Málaga, junio de 2010

Informe Ejecutivo

TÍTULO: BIO-2.0-2010: Clasificación y Selección de Genes en Microarrays de ADN

RESUMEN: En este informe se define el problema de la selección y clasificación de genes en Microarrays de ADN. El objetivo es descubrir pequeños subconjuntos descriptivos de genes, extraídos de grandes bases de datos, que ofrezcan el máximo poder clasificador posible. Así, el especialista podrá trabajar con conjuntos de genes más manejables pero con el mismo poder de clasificación, descartando la información genética errónea y redundante. Este problema se resolverá en el ámbito de DIRICOM empleando una combinación de técnicas de optimización (metaheurísticas) y clasificadores de máquinas de soporte (SVM).

OBJETIVOS:

1. Definir el problema de la selección de genes y la función objetivo.
2. Presentar la estrategia seguida para resolver el problema.
3. Seleccionar de Microarrays con los que trabajar.

CONCLUSIONES:

1. Los resultados obtenidos serán puestos a disposición de los especialistas con el objetivo de que se trabaje con genes más manejables pero con el mismo poder de clasificación, descartando la información genética errónea y redundante.

   Este problema se resolverá en el ámbito de DIRICOM empleando una combinación de técnicas de optimización (metaheurísticas) y clasificadores de máquinas de soporte (SVM).

RELACIÓN CON ENTREGABLES:

PRE: BIO-1.0-2009 (anterior o necesario de leer)

CO: PSO-1.0-2008 (anterior o necesario de leer)

Málaga, June, 2010

Executive Summary

TITLE: BIO-2.0 Gene Selection and Classification in DNA Microarrays

ABSTRACT: In this report, the problem of gene selection and classification in DNA Microarrays is defined. The main goal is to discover small sets of descriptive genes from large gene expression datasets. The obtained sets offering a classification rate as precise as possible. Thus, the specialist can handle small gene subsets by rejecting wrong and redundant information from the initial Microarray. This problem is solved in the scope of DIRICOM by using a series of optimization techniques (metaheuristics) and support vector machines classifiers (SVM).

GOALS:

1. Defining the problem of gene selection and the fitness function.
2. Designing the optimization strategy followed for solving the problem.
3. Selecting case study well-known and real world Microarrays.

CONCLUSIONS:

1. The obtained results will be available to the specialists with the aim of providing easy to use and small gene subsets by rejecting wrong and redundant information from the initial Microarray. This problem is solved in the scope of DIRICOM by using a series of optimization techniques (metaheuristics) and support vector machines classifiers (SVM).

RELATION WITH DELIVERABLES:

PRE: BIO-1.0-2009 (mandatory reading)

CO: PSO-1.0-2008 (advisable reading)

Gene Selection and Classification in DNA Microarrays

DIRICOM

June 2010

1. Introduction

DNA Microarrays ([9]) allow scientists to simultaneously analyze thousands of genes, thus giving important insights about cell's function, since changes in the physiology of an organism are generally associated with changes in large gene ensembles of expression patterns. The vast amount of data involved in a typical Microarray experiment usually requires to perform a complex statistical analysis, with the goal of the classification of the dataset into correct classes. The key issue in this classification is to identify significant and representative gene subsets that may be later used to predict class membership for new external samples. Furthermore, these subsets should be as small as possible in order to develop fast and low consuming processes for the future class prediction. The main difficulty in Microarray classification versus other domains is the availability of a relatively small number of samples in comparison with the number of genes in each sample. In addition, expression data are highly redundant and noisy, and most genes are believed to be uninformative with respect to studied classes, as only a fraction of genes may present distinct profiles for different classes of samples.

In this context, machine learning techniques have been applied to handle with large and heterogeneous datasets, since they are capable to isolate the useful information by rejecting redundancies. Concretely, feature selection (gene selection in Biology) is often considered as a necessary preprocess step to analyze large datasets, as this method can reduce the dimensionality of the datasets and often conducts to better analyses [6].

2. Problem definition

Feature selection for gene expression analysis in cancer prediction often uses wrapper classification methods [7] to discriminate a type of tumor, to reduce the number of genes to investigate in case of a new patient, and also to assist in drug discovery and early diagnosis. Several classification algorithms could be used for wrapper methods, such as K-Nearest Neighbor (K-NN) [3] or Support Vector Machines (SVM) [2]. By defining clusters, a big reduction of the number of considered genes and an improvement of the classification accuracy can be finally achieved. The formal definition of the feature selection problem that we consider here is:

Let $F = \{f_1, ..., f_i, ..., f_n\}$ *be a set of features; find a subset* $F' \subseteq F$ *that maximizes a scoring function* $\Theta : \Gamma \rightarrow G$ *such that* $F' = argmax_{G \subset \Gamma}\{\Theta(G)\}$*; where* Γ *is the space of all possible feature subsets of* F *and* G *a subset of* Γ.■

Optimal feature selection is a complex problem proved to be NP-hard [8]. Therefore, we need efficient automated approaches to tackle it. Metaheuristics algorithms (studied in the DIRICOM project) have proved to be adequate tools for this matter, since they are capable of solving the feature selection accurately and efficiently for the large dimensions needed in Biology. Evolutionary Algorithms (EAs) and, specifically, Genetic Algorithms (GAs) have been successfully used in the past to tackle the gene selection of Microarrays ([1, 5, 4]). All these approaches consist in using single population sequential algorithms which can achieve competitive performances (from the point of view of the quality of solution), but without considering other important aspects such as the computational effort and the time consumption.

2.1. Gene Selection and Classification Scheme

Following the basic scheme of solution encoding used in feature selection, a given metaheuristic algorithm should provide a binary encoded solution (vector) where each bit represents a gene in the dataset. If a bit is 1, it means that this gene is kept in the reduced subset, while 0 indicates that the gene is not included. Therefore, the individual length is equal to the number of genes in the initial Microarray dataset.

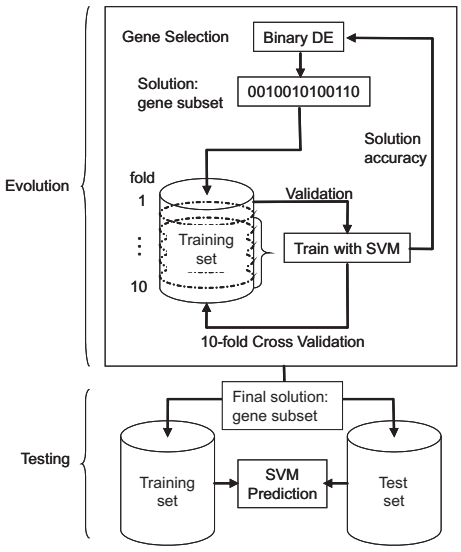

Figura 1: General scheme of our DE-SVM for gene selection and classification of DNA Microarrays

As illustrated in Fig. 1, where as an example, the general model of a Differential Evolution for feature selection [4] is provided, the individuals representing gene subsets, are evaluated by means of a SVM classifier and validated using *10-fold cross-validation* as follows: each gene subset (codified by an individual) is divided into ten subsets of samples, nine of them constituting the training set and the remaining one used as the validation set. The SVM is trained using the training set and then, the accuracy obtained (number of correct sample classifications by the SVM once trained) is evaluated on the validation set [2]. This evaluation in repeated ten times, each one alternating the used validation set. This method reinforces the validation process, so that the final accuracy value is the resulted average of the ten validation folds. Such a strong validation is necessary when the number of samples is low regarding the number of features, which is the goal in this project. Once the evolution process has finished, the resulted subset solution is evaluated on the external test set obtaining the accuracy.

2.2. Fitness Function

A fitness function is needed to guide the search by assigning to any tentative solution a quality value. Once the accuracy value and the number of genes are known, the fitness function is calculated according to:

$$f(x) = (100 - acc) + \lambda \cdot \frac{\#(genes\ in\ subset)}{\#(total\ genes)}, \tag{1}$$

$$being,\ \lambda = 10^{\lfloor log(\#(total\ genes))+1 \rfloor} \tag{2}$$

The objective here consists of maximizing the accuracy and minimizing the number of genes. For convenience (only minimization of fitness), the first factor is presented as $(100 - acc)$ and the second one is normalized in order

to control the trade off between these two factors. A constant value λ (which depends on the total number of genes) is used in this normalization. Therefore, if the number of features in the subset is high (with regards to the total number of genes in the original dataset), then the fitness function promotes the reduction of features. Otherwise, if the number of features in the subset is small, then this fitness function promotes the accuracy improvement.

2.3. Microarray Datasets and Data Preprocessing

Here we present the Microarray instances used in DIRICOM, which can be classified into four well-known datasets got from real-word Microarray experiments. All of them were taken from the public repository of Kent Ridge Biomedical Dataset (`http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html`). In particular:

- The ALL-AML Leukemia dataset consists of 72 Microarrays experiments with 7,129 gene expression levels. Two classes exist: *Acute Myeloid Leukemia* (AML) and A*cute Lymphoblastic Leukemia* (ALL). The complete dataset contains 25 AML and 47 ALL samples. The original dataset is divided into a training set of 38 samples and a test set of 34 samples.
- The Colon Tumor dataset consists of 62 Microarray experiments collected from colon-cancer patients with 2,000 gene expression levels. Among them, 40 tumor biopsies are from *tumors* and 22 (*normal*) are from healthy parts of the colons of the same patients.
- *Types of Diffuse Large B-cell Lymphoma* dataset consists of 47 tissue samples, 24 of them are from *germinal centre B-like group* while the rest 23 are *activated B-like group*. Each sample is described by 4,026 genes.
- The Lung Cancer dataset involves 181 experiments with 12,533 gene expression levels. Classification occurs between *Malignant Pleural Meso-thelioma* (MPM) and *Adenocarcinoma* (ADCA) of the lung. In tissue samples there are 31 MPM and 150 ADCA.

Table 1 summarizes the original organization of training and testing samples of the four used Microarrays. The test sets of Leukemia and Lung were taken from the original repositories provided by the authors. In Colon and Lymphoma, only training sets are available in the original repositories, and for this reason, new test and training sets have been here generated for these two datasets by randomly (uniformly) extracting samples from the original one as stated in Table 1. Expression levels of training and test sets were normalized separately in order to scale their intensities, thus enabling a fair comparison between the different datasets. Each attribute was scaled to $[-1, 1]$ by using (as LIBSVM recommends):

These datasets were selected because of their different dimensions and gene organizations, constituting a heterogeneous tests-bed to better support (robust) our conclusions.

$$a'_j(x_i) = 2 \cdot \frac{a_j(x_i) - min_j}{max_j - min_i} - 1, \tag{3}$$

where max_j and min_j correspond to the maximum and minimum gene expression values for attribute a_j over all samples.

Reductions of genes by removing them according to thresholds were not made previously, and so we make the task of correct classification harder by leaving even clearly non-functional genes for the algorithm to remove. Uninformative genes to the classifier could nevertheless be informative ones to the metaheuristic algorithm, since bad solutions are quite important for avoiding guiding the search towards low quality regions of the search space.

Cuadro 1: Usage details of the four Microarray Datasets

Dataset	#genes	Clases	#Train	#Test	#Total
Colon	2,000	Cancer	20	20	40
		Normal	11	11	22
Lymp.	4,026	Ac B-like	17	6	23
		Ce B-like	19	5	24
Leuk.	7,129	AML	11	14	25
		ALL	27	20	47
Lung	12,533	MPM	16	15	31
		ADCA	16	134	150

3. Conclusions

This report presents the problem of gene selection and classification in DNA Microarrays is defined. The main goal is to discover small sets of descriptive genes from large gene expression datasets. The obtained sets offering a classification rate as precise as possible. The main goal is to provide to the scientific community small easy to use gene subsets by rejecting wrong and redundant information from the initial Microarray.

The report is divided in to three main parts: the first one defines the problem of gene selection and the fitness function, the second one designes the optimization strategy followed for solving the problem, and finally, a series of case study well-known and real world Microarrays are selected to work with.

This problem is solved in the scope of DIRICOM by using a series of optimization techniques (metaheuristics) and support vector machines classifiers (SVM).

Referencias

[1] E. Alba, J. García-Nieto, L. Jourdan, and E.-G. Talbi. Gene Selection in Cancer Classification using PSO/SVM and GA/SVM Hybrid Algorithms. In *IEEE Congress on Evolutionary Computation CEC-07*, pages 284–290, Singapore, Sep 2007.

[2] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mac. Lear.*, 20(3):273–297, 1995.

[3] Fix and J. L. Hodges. Nonparametric discrimination: Consistency properties. Technical report, 4, US Air Force School of Aviation Medicine, R. Field, TX, 1951.

[4] J. Garcia-Nieto, E. Alba, and J. Apolloni. Hybrid de-svm approach for feature selection: Application to gene expression datasets. In *Logistics and Industrial Informatics, 2009. LINDI 2009. 2nd International*, pages 1 –6, sept. 2009.

[5] J. García-Nieto, E. Alba, L. Jourdan, and E. Talbi. Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis. *Information Processing Letters*, 109(16):887 – 896, 2009.

[6] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[7] J. Kohavi and G. H. John. The wrapper approach. In *Feature Selection for Knowledge Discovery and Data Mining*, pages 33–50, 1998.

[8] M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Trans. on Computer*, 26:917–922, 1977.

[9] A.C. Pease, D. Solas, E. Sullivan, M. Cronin, C. P. Holmes, and S. Fodor. Light-generated oligonucleotide arrays for rapid dna sequence analysis. In *Proc. Natl. Acad. Sci.*, volume 96, pages 5022–5026, USA, 1994.