P_G = \frac{|G_{133} \cap M_{133}|}{|M_{133}|} = \frac{124}{133} = 0.93$$

while

$$P_S = \frac{|S_{133} \cap M_{133}|}{|M_{133}|} = \frac{39}{133} = 0.29$$

M_{133} is the set of the relevant genes adopted for the construction of the expressions of f_1 and f_2. As one can note, also in this case the Golub's method achieves by far the best performance.

3 Conclusions

An artificial model for the generation of biologically plausible gene expression data, to be adopted in the evaluation of gene selection and clustering methods, has been proposed.

An application of the proposed artificial model in evaluating the performances of two gene selection techniques, Golub's method [4] and SVM-RFE [5], has been also presented. The analysis of two artificial datasets, where the collection of relevant genes is considerably smaller than the whole set of genes characterizing the virtual tissue, has permitted to derive that the Golub's method performs significantly better than SVM-RFE, being able to retrieve more than 90% of the relevant genes.

The behavior of the proposed model is affected by a collection of parameters, which can be properly fixed by the user to produce a set of data having a high degree of similarity with a specific experiment of interest. A Java program assists the user in the construction of the artificial model and in the generation of the final sample of data.

Due to its flexibility this software can be directly inserted into a distributed environment for microarray analysis, possibly based on a Grid infrastructure.

Acknowledgment

This work was partially supported by the Italian MIUR projects “Laboratory of Interdisciplinary Technologies in Bioinformatics (LITBIO)” and has been developed in the context of *CIMAINA* Center of Excellence.

References

1. A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, and L.M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.
2. P. Baldi and G.W. Hatfield. *DNA Microarrays and Gene Expression.* Cambridge University Press, Cambridge, UK, 2002.
3. S. Dudoit and J. Fridlyand. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9):1090–1099, 2003.
4. T.R. Golub et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286:531–537, 1999.
5. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46:389–422, 2002.
6. T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20:2429–2437, 2004.
7. Y. Lu and J. Han. Cancer classification using gene expression data. *Information Systems*, 28:243–268, 2003.
8. M.D. Onken, L.A. Worley, J.P. Ehlers, and J.W. Harbour. Gene expression profiling in uveal melanoma reveals two molecular classes and predicts metastatic death. *Cancer Research*, 64:7205–7209, 2004.
9. J. Quackenbush. Computational analysis of microarray data. *Nat Rev Genet.*, 2(6):418–427, 2001.
10. G. Valentini, M. Muselli, and F. Ruffino. Cancer recognition with bagged ensembles of Support Vector Machines. *Neurocomputing* 56C:461–466, 2004.
11. J. Weston, A. Elisseeff, B. Scholkopf, and M. Tipping. Use of the zero-norm with linear models and kernels methods. *Journal of Machine Learning Research*, 3:1439–1461, 2003.