BITS 2005 Meeting

Evaluation of gene selection methods through artificial and real-world data concerning DNA microarray experiments

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

[1]Dipartimento di Scienze dell'Informazione, Università di Milano, Milano, Italy
[2]Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, Consiglio Nazionale delle Ricerche, Genova, Italy

Abstract

MOTIVATION

DNA microarrays provide the gene expression level for thousands of genes pertaining a given tissue, thus allowing to understand mechanisms regulating biological processes, such as the onset of a disease or the effects of a drug. An important problem in this analysis is to determine the subset of genes involved in the biological process under examination. Such problem is generally referred to as *gene selection* and several statistic and machine learning techniques have been proposed in literature to face with it.
Golub et al. [1] have obtained interesting results in discriminating two different kinds of leukemia by adopting a simple univariate statistic method (GOLUB). More recently, Guyon et al. [2] have successfully employed a recursive procedure, called Recursive Feature Elimination (RFE), based on the application of linear Support Vector Machines (SVM).
Another promising class of machine learning techniques for gene selection is rule generation methods, which solve a classification problem by generating a collection of intelligible rules in the if-then form. In particular, Switching Neural Networks (SNN) have been shown to obtain an excellent accuracy, when applied to solve real world problems deriving from DNA microarray [3]. The application of a variant of RFE, called Recursive Feature Addition (RFA), allows to use SNN to perform gene selection.

METHODS

If we want to evaluate in an objective way the quality of a gene selection method, such as GOLUB, SVM-RFE, or SNN-RFA, we cannot adopt real data, since we do not know the collection of genes actually involved in the underlying biological process. A valid alternative consists in using artificial datasets that presents a similar statistic behavior as data deriving from DNA microarray experiments. The procedure *TAGGED* (*Technique for Artificial Generation of Gene Expression Data*) allows to achieve this result: it is based on the concept of expression signature, introduced in literature to denote a set of correlated genes with respect to a functional state.

A gene belonging to an expression signature is considered to be *active* if the expression level of that gene is higher (or less) than a given threshold level. A binary variable can therefore be associated to each gene, assuming value 1 if the gene is active and 0 otherwise. By using this coding, the condition of the genes of a cell can be identified by a Boolean string z showing active genes with respect to the functional state of interest.

Now, if we consider a binary variable y assuming value 1 when the cell is in the considered functional state and 0 otherwise, a Boolean function $y = f(z)$, linking the cell state and its genes, can be derived. To generate artificial data for a binary classification problem, TAGGED builds in a

plausible way two Boolean functions $f_1(z)$ and $f_2(z)$ for two virtual functional states. Starting from these two functions the desired dataset is produced.

RESULTS

To evaluate the results obtained by GOLUB, SVM-RFE and SNN-RFA when performing gene selection on real world problems, three datasets containing gene expression levels produced by DNA microarrays have been considered: the Leukemia dataset [1], the Colon cancer dataset [4] and the Lymphoma dataset [5]. In addition, the three techniques have also been applied for the analysis of three artificial datasets generated by TAGGED, which present statistical properties similar to those pertaining to the three real world problems above.

The results obtained for the three datasets show that a good agreement exists between performances on real and artificial data (this points out the validity of TAGGED). Furthermore, GOLUB and SNN-RFA achieve excellent results on artificial data, where the set of relevant genes is known.

SUPPLEMENTARY INFORMATION

[1] T. R. GOLUB ET AL.: "Monotone molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, 286, 531–537, 1999.

[2] I. GUYON ET AL.: "Gene selection for cancer classification using support vector machines", Machine learning, **46,** 389–422, 2002.

[3] M. MUSELLI: "Gene selection through Switched Neural Networks", NETTAB-2003: Workshop on Bioinformatics for Microarrays (Bologna, Italy, 27–28 November 2003).

[4] U. ALON ET AL.: "Broad patterns of gene expressions revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", Proceedings of the National Academy of Science USA, **96,** 6745–6750, 1999.

[5] A. A. ALIZADEH ET AL.: "Different types of diffuse large B-cell lymphoma identified by gene expression profiling", Nature, **403,** 503–511, 2000.