NETTAB 2006 – S. Margherita di Pula, 10-13 luglio 2006

Modeling Gene Expression Data via Positive Boolean Functions

F. Ruffino[1], M. Muselli[2] and G.Valentini[1]

[1] DSI, Dipartimento di Scienze dell'informazione, Università degli studi di Milano.

[2] IEIIT, Istituto di Elettronica e di Ingegneria dell'Informazione e delle Comunicazioni, C.N.R. Sezione di Genova.

SUMMARY

- ✓ Introduction
- Mathematical Model
- Applications
- Results
- Conclusions

INTRODUCTION

DNA microarrays provide the expression level for thousands of genes pertaining to a given tissue and offer the possibility of understanding mechanisms regulating biological processes, such as the onset of a disease.

Machine learning methods have been successfully used in the analysis of gene expression data. However, in some cases their performances cannot be assessed since the correct solution is not available, even in a specific case.

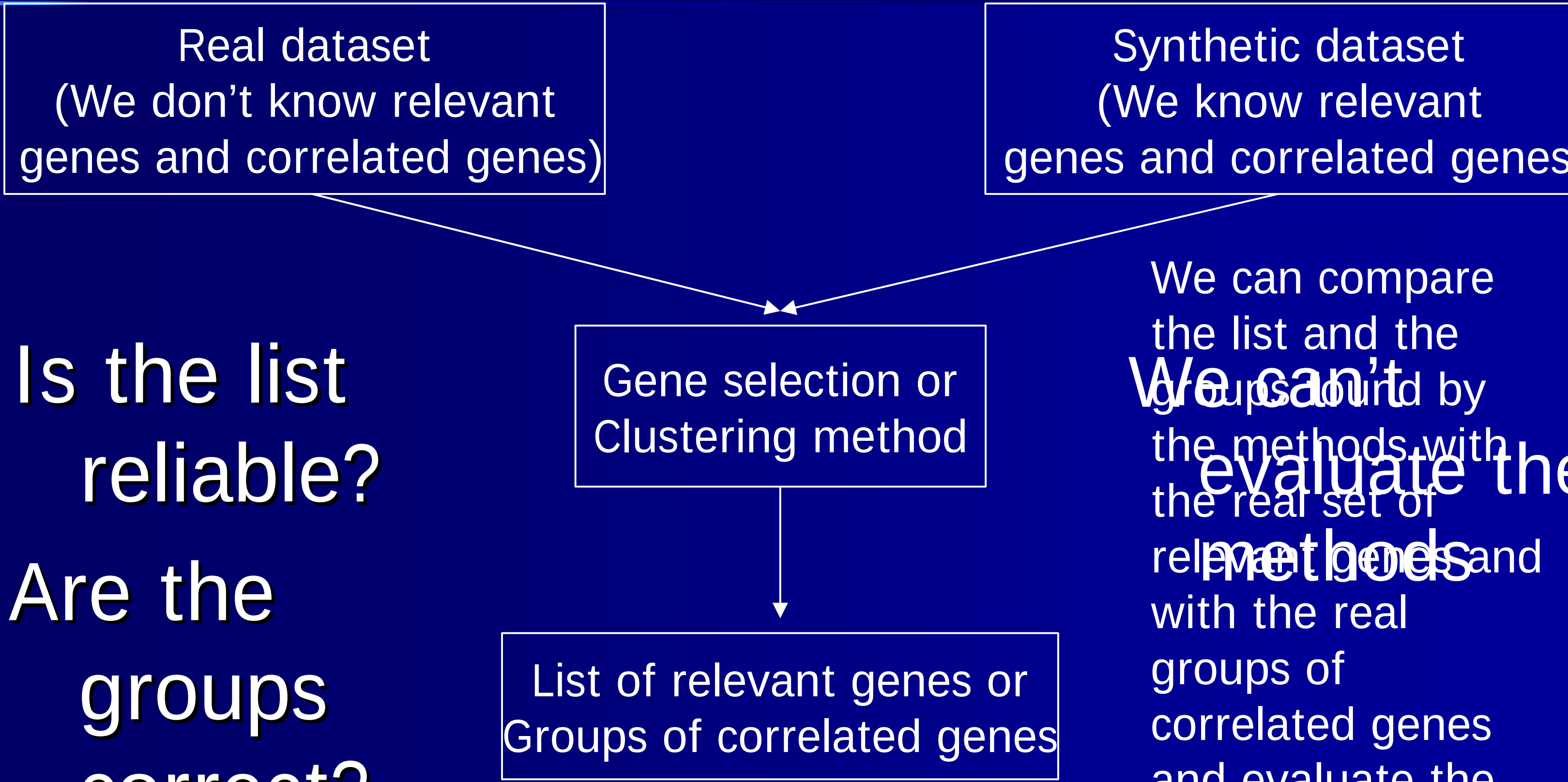
For instance, ***gene selection methods***, determining the subset of genes involved in a biological process from a collection of microarray experiments, cannot be evaluated since a reference case, where the set of relevant genes is known, is not available.

INTRODUCTION

Likewise, ***clustering methods*****,** grouping together correlated genes, cannot be evaluated since, in many cases, the correlations between the genes are unknown.

To provide some kind of performance evaluation, we propose a procedure producing ***synthetic gene expression data*** for classification, clustering and gene selection problems.

INTRODUCTION



Real dataset
(We don’t know relevant
genes and correlated genes)

Synthetic dataset
(We know relevant
genes and correlated genes

Gene selection or
Clustering method

List of relevant genes or
Groups of correlated genes

Is the list reliable?

Are the groups correct?

We can’t evaluate the methods

We can compare the list and the groups found by the methods with the real set of relevant genes and with the real groups of correlated genes and evaluate the performances of the methods.

INTRODUCTION

In contrast with previous proposals, our technique is based on a mathematical model that takes into account the peculiarities of gene expression data.

In particular, proper ***mathematical functions*** are employed to describe the relationship between the expression level of the genes of a virtual tissue and its functional state.

SUMMARY

- Introduction
- ✓ Mathematical Model
- Applications
- Results
- Conclusions

MATHEMATICAL MODEL

Due to biological variability and possible measurement errors in DNA-microarray experiments, a deterministic relation between the gene expression values of a tissue and its functional state does not exist.

Consequently, our model will be formed by:

- a ***deterministic part***, described by a function *f* , that assumes value 1 if the tissue belongs to the functional state of interest and 0 otherwise,
- a ***stochastic component***, concentrated in a random parameter *e*, corresponding to the probability that a tissue is assigned to the wrong state.

MATHEMATICAL MODEL

We define *f* as follows. Consider the vector $\boldsymbol{x} = \{x_1,\ldots,x_m\}$, formed by the expression levels of *m* genes g_i of a tissue, where $i = 1,\ldots,m$.

Suppose that, for each gene g_i, a modulation threshold t_i exists so that the gene g_i is *overexpressed* when the value x_i of its expression level exceeds t_i and *underexpressed* if $x_i < - t_i$.

Then, a gene g_i is *modulated* if it is overexpressed or underexpressed, according to the functional state of interest. This means that when the gene g_i is modulated the probability of observing the considered functional state increases.

MATHEMATICAL MODEL

Suppose that the output is uniquely determined by the state (modulated or not) of the *m* genes and does not depend on their specific expression values.

Then, the function *f* can be written as a composition of

- a ***mapping*** $M : R^m \rightarrow \{0,1\}^m$, assigning value 1 to modulated genes and 0 otherwise, and
- a ***Boolean function*** $f : \{0,1\}^m \rightarrow \{0,1\}$, assigning value 1 to the tissues belonging to the state represented by *f*.

$$\begin{tikzcd}
\boxed{\boldsymbol{x} = \{x_1,\ldots,x_m\}} \arrow[r, "M"] \arrow[rr, bend right, "f"'] & \boxed{\boldsymbol{z} = \{1,0,\ldots,0,1\}} \arrow[r, "f"] & \boxed{\{0,1\}}
\end{tikzcd}$$

MATHEMATICAL MODEL

Consequently, once the mapping *M* is completely described, the deterministic component *f* of our model is uniquely determined by the construction of the Boolean function *f* .

Our model considers only ***positive Boolean functions***, which can be described by a compact mathematical representation, called ***m-of-n expression***.

Although the *m-of-n expression* can be formally defined, it is more convenient to introduce it by using a simple example.

Suppose the number of considered genes is *m* = 10. We can represent a specific positive Boolean function *f* in the following way:

MATHEMATICAL MODEL

$$f\ (\boldsymbol{z}) = \left\{ \left[z_7,\ z_4 \right]_{1} \left[z_1 \right]_{1} \left[z_7,\ z_2,\ z_{10} \right]_{2} \right\}_{1}$$

Each set enclosed into square brackets represents a group of ***correlated genes***.

The number associated with each group represents the minimum number of genes that have to be modulated to make the group ***active***.

The number associated with the braces represents the minimum number of groups that have to be active to make $f\ (\boldsymbol{z}) = 1$.

It follows that the genes included into the representation of *f* , i.e. g_1, g_2, g_4, g_7 and g_{10}, are the ***relevant genes*** for the functional state defined by *f* .

MATHEMATICAL MODEL

It can be shown that every positive Boolean function can be represented through an *m*-of-*n* expression if the groups of inputs (genes) and the indexes (thresholds) are properly chosen.

As a consequence, every virtual functional state can be represented by a positive Boolean function *f* possessing the described structure.

SUMMARY

- Introduction
- Mathematical Model
- ✓ Applications
- Results
- Conclusions

APPLICATIONS

We can associate with every 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,
- the output y can assume the value 1 or - 1, denoting the two possible classes of the tissue.

Our mathematical model can be adopted to build two functions f_1 and f_2, each describing one of the two functional states considered for the output.

Then, the gene expression levels of n virtual tissues can be generated starting from the mathematical expressions of f_1 and f_2.

APPLICATIONS

Randomness inherent the determination of the functional state is 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 the classes are mutually exclusive (as it is usually the case), it should be guaranteed that each tissue belongs to only one functional state.

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.

APPLICATIONS

$$\begin{array}{c} \begin{array}{cc} \overbrace{\hspace{16em}}^{\boldsymbol{X}} & \qquad Y \end{array} \\ \left[\begin{array}{ccccc|c} 0.19 & -0.76 & -0.23 & \cdots & 0.78 & 1 \\ -0.84 & 0.52 & -0.18 & \cdots & 0.21 & 1 \\ 0.37 & 0.25 & 0.76 & \cdots & -0.62 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ -0.27 & -0.94 & 0.31 & \cdots & 0.46 & 0 \end{array}\right] \end{array}$$

Feature selection and clustering methods can be applied to $Z = [X,Y]$ and X respectively.

However, since the <u>relationships among the virtual genes are completely known, these methods can be directly tested and their performances can be easily evaluated.</u>

SUMMARY

- Introduction
- Mathematical Model
- Applications
- ✓ Results
- Conclusions

RESULTS

As an example, we compare two feature selection methods:

- the technique proposed by Golub (a simple variation of the classic *t*-test) and
- the SVM-RFE procedure

on two different collections of examples built by adopting the proposed model.

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 *m* virtual genes.

The list of relevant genes of the two functional states, represented by the functions f_1 and f_2, contains 144 genes in total. For both the functional states the parameter *e* has been fixed to 0.1.

RESULTS

Every 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.

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. Then, let M_{144} be the collection of the true 144 relevant genes.

The greater is the size of the intersection between G_{144} and S_{144} and M_{144}, 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 $G144$ (resp. S_{144}).

RESULTS

The results show that

$$P_G = \frac{|G_{144} \cap M_{144}|}{|M_{144}|} = \frac{132}{144} = 0.92 \qquad Ps = \frac{|S_{144} \cap M_{144}|}{|M_{144}|} = \frac{24}{144} = 0.17$$

having denoted with |*A*| the 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.

RESULTS

The application of the same approach to a second artificial dataset $Z_2=[X_2,Y_2]$ may help to understand if this result is more general.

Now, 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.

In this case we have 133 relevant genes and we obtain

$$P_G = \frac{|G_{133} \cap M_{133}|}{|M_{133}|} = \frac{124}{133} = 0.93 \qquad Ps = \frac{|S_{133} \cap M_{133}|}{|M_{133}|} = \frac{39}{133} = 0.29$$

As one can note, also in this case the Golub's method achieves by far the best performance.

SUMMARY

- Introduction
- Mathematical Model
- Applications
- Results
- ✓ Conclusions

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.

As an application we have considered two artificial datasets, where the collection of relevant genes is considerably smaller than the whole set of genes characterizing the virtual tissue.

The analysis 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.

CONCLUSIONS

A Java code has been developed to allow the user to 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 Grid infrastructure.