Gene Selection for Microarray Data by a LDA-based Genetic Algorithm

Edmundo Bonilla Huerta, Béatrice Duval, and Jin-Kao Hao

LERIA, Université d'Angers
2 Boulevard Lavoisier, 49045 Angers, France
{edbonn,bd,hao}@info.univ-angers.fr

Abstract. Gene selection aims at identifying a (small) subset of informative genes from the initial data in order to obtain high predictive accuracy. This paper introduces a new wrapper approach to this difficult task where a Genetic Algorithm (GA) is combined with Fisher's Linear Discriminant Analysis (LDA). This LDA-based GA algorithm has the major characteristic that the GA uses not only a LDA classifier in its fitness function, but also LDA's discriminant coefficients in its dedicated crossover and mutation operators. The proposed algorithm is assessed on a set of seven well-known datasets from the literature and compared with 16 state-of-art algorithms. The results show that our LDA-based GA obtains globally high classification accuracies (81%-100%) with a very small number of genes (2-19).

Keywords : Linear discriminant analysis, genetic algorithm, gene selection, classification, wrapper.

1 Introduction

The DNA Microarray technology permits to monitor and to measure gene expression levels for tens of thousands of genes simultaneously in a cell mixture. Several studies have demonstrated that expression profiles provide valuable information for cancer diagnosis and prognosis [1,2,3,9]. The ability to distinguish a cancer from morphologically similar tissues using their gene expression profiles is important to propose appropriate therapies. Classification of different tumor types is intertwined with the problem of gene selection, which aims to extract from a great number of genes monitored by a Microarray chip, a small subset of discriminant genes. Gene selection is thus of practical and fundamental interest. The identification of relevant biomarkers is necessary for the elaboration of medical diagnostic tests. Knowledge about discriminant gene subsets may confirm the understanding of cancer mechanisms and suggest new ideas to explore.

Two main approaches have been proposed for gene selection. Filter methods rely on a criterion that depends only on the data to assess the importance or relevance of each gene for class discrimination. A relevance scoring provides a ranking of the genes from which the top-ranking ones are generally selected as the most relevant genes. Filter methods ignore the correlations among genes and the

M. Chetty, A. Ngom, and S. Ahmad (Eds.): PRIB 2008, LNBI 5265, pp. 252–263, 2008.

interaction of the selected genes with the classifier. Wrapper approaches embed gene subset selection and evaluation with the same process and consequently overcome the above mentioned inconvenient.

In this paper, we propose a new wrapper approach for gene subset selection and classification of Microarray data. Our approach uses Fisher's Linear Discriminant Analysis (LDA) to provide useful information to a Genetic Algorithm (GA) for an efficient exploration of gene subsets space. LDA is a well-known method of dimension reduction and classification, where the data vectors are transformed into a low-dimensional subspace such that the class centroids are spread out as much as possible. It has been used for several classification problems and recently for Microarray data [8,27,28].

Our approach first extracts a set of interesting genes (about 100 genes) by a filter method in order to limit the search space. Then we use a dedicated GA to determine a small subset of genes that allows a high classification accuracy. Contrary to most previously GAs for gene selection that rely essentially on random genetic operators, we devise a problem specific GA that takes into account useful knowledge of the gene selection and classification problem. Our GA uses a LDA classifier to assess the fitness of a given candidate gene subset and LDA's discriminant coefficients in its crossover and mutation operators.

To evaluate the usefulness of the proposed approach, we carry out extensive experiments on seven public datasets and compare our results with 16 best performing algorithms from the literature. We observe that our approach is able to achieve a high prediction accuracy (from 81% to 100%) with a very small number of informative genes (from 2 to 19). Moreover, our approach enables to propose different subsets of discriminant genes, which may be of a great interest for biological research.

The remainder of this paper is organized as follows. Section 2 recalls the main characteristics of Fisher's LDA and discusses the calculus that must be done in the case of small sample size. Section 3 presents our LDA-based GA for gene selection. Section 4 shows the experimental results and comparisons. Finally conclusions are presented in Section 5.

2 LDA and Small Sample Size Problem

2.1 Linear Discriminant Analysis

LDA is a dimension reduction and classification method, where the data are projected into a low dimension space such that the classes are well separated. As we use this method for binary classification problems, we shall restrict the explanations to this case. We consider a set of n samples belonging to two classes C_1 and C_2, with n_1 samples in C_1 and n_2 samples in C_2. Each sample is described by q variables. So the data form a matrix $X = (x_{ij}), i = 1, \ldots, n; j = 1, \ldots, q$. We denote by μ_k the mean of class C_k and by μ the mean of all the samples:

$$\mu_k = \frac{1}{n_k} \sum_{x_i \in C_k} x_i \text{ and } \mu = \frac{1}{n} \sum_{x_i} x_i = \frac{1}{n} \sum_k n_k \mu_k$$

The data are described by two matrices S_B and S_W, where S_B is the between-class scatter matrix and S_W the within-class scatter matrix defined as follows:

$$S_B = \sum_k n_k(\mu_k - \mu)(\mu_k - \mu)^t \tag{1}$$

$$S_W = \sum_k \sum_{x_i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^t \tag{2}$$

If we denote by S_V the covariance matrix for all the data, we have $S_V = S_B + S_W$.

LDA seeks a linear combination of the initial variables on which the means of the two classes are well separated, measured relatively to the sum of the variances of the data assigned to each class. For this purpose, LDA determines a vector w such that $w^t S_B w$ is maximized while $w^t S_W w$ is minimized. This double objective is realized by the vector w_{opt} that maximizes the criterion:

$$J(w) = \frac{w^t S_B w}{w^t S_W w} \tag{3}$$

One can prove that the solution w_{opt} is the eigen vector associated to the sole eigen value of $S_W^{-1} S_B$, when S_W^{-1} exists. Once this axis w_{opt} is determined, LDA provides a classification procedure (classifier), but in our case we are particularly interested in the *discriminant coefficients* of this vector: the absolute value of these coefficients indicates the importance of the q initial variables for the class discrimination.

2.2 Generalized LDA for Small Sample Size Problems

When the sample size n is smaller than the dimensionality of samples q, S_W is singular. In this case, it is not possible to compute S_W^{-1}. To overcome the singularity problem, recent works have proposed different methods like the null space method [28], orthogonal LDA [26], uncorrelated LDA [27,26] (see also [17] for a comparison of these methods). The two last techniques use the pseudo inverse method to solve the small sample size problem and this is the approach we apply in this work. When S_w is singular, the eigen problem is solved for $S_w^+ S_b$, where S_w^+ is the pseudo inverse of S_w. The pseudo-inverse of a matrix can be computed by Singular Value Decomposition. More specifically, for a matrix A of size $m \times p$ such that $rank(A) = r$, if we denote by $A = U \Sigma V T$ the singular value decomposition of A, where U of size $m \times r$ and V of size $r \times p$ have orthonormal columns, Σ of size $r \times r$, is diagonal with positive diagonal entries, then the pseudo-inverse of A is defined as $A^+ = V \Sigma^{-1} U T$.

2.3 Application to Gene Selection

Microarray data generally contain less than one hundred samples described by at least several thousands of genes. We limit this high dimensionality by a first pre-selection step, where a filter criterion (t-statistic) is applied to determine a subset of relevant genes. In this work, we typically retain 100 genes from which

an intensive exploration is performed using a genetic algorithm to select smaller subsets. In this process, LDA is used as a classification method to evaluate the classification accuracy that can be achieved on a selected gene subset. Moreover the coefficients of the eigen vector calculated by LDA are used to evaluate the importance of each gene for class discrimination.

For a selected gene subset of size p, if $p \leq n$ we rely on the classical LDA (Section 2.1) to obtain the projection vector w_{opt}, otherwise we apply the generalized LDA (Section 2.2) to obtain this vector. We explain in Section 3 how the LDA-based GA reduces progressively the number of selected genes.

3 LDA-based Genetic Algorithm

In this section we describe our LDA-based Genetic Algorithm (LDA-GA) for gene subset selection. Notice that prior to the LDA-GA search, a filter (t-statistic) is first applied to retain a group G_p of p top ranking genes (typically $p \geq 100$, in this work, $p = 100$). Then, the LDA-based GA is used to conduct a combinatorial search within the space of size 2^p. The purpose of this search is to determine from this large search space small sized gene subsets allowing a high predictive accuracy. In what follows, we present the general procedure and then show the components of the LDA-based Genetic Algorithm. In particular, we explain how LDA is combined with the Genetic Algorithm.

3.1 General GA Procedure

Our LDA-based Genetic Algorithm follows the conventional scheme of a generational GA and uses also an elitism strategy.

- *Initial population*: the initial population is generated randomly in such a way that each chromosome contains a number of genes ranging from $p \times 60\%$ to $p \times 75\%$. The population size is fixed at 100 in this work.
- *Evolution*: the chromosomes of the current population P are sorted according to the fitness function (see Section 3.3). To generate the next population P', |P| new chromosomes are first created using crossover and mutation (see next point). These new chromosomes are then merged with the "best" 10% chromosomes of P to form P' while deleting the worst chromosomes to keep the population size constant.
- *Crossover and mutation*: mating chromosomes are determined from P by considering each pair of adjacent chromosomes (the last one is mated with the first one). By applying our specialized crossover operator (see Section 3.4), one child is created. This child then undergoes a mutation operation (see Section 3.5).
- *Stop condition*: the evolution process ends when a pre-defined number of generations is reached or when one finds a chromosome in the population having a very small gene subset (fixed at 2 genes in this work).

3.2 Chromosome Encoding

Conventionally, a chromosome is used simply to represent a candidate gene subset. Following the idea of [11], a chromosome in our GA encodes more information and is defined by a couple:

$$I = (\tau; \phi)$$

where τ and ϕ have the following meaning. The first part (τ) is a *binary vector* and effectively represents a *candidate gene subset*. Each allele τ_i indicates whether the corresponding gene g_i is selected (τ_i=1) or not (τ_i=0). The second part of the chromosome (ϕ) is a real-valued vector where each ϕ_i corresponds to the *discriminant coefficient* of the eigen vector for gene g_i. As explained in Section 2, the discriminant coefficient defines the contribution of gene g_i to the projection axis w_{opt}. A chromosome can thus be represented as follows:

$$I = (\tau_1, \tau_2, \ldots, \tau_p; \phi_1, \phi_2, \ldots, \phi_p)$$

The length of τ and ϕ is defined by p, the number of the pre-selected genes with a filter (t-statistics) (see beginning of this Section).

Notice that this chromosome encoding is more general and richer than those used in most genetic algorithms for feature selection in the sense that in addition to the candidate gene subset, the chromosome includes other information (LDA discriminant coefficients here) which are useful for designing powerful crossover and mutation operators (see Section 3.4 and 3.5).

3.3 Fitness Evaluation

The purpose of the genetic search in our LDA-GA approach is to seek "good" gene subsets having the minimal size and the highest prediction accuracy. To achieve this double objective, we devise a fitness function taking into account these (somewhat conflicting) criteria.

To evaluate a chromosome I=(τ;ϕ), the fitness function considers the classification accuracy of the chromosome (f_1) and the number of selected genes in the chromosome (f_2). More precisely, f_1 is obtained by evaluating the classification accuracy of the gene subset τ using the LDA classifier on the training dataset and is formally defined as follows[1]:

$$f_1(I) = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

where TP and TN represent respectively the true positive and true negative samples, *i.e.* the correct classifications; FP (FN) is the number of false (true) samples misclassified into the positive (negative) samples.

The second part of the fitness function f_2 is calculated by the formula:

$$f_2(I) = \left(1 - \frac{m_\tau}{p}\right) \tag{5}$$

[1] For the sake of simplicity, we use I (chromosome) instead of τ (gene subset part of I) in the fitness function even if it is the gene subset τ that is effectively evaluated.

where m_τ is the number of bits having the value "1" in the candidate gene subset τ, *i.e.* the number of selected genes; p is the length of the chromosome corresponding to the number of the pre-selected genes from the filter ranking.

Then the fitness function f is defined as the following weighted aggregation:

$$f(I) = \alpha f_1(I) + (1-\alpha) f_2(I) \text{ subject to } 0 < \alpha < 1$$

where α is a weighted parameter that allows us to allocate a relative importance factor to f_1 or f_2. Assigning to α a value greater than 0.5 will push the genetic search toward solutions of high classification accuracy (probably at the expense of having more selected genes). Inversely, using small values of α helps the search go toward small sized gene subsets. So variations of α will change the search direction of the genetic algorithm.

3.4 LDA-based Crossover

It is now widely acknowledged that, whenever it is possible, genetic operators such as crossover and mutation should be tailored to the target problem. In other words, in order for genetic operators to fully play their role, it is preferable to integrate problem-specific knowledge into these operators. In our case, we use the discriminant coefficients from the LDA classifier to design our crossover and mutation operators. Here, we explain how our LDA-based crossover operates (denoted by LDA-X hereafter).

LDA-X combines two parent chromosomes I^1 and I^2 to generate a new chromosome I^c in such a way that 1) top ranking genes in both parents are conserved in the child and 2) the number of selected genes in the child I^c is not greater than the number of selected genes in the parents. The first point ensures that "good" genes are transmitted from one generation to another while the second property is coherent with the optimization objective of small-sized gene subsets.

More formally, let $I^1=(\tau^1;\phi^1)$ and $I^2=(\tau^2;\phi^2)$ be two parent chromosomes, $I^c=(\tau^c;\phi^c)$ the child which will be generated by crossover, $\kappa \in [0,1)$ a parameter indicating the percentage of genes that will not be transmitted from the parents to the child. Then our LDA-X crossover performs the following steps to generate I^c, the child chromosome.

1. According to κ determine the number of genes of I^1 and I^2 (more precisely, τ^1 and τ^2) that will be discarded, denote them by n_1 and n_2;
2. Remove respectively from τ^1 and τ^2, the n_1 and n_2 least ranking genes according to the LDA discriminant coefficients;
3. Merge the modified τ^1 and τ^2 by the logic AND operator to generate τ^c;
4. Apply the LDA classifier to τ^c, fill ϕ^c by the resulting LDA discriminant coefficients;
5. If needed, remove the least discriminative genes from τ^c until τ^c contains no more genes than I^1 or I^2 does; update ϕ^c accordingly;
6. Create the child $I^c=(\tau^c;\phi^c)$.

Before inserting the child into the next population, I^c undergoes a mutation operation.

3.5 LDA-based Mutations

In a conventional GA, the purpose of mutation is to introduce new genetic materials for diversifying the population by making local changes in a given chromosome. For binary coded GAs, this is typically realized by flipping the value of some bits ($1 \rightarrow 0$ or $0 \rightarrow 1$). In our case, mutation is used for dimension reduction; each application of mutation eliminates a single gene ($1 \rightarrow 0$). To determine which gene is discarded, two criteria are used, leading to two mutation operators.

- *Mutation using discriminant coefficient* (M1): Given a chromosome $I=(\tau; \phi)$, we identify the smallest LDA discriminant coefficient in ϕ and remove the corresponding gene (this is the least informative genes among the current candidate gene subset τ).
- *Mutation by discriminant coefficient and frequency* (M2): This mutation operator relies on a frequency information of each selected gene. More precisely, a frequency counter is used to count the number of times a selected gene is classified (according to the LDA classifier) as the least informative gene within a gene subset. Based on this information, we remove the gene that has the highest counter, in other words, the gene that is frequently considered as a poor predictor by the classifier.

4 Datasets and Experimental Setup

4.1 Microarray Gene Expression Datasets

To assess the performance of our LDA-based genetic algorithm, we performed our experiments on seven well-known public datasets, namely Leukemia, Colon cancer, DLBCL, CNS embryonal tumor, Lung, Prostate and Ovarian cancer. A summary of the datasets is provided in Table 1.

4.2 Experimental settings

For our experimentations, we used the following experimental settings. Each initial dataset is split into a training set and a test set according to the literature. LDA-GA is applied on the training set in order to select relevant gene subsets. Because our fitness function relies on two criteria (3.3), we carry out two types of experiments. In the first one, named Exp1 hereafter, we select the gene subset according to the second criterion trying to minimize the number of selected genes. In the second type of experiments, named Exp2, we focus on the accuracy achieved by the different solutions obtained by LDA-GA and we retain the gene subsets that provide the best accuracy. Because of the stochastic nature of our LDA-GA algorithm, we run 10 executions of the GA and we retain the best solution found during these 10 executions.

In both experiments, the final predictive accuracy of a selected gene subset is estimated by the LDA-classifier built on the gene subset obtained by the training

Table 1. Summary of datasets used for experimentation

Dataset	Genes	Samples	References
Leukemia	7129	72	Golub et al [9]
Colon	2000	62	Alon et al [2]
Lung	12533	181	Gordon et al [10]
Prostate	12600	109	Singh et al [21]
CNS	7129	60	Pomeroy et al [20]
Ovarian	15154	253	Petricoin et al [19]
DLBCL	4026	47	Alizadeh et al [1]

step. As the data contain few samples, we use a 10-fold cross validation on the whole dataset to obtain a reliable estimation of the classification accuracy.

We have explained in Section 3 that our LDA-GA can apply two kinds of mutation (M1 and M2). That is why we report in the following subsection four results for each dataset: GA-M1/Exp1 is our GA with M1 mutation and we select the gene subset according to the conditions of Exp1 (focusing on the number of genes); GA-M1/Exp2 is the GA with M1 mutation and we select the gene subset according to the conditions of Exp2 (the best accuracy). Similarly, two results are reported with M2 mutation (named GA-M2/Exp1, GA-M2/Exp2).

4.3 Results and Comparisons

In this section, we propose a comparison of our LDA-GA with some state-of-the-art methods for gene selection and classification. A reliable comparison between two methods is only possible if they use the same experimental conditions. For this reason, we select 16 recent methods (since 2004) that seem to fulfill this condition.

We show in Table 2 the best results (in bold) obtained by these methods and by our LDA-based GA approach on the seven datasets presented previously. An entry with the symbol (-) in this table means that the paper does not treat the corresponding dataset. All the methods reported in this table use a process of cross validation, notice however that in some cases, the papers do not explain precisely how the experimentation is conducted.

From the results of Table 2, one observes that the proposed approach (last four lines) gives very competitive results compared with these reference methods. Indeed, our LDA-based GA achieves globally very high predictive accuracy (from 81.6% to 100%) with a very small number of selected genes (from 2 to 19).

The most remarkable results for our approach concern the DLBCL dataset. We obtain a perfect prediction with only 4 genes while the previously methods reach a prediction rate no greater than 98% with at least 20 genes. For the Ovarian cancer dataset, the LDA-GA gives a prediction accuracy of 98.4% with a subset of only 4 genes. The reference algorithms have a slightly better classification rate, but select much more genes (20, 26, 75). Notice that a perfect rate is reported in [15] with 50 genes. However the dataset used in [15] (30 cancerous and 24 normal samples, 1536 genes) is different from the Ovarian cancer dataset described in Table 1 (91 normal and 162 cancerous samples, 15154 genes).

Table 2. Results of our LDA-based GA (four last lines) compared to the most relevant works on cancer classification. The figures give the classification accuracy and in brackets, the number of genes when this is available.

Authors	Leukemia	Colon	Lung	Prostate	CNS	Ovarian	DLBCL
Ye et al [27]	97.5	85.0	–	92.5	–	–	–
Liu et al [14]	**100**(30)	91.9(30)	**100**(30)	97.0(30)	–	99.2(75)	98(30)
Tan & Gilbert [22]	91.1	95.1	93.2	73.5	88.3	–	–
Ding & Peng [7]	**100**	93.5	97.2	–	–	–	–
Cho & Won [6]	95.9 (25)	87.7(25)	–	–	–	–	93.0(25)
Yang et al [25]	73.2	84.8	–	86.88	–	–	–
Peng et al [18]	98.6 (5)	87.0(4)	**100**(3)	–	–	–	–
Wang et al [24]	95.8 (20)	**100**(20)	–	–	–	–	95.6(20)
Huerta et al [4]	**100**	91.4	–	–	–	–	–
Pang et al [16]	94.1(35)	83.8(23)	91.2(34)	–	65.0(46)	98.8(26)	–
Li et al [12]	97.1(20)	83.5(20)	–	91.7(20)	68.5(20)	99.9(20)	93.0(20)
Zhang et al [29]	**100**(30)	90.3(30)	**100**(30)	95.2(30)	80(30)	–	92.2(30)
Yue et al [28]	83.8(100)	85.4(100)	–	–	–	–	–
Hernandez et al [11]	91.5(3)	84.6(7)	–	–	–	–	–
Li et al [13]	**100**(4)	93.6(15)	–	–	–	–	–
Wang et al [23]	**100**(375)	93.5(35)	–	–	–	–	–
GA-M1/Exp1	97.2(**2**)	90.3(**2**)	97.7(**2**)	94.1(**2**)	78.3(**4**)	96.0(**2**)	91.4(**2**)
GA-M2/Exp1	97.2(**2**)	91.9(3)	98.3(2)	94.1(**2**)	85.0(**4**)	96.4(**2**)	93.6(**2**)
GA-M1/Exp2	98.6(5)	91.9(3)	97.7(**2**)	94.8(6)	81.6(8)	98.4(4)	**100**(8)
GA-M2/Exp2	**100**(5)	93.5(9)	98.3(**2**)	95.5(18)	86.6(7)	98.8(19)	**100**(4)

Finally, notice that the LDA classifier used in this paper is not the most powerful classifier. Effectively, in another experimentation, we also used a linear SVM classifier to estimate the predictive accuracy of the gene subsets selected by the LDA-GA, leading to slightly better results.

4.4 Discussion

We now discuss about two important issues of the LDA-GA approach: possible influences of the pre-selection on the prediction accuracy and the capacity of the approach to explore large sets of genes.

The search space of our LDA-GA is delimited by a first step which pre-selects a limited number (100 in this paper) of genes with the t-statistic filter criterion. One may wonder whether changing the filtering criterion and the number of selected genes affects the performance of the approach. In [5], an exhaustive study is presented concerning the influence of data pre-processing and filtering criteria on the classification performance. Three filtering criteria, BSS/WSS, t-statistic and Wilcoxon test were compared, and the results did not show any clear dominance of one criterion with respect to the others. However, the fuzzy pre-processing for date normalization and redundancy reduction presented in [5] does show a positive influence on the classification performance whatever the filtering criterion that is applied after.

The main interest of this genetic approach for gene selection is its ability to propose a combinatorial exploration of gene subsets. Clearly, this is not the case in classical approaches like backward selection. In recursive feature elimination for example, once a gene is discarded by the selection process, it is definitively ignored in the further steps even if its association to other genes can improve the classification result. Consider the Leukemia dataset, a perfect performance of 100% is reached with 5 genes (Table 2). LDA-GA also finds other gene subsets (with 5 to 10 genes) achieving a perfect cross-validation classification. More precisely, one of these subsets contains the genes placed in positions 3, 12, 63, 72, and 81 by the filter ranking criterion. Another gene subset that achieves a perfect classification contains the genes ranked in positions: 1, 2, 19, 72 and 81. Generally filter methods retain a small number of genes for classification (typically 30). Our observation shows that it is interesting and useful to explore a large set of genes because relevant subsets can contain genes that are not in the 30 top-ranking ones. Moreover, the possibility to examine diverse solutions constitutes a valuable feature for further biological investigations.

5 Conclusions

In this paper, we have introduced a new wrapper approach for selecting small gene subsets able to lead to high prediction accuracy. Our approach begins with a t-statistic filter that pre-selects a first set of genes (100 in this paper). To further reduce the gene dimension, we use a hybrid Genetic Algorithm to explore the gene subset space. The hybrid GA includes some original features that make it highly effective for identifying small sized and informative gene subsets. In particular, it uses Fisher's Linear Discriminant Analysis as its fitness function to assess the quality of each candidate gene subset. Moreover, useful discriminant information provided by the LDA classifier is directly integrated into its crossover and mutation operators. Indeed, the discriminant coefficients of LDA's eigen vector constitute a valuable indicator for recombining gene subsets (crossover) and for gene dimension reduction (mutation). The bi-criteria fitness function provides a very flexible way for the LDA-GA to explore the gene subset space either for the minimization of the selected genes or for the maximization of the prediction accuracy.

We have evaluated extensively our LDA-based GA approach on seven public datasets (Leukemia, Colon, DLBCL, Lung, Prostate, CNS and Ovarian) using a 10-fold cross validation process. A large comparison was carried out with 16 state-of-art algorithms that are based on a variety of methods. The results show clearly the interest of the LDA-GA approach for finding small sized informative gene subsets leading to high prediction accuracy. For all the datasets, our approach is able to select gene subsets of the smallest size while ensuring the best or the second best classification rate. For one dataset (DLBCL), we obtain the best result ever found with a perfect prediction with only 4 informative genes.

Finally, the proposed approach has another practically useful feature for biological analysis. In fact, instead of producing a single solution (gene subset),

our approach can easily and naturally provide multiple non-dominated solutions that constitute valuable candidates for further biological investigations.

Acknowledgments: This work is partially supported by the French Ouest Genopole and the Bioinformatics Program of the Region "Pays de La Loire". The first author of the paper is supported by a CoSNET scholarship. We would like to thank the referees for their helpful questions and suggestions.

References

1. A. Alizadeh, B.M. Eisen, R.E. Davis et al. Distinct types of diffuse large (b)–cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, February 2000.
2. U. Alon, N. Barkai, D. Notterman et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci. USA.*, 96:6745–6750., 1999.
3. A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7(3-4):559–583, 2000.
4. E. Bonilla Huerta, B. Duval, and J.K. Hao. A hybrid ga/svm approach for gene selection and classification of microarray data. In Proc. of EvoBIO'06, *Lecture Notes in Computer Science* 3907: 34–44, 2006.
5. E. Bonilla Huerta, B. Duval, and J.K. Hao. Fuzzy logic for elimination of redundant information of microarray data. *Genomics, Proteomics and Bioinformatics*. To appear, June 2008.
6. S.-B. Cho and H.-H. Won. Cancer classification using ensemble of neural networks with multiple significant gene subsets. *Applied Intelligence*, 26(3):243–250, 2007.
7. C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinformatics and Computational Biology*, 3(2):185–206, 2005.
8. S. Dudoit, J. Fridlyand, and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87, 2002.
9. T. Golub, D. Slonim, P. Tamayo et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
10. G.J. Gordon et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res.*, 62:4963–4963, 2002.
11. J. C. Hernandez Hernandez, B. Duval, and J.K. Hao. A genetic embedded approach for gene selection and classification of microarray data. In Proc. of EvoBIO'07, *Lecture Notes in Computer Science* 4447: 90–101, 2007.
12. G-Z. Li, X-Q. Zeng, J.Y. Yang, and M.Q. Yang. Partial least squares based dimension reduction with gene selection for tumor classification. In *Proc. of 7th IEEE Intl. Symposium on Bioinformatics and Bioengineering*, pages 1439–1444, 2007.
13. S. Li, X. Wu, and X. Hu. Gene selection using genetic algorithm and support vectors machines. *Soft Comput.*, 12(7):693–698, 2008.
14. 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*, 5:136:1–12, 2004.

15. E. Marchiori and M. Sebag. Bayesian learning with local support vector machines for cancer classification with gene expression data. In Proc. of EvoWorkshops'05, *Lecture Notes in Computer Science* 3449: 74–83, 2005.
16. S. Pang, I. Havukkala, Y. Hu, and N. Kasabov. Classification consistency analysis for bootstrapping gene selection. *Neural Computing and Appli.*, 16:527,539, 2007.
17. H. Park and C. Park. A comparison of generalized linear discriminant analysis algorithms. *Pattern Recognition*, 41(3):1083–1097, 2008.
18. Y. Peng, W. Li, and Y. Liu. A hybrid approach for biomarker discovery from microarray gene expression data. *Cancer Informatics.*, pages 301–311., 2006.
19. E.F. Petricoin, A.M. Ardekani, B. Hitt, P. Levine, S. Steinberg, G. Mills, C. Simone, D. Fishman, E. Kohn, and L.A. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359:572–577, 2002.
20. S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, T.R., and Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.
21. D. Singh, P.B. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. Amico, and J.P. Richie et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002.
22. A.C. Tan and D. Gilbert. Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinformatics* 2(3 Suppl):75–83, 2003.
23. S. Wang, H. Chen, S. Li, and D. Zhang. Feature extraction from tumor gene expression profiles using DCT and DFT. In Proc. of EPIA Workshops'07, *Lecture Notes in Computer Science* 4874: 485–496, 2007.
24. Z. Wang, V. Palade, and Y. Xu. Neuro-fuzzy ensemble approach for microarray cancer gene expression data analysis. In *Proc. Evolving Fuzzy Systems.*, pages 241–246, 2006.
25. W-H. Yang, D-Q. Dai, and H. Yan. Generalized discriminant analysis for tumor classification with gene expression data. *Machine Learning and Cybernetics.*, pages 4322–4327, 2006.
26. J. Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6:483–502, 2005.
27. J. Ye, T. Li, T. Xiong, and R. Janardan. Using uncorrelated discriminant analysis for tissue classification with gene expression data. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 1(4):181–190, 2004.
28. F. Yue, K. Wang, and W. Zuo. Informative gene selection and tumor classification by null space LDA for microarray data. In Proc. of ESCAPE'07, *Lecture Notes in Computer Science* 4614: 435–446, 2007.
29. L. Zhang, Z. Li, and H. Chen. An effective gene selection method based on relevance analysis and discernibility matrix. In Proc. of PAKDD'07, *Lecture Notes in Computer Science* 4426: 1088–1095, 2007.