Gene Selection in Cancer Classification using PSO/SVM and GA/SVM Hybrid Algorithms

Enrique Alba, José García-Nieto, Laetitia Jourdan and El-Ghazali Talbi

***Abstract*— In this work we compare the use of a Particle Swarm Optimization (PSO) and a Genetic Algorithm (GA) (both augmented with Support Vector Machines SVM) for the classification of high dimensional Microarray Data. Both algorithms are used for finding small samples of informative genes amongst thousands of them. A SVM classifier with 10-fold cross-validation is applied in order to validate and evaluate the provided solutions. A first contribution is to prove that PSO_{SVM} is able to find interesting genes and to provide classification competitive performance. Specifically, a new version of PSO, called Geometric PSO, is empirically evaluated for the first time in this work using a binary representation in Hamming space. In this sense, a comparison of this approach with a new GA_{SVM} and also with other existing methods of literature is provided. A second important contribution consists in the actual discovery of new and challenging results on six public datasets identifying significant in the development of a variety of cancers (leukemia, breast, colon, ovarian, prostate, and lung).**

I. Introduction

Microarray technology (DNA microarray) [1] allows to simultaneously analyze thousands of genes and thus can give important insights about cell's function, since changes in the physiology of an organism are generally associated with changes in gene expression patterns. Several gene expression profiles obtained from tumors such as Leukemia [2], Colon [3] and Breast [4] have been studied and compared to expression profiles of normal tissues. However, 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. So, tools to deal with these issues are critically important. These tools should learn to robustly identify a subset of informative genes embedded out of a large dataset which is contaminated with high-dimensional noise [5].

In this context, feature selection is often considered as a necessary preprocess step to analyze these data, as this method can reduce the dimensionality of the datasets and often conducts to better analyses [6].

Two models of feature selection exist depending on whether the selection is coupled with a learning scheme or not. The first one, *filter model*, which carries out the feature subset selection and the classification in two separate phases, uses a measure that is simple and fast to compute. Hence, a filter method, is in definition, independent of the learning algorithm used after it. The second one, the *wrapper method*, which carries out the feature subset selection and classification in the same process, engages a learning algorithm to measure the classification accuracy. From a conceptual point of view, wrapper approaches are clearly advantageous, since the features are selected by optimizing the discriminate power of the finally used induction algorithm.

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) [8] or Support Vector Machines (SVM) [9]. By creating clusters a big reduction of the number of considered genes and an improvement of the classification accuracy can be finally achieved.

The definition of the feature selection problem is this: given a set of features $F = \{f_1, ..., f_i, ..., f_n\}$, find a subset $F' \subseteq F$ that maximizes a scoring function $\Theta : \Gamma \rightarrow G$ such that

$$F' = argmax_{G \subset \Gamma}\{\Theta(G)\}, \quad (1)$$

where Γ is the space of all possible feature subsets of F and G a subset of Γ. The optimal feature selection problem has been shown to be NP-hard [10]. Therefore, only heuristics approaches are able to deal with large size problems. Recently, such advanced structured methods have been used to explore the huge space of feature subsets, like for example metaheuristics as Evolutionary Algorithms and, specifically, Genetic Algorithms (GAs) [11], [5], [12].

In this work, we are interested in gene selection and classification of DNA Microarray data in order to distinguish tumor samples from normal ones. For this purpose, we propose two hybrid models that use metaheuristics and classification techniques. The first one consists of a Particle Swarm Optimization (PSO) [13] combined with a SVM approach. PSO is a population based metaheuristic inspired by the social behavior of bird flocking or fish schooling. Specifically, a recent version called Geometric PSO [14] (explained in Section II) has been used in this work. The second model is based on the popular GA using a specialized Size-Oriented Common Feature Crossover Operator (SSOCF) [15], which keeps useful informative blocks and produces offsprings which have the same distribution than the parents. This model will be also combined with SVM in our approach.

E. Alba and J. García-Nieto are with the Department of Lenguajes y Ciencias de la Computación, Universidad de Málaga, (email: {eat,jnieto}@lcc.uma.es, url http://neo.lcc.uma.es)

L. Jourdan and E-G. Talbi are with the LIFL/INRIA Futurs-Université de Lille 1, Bât M3-Cité Scientifique, (email: {jourdan,talbi}@lifl.fr, url http://www2.lifl.fr/OPAC/)

Both proposed approaches are experimentally assessed on six well-known cancer datasets (Leukemia, Colon, Breast, Ovarian [16], Prostate [17] and Lung [18]), discovering new and challenging results and identifying specific genes that our work suggests as significants. Performances of proposed PSO and GA algorithms solving the gene extraction problem (using SVM) are compared in this paper. Specifically, we focused in the capacity of the PSO_{SVM} combination in order to provide considerable performance in this matter. In this sense, comparisons with several state of art methods show competitive results according to the conventional criteria.

The outline of this work as follows. We review the PSO and the SVM techniques in order to introduce our PSO_{SVM} hybrid model in Section II. In Section III, the six microarray datasets used in this study are described. Experimental results are presented in Section IV, including biological descriptions of several obtained genes. Finally, we summarize our work and present some conclusions and possible future work in Section V.

II. Gene Selection and Classification by PSO_{SVM}

In this section, we describe the hybrid PSO_{SVM} approach for gene selection and classification of Microarray data. The PSO algorithm is designed for obtaining gene subsets as solutions in order to reduce the high number of genes to be later classified. The SVM classifier is used whenever the fitness evaluation of a tentative gene subset is required.

A. Particle Swarm Optimization

Particle Swarm Optimization was first proposed by Kennedy and Eberhart in 1995 [13]. PSO is a population based evolutionary algorithm inspired in the social behavior of bird flocking or fish schooling. In the description of PSO, the swarm is made up of a certain number of particles (similar to population of individuals in EAs). At each iteration, all the particles move in the problem space to find the global optima. Each particle has a current position vector and a velocity vector for directing its movement.

$$v_i^{k+1} = \omega \cdot v_i^k + \varphi_1 \cdot rnd_1 \cdot (pBest_i - x_i^k) + \varphi_2 \cdot rnd_2 \cdot (g_i - x_i^k) \quad (2)$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \quad (3)$$

Equations 2 and 3 describe the velocity and position update of a given particle i at a certain iteration k. Equation 2 calculates a new velocity v_i for each particle (potential solution) based on its previous velocity, the particle's location at which the best fitness so far has been found $pBest_i$, and the population global (or local neighborhood, in the neighborhood version of the algorithm) location at which the best fitness so far has been achieved g_i. Individual and social weight are represented by means of φ_1 and φ_2 factors respectively. Finally, rnd_1 and rnd_2 are random numbers in range $\{0, 1\}$, and ω represents the inertia weight factor. Equation 3 updates each particle's position x_i in solution space.

B. The SVM Classifier

Support Vector Machines, a technique derived from statistical learning theory, is used to classify points by assigning them to one of two disjoint half spaces [9]. So, SVM performs mainly a (binary) 2-class classification. For linearly separable data, SVM obtains the hyperplane which maximizes the margin (distance) between the training samples and the class boundary. For non linearly separable cases, samples are mapped to a high dimensional space where such a separating hyperplane can be found. The assignment is carried out by means of a mechanism called the *kernel* function.

SVM is widely used in the domain of cancer studies, protein identification and specially in Microarray data [6], [12], [19]. Unfortunately, in many bioinformatics problems the number of features is significantly larger than the number of samples. For this reason, tools for decreasing the number of features in order to improve the classification or to help to identify interesting features (genes) in noisy environments are necessary. In addition, SVM can treat data with a large number of genes, but it has been shown that its performance is increased by reducing the number of genes [20]. The hybrid PSO and hybrid GA approaches next proposed contribute notably in this sense.

C. The Hybrid PSO_{SVM} Approach

In order to offer a basic idea of the operation of our PSO_{SVM} approach, in Figure 1, we can observe a simple scheme of how features are extracted from the initial microarray dataset and how the resulted subset is evaluated.

In a first phase, the metaheuristic algorithm involved, PSO in this case, provides a binary encoded particle[1] where each bit[2] represents a gene. If a bit is 1, it means that this gene is kept in the subset and 0 indicates that the gene is not included in the subset. Therefore, the particle length is equal to the number of genes in the initial microarray dataset.

The original PSO was initially developed for continuous optimization problems. However, lots of practical engineering problems are formulated as combinatorial optimization problems and specifically as binary decisions. Several binary versions of PSO can be found in present literature [21], [22]. Nevertheless, these versions consist on *ad hoc* adaptations from the original PSO and therefore their performances are usually improvable.

With the aim of facing the gene selection problem, an innovative version of PSO, based on the geometric framework presented in [14], has been developed in this work. This version, called Geometric Particle Swarm Optimization (GPSO), enables to us to generalize PSO to virtually any solution representation in a natural and straightforward way. This property was demonstrated for the cases of Euclidean, Manhattan and Hamming spaces in the referenced work. Even a recently appeared work [23], uses the GPSO for solving the Sudoku Puzzle by means of permutations

[1] chromosome in GA and solution (S) in Figure 1

[2] allele in GA

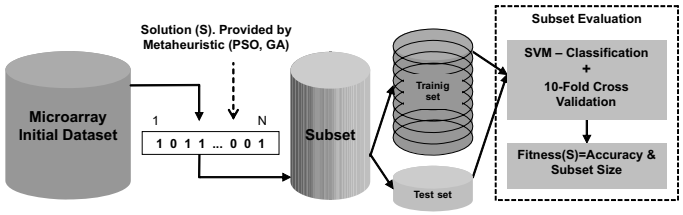

Fig. 1. A simple scheme of how features (genes) are selected out from the original microarray dataset using a particle with binary encoding. In a second phase, the resulted subset is evaluated by means of a SVM classifier and 10-fold cross validation to obtain the fitness value (accuracy) of such particle.

representation. Since the gene selection problem has been represented by binary way, specific operators for Hamming space were used in the PSO described here.

D. Geometric Particle Swarm Optimization

In this version, the location of each particle i is represented as vector $x_i = \langle x_{i1}, x_{i2}, ..., x_{iN} \rangle$ taking each bit x_{ij} (with j in $\{1, N\}$) binary values 0 or 1. The key issue of the GPSO is the concept of particle movement. In this approach, instead of the notion of velocity added to the position, a *three-parent mask-based crossover* (3PMBCX) operator is applied to each particle in order to "move" it. According to the definition of 3PMBCX [14], given three parents a, b and c in $\{0, 1\}^n$, generate randomly a crossover mask of length n with symbols from the alphabet $\{a, b, c\}$. Build the offspring filling each element with the bit from the parent appearing in the crossover mask at the position.

The pseudocode of the GPSO algorithm for Hamming spaces is illustrated in Algorithm 1. For a given particle i, three parents take part in the 3PMBCX operator (line 13): the current position x_i, the social best position g_i and the historical best position found h_i (of this particle). The weight values $w1$, $w2$ and $w3$ indicate for each element in the crossover mask the probability of having values from the parents x_i, g_i or h_i respectively. These weight values associated to each parent represent the *inertia* value of the current position ($w1$), the *social* influence of the global/local best position ($w2$) and the *individual* influence of the historical best position found ($w3$). A constriction of the geometric crossover forces $w1$, $w2$ and $w3$ to be non-negative and add up to one.

In summary, the GPSO developed in this study operates as follows: In a first phase of the pseudocode, the initialization of particles are carried out by means of the $SwarmInitialization()$ function (Line 1). This special initialization method (used also in our GA approach) was adapted to gene selection as follows. The swarm (population) was divided into four subsets of particles (chromosomes) initialized in different ways depending on the number of features in each particle. That is, 10% of particles were initialized with N (prefixed value) selected genes (1s) located randomly. Another 20% of particles were initialized with $2N$ genes, 30% with $3N$ genes and finally, the rest of particles (40%) were initialized randomly and 50% of the genes were turned on. In these experiments N will be equal to 4. In

Algorithm 1 Pseudocode of the GPSO for Hamming space.

1: $S \leftarrow SwarmInitialization()$
2: **while** not stop condition **do**
3: **for** each particle x_i of the swarm S **do**
4: evaluate(x_i)
5: **if** $fitness(x_i)$ is better than $fitness(h_i)$ **then**
6: $h_i \leftarrow x_i$
7: **end if**
8: **if** $fitness(h_i)$ is better than $fitness(g_i)$ **then**
9: $g_i \leftarrow h_i$
10: **end if**
11: **end for**
12: **for** each particle x_i of the swarm S **do**
13: $x_i \leftarrow 3PMBCX((x_i, w_1), (g_i, w_2), (h_i, w_3))$
14: mutate(x_i)
15: **end for**
16: **end while**
17: **Output:** best solution found

a second phase, after the evaluation of particles (line 4), historical and social position are updated (lines 5 to 10). Finally, particles are "moved" by means of the 3PMBCX operator (line 13). In addition, with a probability of 10%, a simple bit-mutation operator (line 14) is applied in order to avoid the early convergence. This process is repeated until reach the stop condition fixed to a certain number of evolutions.

E. Evaluation Function

Since the position of a particle x_i represents a gene subset, the evaluation of each particle is carried out by means of the SVM classifier to assess the quality of the represented gene subset. The fitness of a particle x_i is calculated applying a *10-fold Cross Validation* (10FCV) method to calculate the rate of correct classification (accuracy) of a SVM trained with this gene subset. In 10FCV, the data set is divided into 10 subsets. Each time, one of the 10 subsets is used as the test set and the other 9 subsets are put together to form the training set. Then the average error across all 10 trials is computed. The complete fitness function is described in Equation 4.

$$fitness(x) = \alpha \cdot (100/accuracy) + \beta \cdot \#features, \quad (4)$$

where α and β are weight values set to 0.75 and 0.25 respectively in order to control that the accuracy value takes priority over the subset size, since high accuracies are preferred when guiding the search process. The objective here consists of maximizing the accuracy and minimizing

the number of genes ($\#features$). For convenience (only minimization of fitness) the first factor is presented as $(100/accuracy)$.

III. Data Sets

Instances used in this study consists of six well-known datasets issued of microarray experiments, ALL-AML Leuke-mia dataset, Breast cancer dataset, Colon tumor dataset, Ovarian cancer dataset, Prostate cancer dataset, and Lung cancer dataset. All of them were taken from the public Kent Ridge Bio-medical Data Repository with url `http://sdmc.lit.org.sg/GEDatasets/Datasets.html`.

- The ALL-AML Leukemia dataset consists of 72 microarray experiments with 7129 gene expression levels. Two classes for distinguishing: *Acute Myeloid Leukemia* (AML) and A*cute Lymphoblastic Leukemia* (ALL). The complete dataset contains 25 AML and 47 ALL samples.
- The Breast cancer dataset consists of 97 experiments with 24481 gene expression levels. Patients studied show two classes of diagnosis called *relapse* with 46 patients and *non-relapse* with 51 ones.
- The Colon tumor dataset consists of 62 microarray experiments collected from colon-cancer patients with 2000 gene expression levels. Among them, 40 tumor biopsies are from *tumors* and 22 (*normal*) biopsies are from healthy parts of the colons of the same patients.
- The Lung cancer dataset involves 181 microarray experiments with 12533 gene expression levels. Classification occurs between *Malignant Pleural Mesothelioma* (MPM) and *Adenocarcinoma* (ADCA) of the lung. In tissue samples there are 31 MPM and 150 ADCA.
- The Ovarian cancer dataset consists of 253 microarray experiments with 15154 gene expression levels. The goal of this experiment is to identify proteomic patterns in serum that distinguish cancer from non-cancer scenarios. The dataset includes 162 (of 253) ovarian *cancers* and 91 *normal* ones.
- The Prostate cancer dataset involves 136 microarray experiments with 12600 gene expression levels. Two classes must be differentiated: *tumor* with 77 $(52 + 25)$ samples and *non-tumor* with 59 (50+9) samples.

IV. Experimental Results and Comparisons

For our PSO_{SVM} approach, the PSO was implemented in C++ following the *skeleton* architecture of the MALLBA [24] library. For the GA_{SVM} approach the GA was implemented in C++ using the ParadisEO [25] Framework. The GA implements a generational evolution strategy (offspring replacement with elitism) and uses the following operators: deterministic tournament selection, SSOCF crossover, and uniform mutation. The SVM classifier used in both approaches is based on the LIBSVM [26] library. For the SVM confi-guration, the same parameters were used in PSO and GA algorithms and the Kernel function was configured as Linear. The fitness function used in GA_{SVM} is the same one (described in Section II) as in PSO_{SVM}.

All experiments were carried out using a PC with Linux O.S (Suse 9.0 with kernel 2.4.19) and a Pentium IV 2.8GHz processor, with 512MB of RAM. PSO_{SVM} and GA_{SVM} algorithms on six cancer related microarray datasets were independent executed 10 times over each dataset, in order to have statistically meaningful conclusions as both algorithms are stochastics.

A. Parameters Settings

The parameters used in our PSO and GA algorithms are shown in Table I. These parameter were selected after several test evaluations of each algorithm and dataset instance until reach the best configuration in terms of the quality of solutions and the computational effort.

TABLE I
PSO and GA parameters for gene subset selection and classification

PSO		**GA**	
Parameter	**Value**	**Parameter**	**Value**
Swarm size	40	Population size	40
Number of generations	100	Number of generations	100
Neighborhood size	20	Probability of crossover	0.9
Probability of mutation	0.1	Probability of mutation	0.1
(w1, w2, w3)	(0.33, 0.33, 0.34)	-	-

B. Discussion and Analysis

Several observations can made based on the above experiments, so we tackle the analysis of results focusing on the performance and robustness of our algorithms, as well as the quality of the obtained solutions providing a biological description of most significant ones.

1) Performance Analysis: From the point of view of the performance, both algorithms obtain in a few iterations acceptable results in gene selection, providing reduced subsets with high classification rates. However, the behavior is slightly different. Figure 2 shows a graphical evolution, in terms of the average of the fitness value, of a typical execution of PSO_{SVM} and GA_{SVM}. It is noticeable that in few iterations (4 or 5) the average of fitness decrease quickly and then stop in similar solutions. The large diversity of solutions provided in the initialization method (Section II-B) provokes fast of good solutions and the early convergence of both methods. Although the GA_{SVM} generally obtains lower average than PSO_{SVM}, whose solutions have in turn higher diversity.

Results for all the datasets are shown in Table II. Columns 2 and 3 contain the average of the best solutions obtained in 10 independent executions of PSO_{SVM} and GA_{SVM} respectively. Six state of the art methods from literature are presented in columns 4 to 10 in order to show how our proposals actually push forward the research in this area. Cells in - haven't values to our knowledge. Standard criteria are used to compare the results: the *classification accuracy* in terms of the rate of correct classification (first value in every table cell) and the *number of selected genes* (the value in parenthesis).

TABLE II

COMPARISON OF RELEVANT WORKS ON CANCER CLASSIFICATION WITH PROPOSED MODELS GA_{SVM} AND PSO_{SVM}. IN BOLD WE MARK THE MOST ACCURATE RESULTS. CELLS WITHOUT KNOWN VALUE (TO US) ARE MARKED WITH - CHARACTER

Dataset	PSO_{SVM}	GA_{SVM}	[5]	[6]	[12]	[27]	[28]	[29]	[30]
Luekemia	97.38(3)	97.27(4)	-	100(2)	100(25)	100(4)	87.55(4)	-	-
Breast	86.35(4)	95.86(4)	-	-	-	-	79.38(67)	-	-
Colon	100(2)	100(3)	94.12(37)	98.0(4)	99.41(10)	97.0(7)	93.55(4)	85.48(-)	94.00(4)
Lung	99.00(4)	99.49(4)	-	-	-	-	98.34(6)	-	-
Ovarian	99.44(4)	98.83(4)	-	-	-	-	-	99.21(75)	-
Prostate	98.66(4)	98.65(4)	88.88(20)	-	-	-	-	-	-

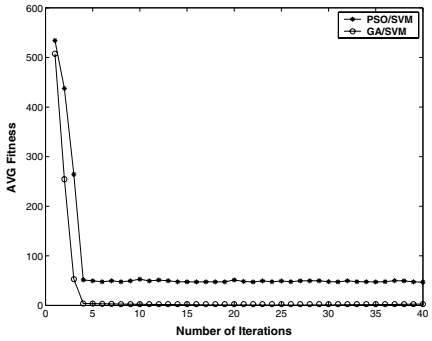

Fig. 2. Evolution of the average fitness (AVG_Fitness) in a typical execution of PSO_{SVM} and GA_{SVM} approaches using the Leukemia dataset.

In this comparison, we can observe that all solutions provided by our PSO_{SVM} and GA_{SVM} algorithms present a classification rate higher than 86%, and subsets with four and less than four selected genes are common. We outperform all the existing results (to our knowledge) but one case [6] presents a smaller subset of 2 genes. We suspect that the initialization method used in our work helps the performance of the algorithms significantly, finding small subsets with a high classification accuracy. PSO with SVM algorithm was also used in [30] (PSO/SVM from now on), although showing several differences with respect to our PSO_{SVM}. In the first place, in PSO/SVM the binary values of each particle are chosen at each iteration by means of a decision function based on threshold parameters, whereas in PSO_{SVM}, binary particles evolve following a completely different mechanism, that is, 3PMBCX crossover and mutation operators (Subsection II-D). In second place, during the evaluation phase, a *leave-half-out cross-validation* method was used in PSO/SVM whereas the PSO_{SVM} validates the selected subsets by means of *10 k fold cross-validation*.

If we compare our GPSO and GA metaheuristics combined with SVM similar results are found. In general, the GA approach obtain better best solutions (Hits) although the best classification was provided by $GPSO_{SVM}$ for the Colon tumor dataset (100% accuracy and 2 genes in Table IV). From the point of view of the accuracy average in all independent runs, the GPSO obtain a better performance although the difference with regard to the GA (as shows Fig. 3) is insignificant.

2) Algorithm Robustness: One of the most important criteria in evaluating any proposed algorithm is the quality of the algorithm and its ability to generate similar (identical) outcomes when executed several times. This factor is very important for metaheuristics which is the case in this work. To examine the robustness of the two proposed approaches, in some instances, in all ten runs both algorithms manages to find the same answer or similar ones (not identical). However, it is worthwhile mentioning that the total accuracy and number of selected features in all the cases did not deviate from each other by more than 5.5. Table III shows the result of running the GA_{SVM} and PSO_{SVM} algorithms in terms of statistical results, reporting the *Best* solution found, *Mean* and *Standard Deviation* of ten independent runs.

TABLE III

COMPARISON IN TERMS OF STATISTICAL RESULTS OF THE GA_{SVM} AND PSO_{SVM} APPROACHES. THE *Best* SOLUTION FOUND, *Mean* AND *Standard Deviation* OF 10 INDEPENDENT RUNS WERE REPORTED

Dataset	PSO_{SVM}			GA_{SVM}		
	Best	Mean	Std Dev.	Best	Mean	Std Dev.
Leukemia	100(3)	97.38(3)	3.80	100(4)	97.27(4)	3.82
Breast	90.72(4)	86.35(4)	4.11	100(4)	95.86(4)	5.33
Colon	100(2)	100(2)	0.0000	100(3)	100(3)	0.0000
Lung	99.44(4)	99.00(4)	0.50	100(4)	99.49(4)	0.41
Ovarian	100(4)	99.44(4)	0.38	100(4)	98.83(4)	3.18
Prostate	100(4)	98.66(4)	1.14	100(4)	98.65(4)	3.24

3) Brief Biological Analysis of Selected Genes: Finally, a summary of the best subsets of genes found for each dataset is shown in Table IV. All subset of genes which reported close to 100% test accuracy and the minimum number of genes. It is remarkable that apparently (to our knowledge) several discovered genes that has not been seen in any past studies. In this sense, we can provide a brief biological description of some of the most frequently obtained genes since they are currently used in the design of drugs and cancers treatment. Some of which are listed below:

- Gene *L12052_at* is "CAMP phosphodiesterase mRNA, 3' end" which is used in drugs like Anagrelide or Milrinone. Specifically the Anagrelide is used for the treatment of essential thrombocytosis, and it was proved to be effective in treating patients with certain kinds of leukemia such as chronic myeloid leukemia [31]. This gene belongs to a set of 3 genes (reported from leukemia dataset in Table IV) with 100% accuracy selected by the PSO_{SVM}.
- Gene *AB022847* is a "Solute carrier family 6 (neu-

TABLE IV

SUBSETS OF GENES REPORTED WITH 100% TEST ACCURACY

Dataset	PSO_{SVM}		GA_{SVM}	
Leukemia	100(3)	*U39226_at, L12052_at, X99101_at*	100(4)	*Z26634_at, HG870-HT870_at X52005_at, L02840_at*
Breast	90.72(4)	*NM_012269, NM_002850 AL162032, AB022847*	100(4)	*NM_005014, AF060168 NM_021176, NM_013242*
Colon	100(2)	*U29092, M55543*	100(3)	*M90684, M94132 X62025*
Lung	99.44(4)	*31820_at, 33389_at 39057_at, 40772_at*	100(4)	*31573_at, 33226_at 36245_at, 37076_at*
Ovarian	100(4)	*MZ49.784115, MZ3546.2884 MZ4362.0866, MZ9159.3641*	100(4)	*MZ420.40671, MZ825.16557 MZ1024.6857, MZ1166.0749*
Prostate	100(4)	*35106_at, 35869_at 36754_at, 37107_at*	100(4)	*41447_at, 34299_at 39556_at, 39813_s_at*

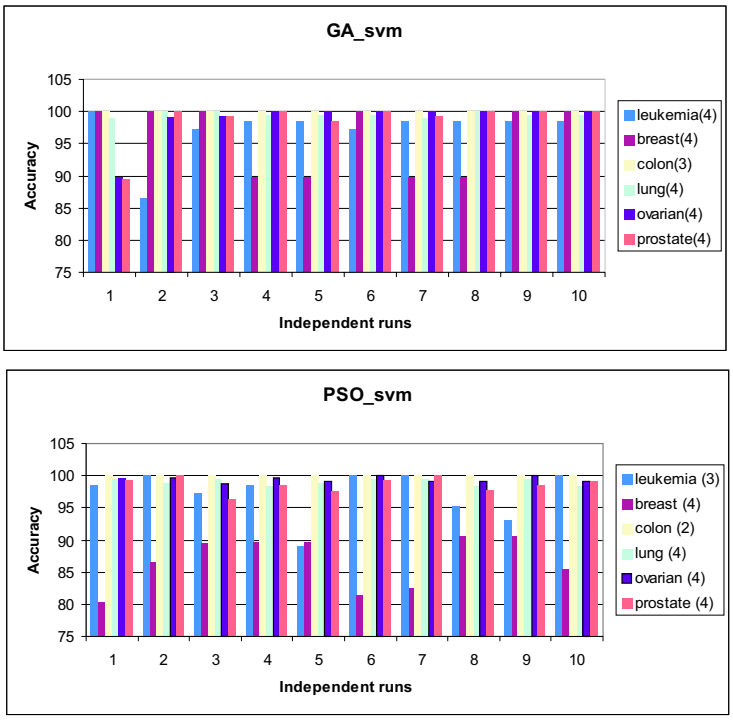

Fig. 3. Accuracy obtained by PSO_{SVM} and GA_{SVM} in each independent run. Legends specify the datasets with the number of features in parenthesis.

rotransmitter transporter, noradrenalin), member 2" located in plasma membrane. Current drugs like Radaxafine, Amphetamine or Venlafaxine are associated with this gene. Specifically, Venlafaxine is a prescription antidepressant first introduced by Wyeth in 1993. It is specifically used in management of hot flashes in survivors of breast cancer [32]. This gene belongs to a set of 3 genes (reported from breast dataset in Table IV) with 95.8763% accuracy selected by the PSO_{SVM}.

- Gene *36245_at* is "5-hydroxytryptamine (serotonin) receptor 2B" located in human plasma membrane. There are several drugs where this gene is used like Risperidone, Blonanserin and Mirtazapine. Some studies consider Mirtazapine as the first-choice agent for anxiety and depression after lung transplantation. This gene belongs to a set of 4 genes (reported from lung dataset in Table IV) with 100% accuracy selected by the GA_{SVM}.

V. CONCLUSIONS AND FUTURE WORK

In this work, two hybrid techniques for gene selection and classification of high dimensional DNA Microarray data were presented and compared. These techniques are based on different metaheuristic algorithms such as GPSO and GA used for feature selection using the SVM classifier to identify potentially good gene subsets. Specifically, the Geometric PSO algorithm for Hamming space was used to solve a real problem (gene selection in this case) for the first time (to our knowledge). In addition, genes selected are validated by an accurate 10-fold cross validation method to improve the actual classification.

Both approaches (PSO_{SVM} and GA_{SVM}) were experimentally assessed on six well-known cancer datasets discovering new and challenging results, and identifying specific genes that our work suggests as significant ones. In this sense, comparisons with several state of art methods show competitive results according to standard evaluation. Results of 100% classification rate and few genes per subset (3 and 4) are obtained in most of our executions. The use of an adapted initialization method has shown a great influence on the performance of proposed algorithms, since it introduces an early set of acceptable solutions in their evolution process.

Continuing the line of this work, we are interested in developing and testing several combinations of other metaheuristics with classification methods in order to discover new and better subsets of genes using specific Microarray datasets. In this sense, the utilization of multiobjective approaches could contribute notably in gene subset selection.

ACKNOWLEDGMENT

Authors acknowledge founds from the INRIA-PERFOM 3+3 Mediterranean project. E. Alba and J. García-Nieto acknowledge founds from the Spanish Ministry of Education and European FEDER under contract TIN2005-08818-C04-01 (the OPLINK project, `http://oplink.lcc.uma.es`).

REFERENCES

[1] A. 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.*, vol. 96, USA, 1994, pp. 5022–5026.

[2] R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.

[3] Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci*, vol. 96, pp. 6745–6750, 1999.

[4] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–536, 2002.

[5] Juliusdottir, D. Corne, E. Keedwell, and A. Narayanan, "Two-phase EA/K-NN for feature selection and classification in cancer microarray datasets," in *CIBCB*, 2005, pp. 1–8.

[6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002. [Online]. Available: citeseer.ist.psu.edu/guyon02gene.html

[7] J. Kohavi and G. H. John, "The wrapper approach," in *Feature Selection for Knowledge Discovery and Data Mining*, 1998, pp. 33–50. [Online]. Available: citeseer.ist.psu.edu/article/kohavi97wrapper.html

[8] Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," 4, US Air Force School of Aviation Medicine, Randolph Field, TX, Tech. Rep., 1951.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

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

[11] J. Yang and V. Honavar, *Feature Extraction, Construction and Selection: A Data Mining Perspective*, 1998, ch. Feature Subset Selection Using a Genetic Algorithm, pp. 177–136.

[12] E. B. Huerta, B. Duval, and J.-K. Hao, "A Hybrid GASVM Approach for Gene Selection and Classification of Microarray Data," in *Lecture Notes in Computer Science of EvoWorkshops*, F. Rothlauf, J. Branke, S. Cagnoni, E. Costa, C. Cotta, R. Drechsler, E. Lutton, P. Machado, J. H. Moore, J. Romero, G. D. Smith, G. Squillero, and H. Takagi, Eds., vol. 3907. Springer, 2006, pp. 34–44.

[13] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," in *Proc. of the IEEE International Conference on Neural Networks*, vol. 4, 1995, pp. 1942–1948.

[14] A. Moraglio, C. D. Chio, and R. Poli, "Geometric Particle Swarm Optimization," in *10th European conference on Genetic Programming (EuroGP 2007)*, ser. Lecture Notes in Computer Science, vol. 4445. Springer, Abril 2007.

[15] L. Jourdan, C. Dhaenens, and E.-G. Talbi, "A genetic algorithm for feature selection in data-mining for genetics," in *Proceedings of the 4th Metaheuristics International ConferencePorto (MIC'2001)*, Porto, Portugal, 2001, pp. 29–34.

[16] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, vol. 359, pp. 572–577, 2002.

[17] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, and J. Richie, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, pp. 203–209, 2002.

[18] G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Res*, vol. 62, pp. 4963–4967, 2002.

[19] S. Mukherjee, *Classifying Microarray data using Support Vector Machines*, Boston, MA, 2003, ch. 9, pp. 166–185.

[20] T. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machines classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[21] J. Kennedy and R. Eberhart, "A Discrete Binary Version of the Particle Swarm Algorithm," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 5, 1997, pp. 4104–4109.

[22] M. Clerc, "Binary Particle Swarm Optimisers: Toolbox, Derivations, and Mathematical Insights," 2005. [Online]. Available: http://clerc.maurice.free.fr/pso/

[23] A. Moraglio and J. Togelius, "Geometric particle swarm optimization for the sudoku puzzle," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2007)*. London, UK: ACM Press, July 2007.

[24] E. Alba and M. Group, "Mallba: A Library of Skeletons for Combinatorial Optimisation," in *Proceedings of the Euro-Par*, B. Monien and R. Feldmann, Eds., vol. LNCS 2400, 2002, pp. 927–932.

[25] S. Cahon, E.-G. Talbi, and N. Melab, "Paradiseo: A framework for parallel and distributed metaheuristics," in *Proc. of Internationa the Parallel and Distributed Processing Symposium*, 2003, pp. 144–155.

[26] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," Software available at URL http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2002.

[27] K. Deb and A. R. Reddy, "Classification of two-class cancer data reliably using evolutionary algorithms," KanGAL Report No. 2003001, Tech. Rep., 2003.

[28] L. Yu and H. Liu, "Redundancy based feature selection for microarry data," in *Proceedings of International Conference on Knowledge Discovery and Data Mining*, Seattle, Washington, 2004, pp. 22–25.

[29] B. Liu, Q. Cui, T. Jiang, and S. Ma, "A combinational feature selection and ensemble neural network method for classification of gene expression data," *BMC Bioinformatics*, vol. 5, pp. 136–148, 2004.

[30] Q. Shen, W. M. Shi, W. Kong, and B. X. Ye, "A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and classification," *Talanta*, 2006.

[31] R. T. Silver, "Anagrelide is effective in treating patients with hydroxyurea-resistant thrombocytosis in patients with chronic myeloid leukemia," *Leukemia*, vol. 19, no. 3, p. 3943, 2005. [Online]. Available: http://dx.doi.org/10.1038/sj.leu.2403556

[32] C. L. Loprinzi, "Venlafaxine in management of hot flashes in survivors of breast cancer: a randomised controlled trial," *The Lancet*, vol. 356, no. 9247, pp. 2059–2063, 2000.