An efficient embedded gene selection method for microarray gene expression data

Edmundo Bonilla Huerta, José C. Hernández Hernández,
Roberto Morales Caporal, José F. Ramírez Cruz and
Luis A. Hernández Montiel

LITI, Instituto Tecnológico de Apizaco,
Av. Instituto Tecnológico S/N. C.P. 90300
{edbonn,josechh,robertomorales,framirez}@itapizaco.edu.mx
Paper received on 17/07/10, Accepted on 23/09/10.

Abstract. In this paper an embedded LDA based gene selection method is proposed. We combine a Genetic Algorithm (GA) with Fisher Linear Discriminant Analysis (LDA) for selecting a small subset of genes of microarray gene expression data. This LDA-based GA algorithm uses a LDA classifier in its fitness function and uses the discriminant coefficients of LDA in its dedicated crossover and mutation operators. This paper studies the effect of these specialized operators on the evolutionary process. The proposed algorithm is assessed on a several well-known datasets from the literature and compared with recent algorithms. The results obtained show that our filter-embedded approach obtains globally high classification accuracies with very small number of genes to those obtained by other methods.

Key words: LDA, genetic algorithm, embedded, filter, gene selection, microarray data.

1. Introduction

Feature selection consists to select a minimal subset of *m* features from the original set of n features ($m << n$)[11]. Recently, those methods have been utilized for gene selection in microarray data. Feature selection methods may be categorized into three main families [9] 1) the filter approach, 2) the wrapper approach and 3) the embedded approach.

The filter approach select gene subsets independently of the learning algorithm that is used for classification. In most cases, the selection relies on an individual evaluation of each gene [2], therefore the interactions between genes are not taken into account. The filter methods separate the gene selection process from the classification process.

In contrast, the embedded approach evaluates each candidate gene subset according to their quality to improve sample classification accuracy. Embedded methods are generally computation intensive since the classifier must be trained for each candidate subset. Several strategies can be considered to explore the space of

Advances in Intelligent and Information Technologies.
Research in Computing Science 50, 2010, 2010, pp.289 -299.

possible subsets. Recently, evolutionary algorithms have been proposed for the analysis of microarray gene expression data [12, 22, 13, 15].

Finally, in embedded methods, an inductive algorithm is used as feature selector and classifier. A representative work of this approach is the method that uses support vector machines with recursive feature elimination (SVM/RFE) [1].

In this paper, we propose an embedded GA-LDA-based gene selection approach for gene subset selection for microarray gene expression data where Fisher's Linear Discriminant Analysis (LDA) is used to provide useful information to a Genetic Algorithm (GA) for an efficient exploration of gene subsets space. LDA has been used for several classification problems and recently for microarray data [7, 27, 28]. Experimental results show that our approach obtains globally high classification accuracies with very small number of genes to those obtained by others similar methods.

The organization of the rest of this paper is as follows. Section 2 describe in detail the Fisher's LDA method and discusses the calculus that must be done in the case of the small sample size problem. Section 3 presents our embedded method for gene selection. Section 4 shows the experimental results on seven microarray datasets and presents a table of comparisons with other well-known gene selection methods. Finally, conclusions are presented in Section 5.

2. LDA and Small Sample Size Problem

2.1. Linear discriminant analysis

LDA is one of the most effective dimension reduction and classification methods, where the data are projected into a low dimension space according to Fisher's criterion, such that the classes are well separated. As we use this method for binary classification problems, we shall restrict the explanations to this case. Given a data set of n samples consisting of 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:

$$S_B = \sum_k n_k(\mu_k - \mu)(\mu_k - \mu)^t \quad (1)$$

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

If we denote by SV 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:

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

One can prove that the solution w_{opt} is the eigen vector associated to the sole eigen value of $S^{-1}{}_W S_B$, when $S^{-1}{}_W$ 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, and it is not possible to compute $S^{-1}{}_W$. 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 $A = U\_V\, T$ the singular value decomposition of A, where U of size $m \times r$ and V of size $r \times p$ have orthonormal columns, $\sum$ of size $r \times r$, is diagonal with positive diagonal entries, then the pseudo-inverse of A is defined as $A^+ = V \sum^{-1} UT$.

3. Embedded method

In this section we describe our embedded method (LDA-GA) for gene subset selection. First, we apply a filter Bss/Wss [7] to retain a group *Gp* of *p* top ranking genes (In this work about p = 150). Then, the LDA-based GA is used to conduct a combinatorial search within the space of size 2p. 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 schema 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 $p \times 0.6$ to $p \times 0.75$. 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). The "best" 10% chromosomes of *P* are directly copied to the next population P^1 and removed from *P*. The remaing 90% chromosomes of P^1 are then generated by using crossover and mutation.
- Crossover and mutation: Mating chromosomes are determined from the remaining chromosomes of *P* by considering each pair of adjacent chromosomes. By applying our multi-parent recombination operator (see Section 3.4), one child is created each time. This child undergoes then a mutation operation (see Section 3.5) before joining the next population P^1.
- Stop condition: The evolution process ends when a pre-defined number of generations is reached (fixed at 250 generations in this work).

3.2. Chromosome encoding

Conventionally, a chromosome is used simply to represent a candidate gene subset. In our GA, a chromosome 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 represents effectively a candidate gene subset. Each allele τ_i indicates whether the corresponding gene g_i is selected (τ_i =1) or not selected (_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 be thus 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. 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]:

$$f_1(I) = AccuracySVM \tag{4}$$

We apply 10-fold cross-validation error estimation to calculate the accuracy of the classifier SVM for each candidate gene subset. The second part of the fitness function *f2* is calculated by the formula:

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

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

[1] For simplicity reason, 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.

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 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 toward small sized gene subsets. So varying α will change the search direction of the genetic algorithm.

3.4 Informative LDA-based Crossover

We use the discriminant coefficients from the LDA classifier to design our specialized operators (crossover and mutation). Here, we explain how our LDA-based specialized genetic operators operates.

The crossover combines two parent chromosomes I_1 and I_2 to generate a new chromosome I_{new} 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_{new} is no 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.

Before inserting the child into the next population, I^c undergoes a mutation operation based in the LDA-coefficients to remove the gene having the lowest discriminant coefficients.

3.5 Informative LDA-based Mutation

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→0, or 0→1). In our case, mutation is used for dimension reduction; each application of mutation eliminates a single gene (1→0). To determine which gene is removed, we use discriminant coefficients obtained from LDA classifier. Given a candidate gene subset, we identify the smallest LDA discriminant coefficient and remove the corresponding gene, this is the least informative genes among the current candidate gene subset.

4. Experiments on microarray datasets

4.1 Microarray gene expression datasets

In this section, we use 7 public microarray datasets for our experiments (more details in table 1). In this table is shown the number of genes, the number of samples and the first publication that has presented an analysis of this dataset.

Table 1. Summary of datasets used for experimentation.

Dataset	Genes	Samples	References
Leukemia	7129	72	Golub et al [2]
Colon	2000	62	Alon et al [4]
Lung	12533	181	Gordon et al [8]
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 [3]

4.2 Experimental results

In order to enable a fair comparison, all the crossover operators were tested under the same conditions on seven microarray datasets (Leukemia, Colon cancer, Lung cancer, Prostate cancer, Ovarian, CNS and DLBCL). The following parameters were used in this experiment: a) population size $|P| = 50$, b) maximal number of generations is fixed at 250, c) individual length (number of pre-selected genes) $p = 150$. We use a classical mutation where each bit of an individual has a mutation probability of 0.01. For the single point and uniform crossover operators, we use a crossover probability of 0.875, whereas the general settings for our LDA based crossover operator are explained in subsection 3.4.

The results of figure 1 shown that α does have a clear influence on the search direction of the genetic algorithm. A smaller value of α allows the search to obtain smaller sized gene subsets at the price of lower classification accuracy and vice-versa.

We note that the embedded approach is able to achieve very good performance even with a small population of 30 individuals. However the population size of 100 provides a slight improvement in the sense that it offers a better compromise between two objectives: good classification accuracy and small number of genes.

Table 2 summarizes the best accuracies (in bold) obtained by other methods and by our filter-embedded 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. Each cell contains the classification accuracy and the number of genes when this is available. We remark that each cell contains the classification accuracy and the number of genes when this is available.

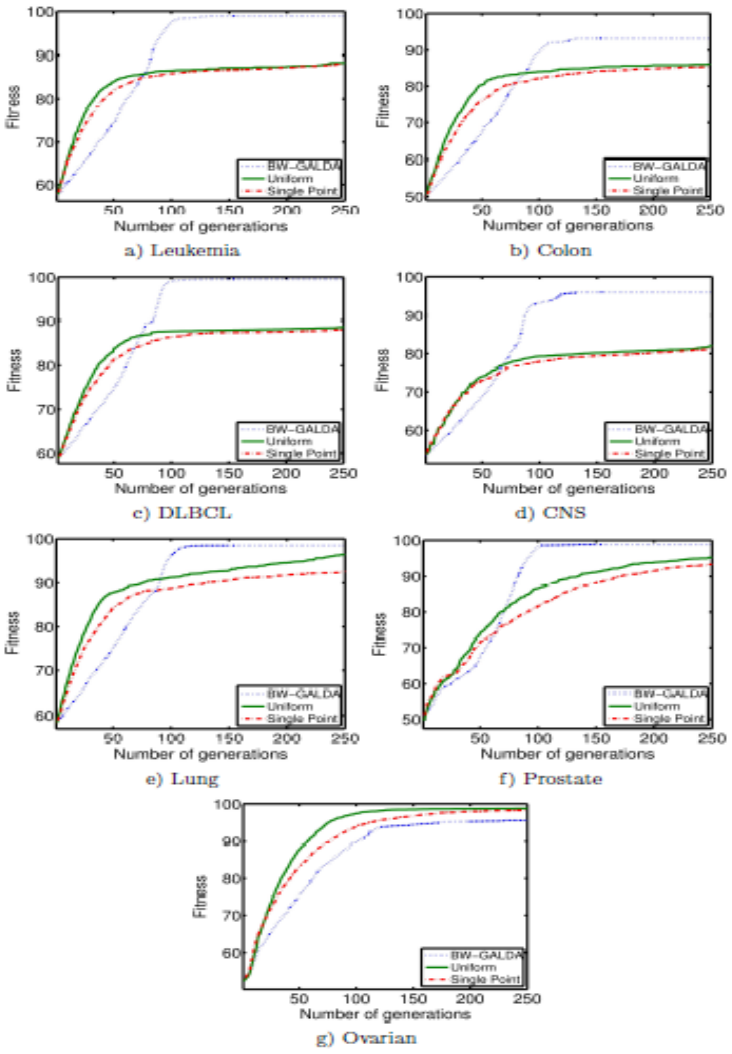

Fig. 1. Evaluation of our method with $\alpha = 0.50$

According to these observations, it seems clear that our embedded approach is very competitive in comparison with some top-performing methods. It is difficult to assess this statement by statistical tests since many methods only deal with two datasets

Table 2. Results of our embedded LDA-based GA (two last lines) compared to the most relevant works on cancer classification.

Author	Leukemia	Colon	Lung	Prostate	CNS	Ovarian	DLBCL
[6]	**100**	93.5	97.2	–	–	–	–
[5]	95.9(25)	87.7(25)	–	–	–	–	93.0(25)
[25]	73.2	84.8	–	86.88	–	–	–
[23]	95.8(20)	**100**(20)	–	–	–	–	95.6(20)
[16]	94.1(35)	83.8(23)	91.2(34)	–	65.0(46)	98.8(26)	–
[14]	97.1(20)	83.5(20)	–	91.7(20)	68.5(20)	**99.9**(20)	93.0(20)
[29]	**100**(30)	90.3(30)	**100**(30)	95.2(30)	80(30)	–	92.2(30)
[28]	83.8(100)	85.4(100)	–	–	–	–	–
[15]	**100**(4)	93.6(15)	–	–	–	–	–
our model	**100**(**4**)	95.1(**3**)	99.3(**2**)	**98.0**(**4**)	**98.8**(**2**)	99.3(**2**)	**100**(**3**)

5. Conclusions and discussion

In this paper we proposed an embedded method with specialized genetic operators for the gene selection and classification of microarray gene expression. The propose approach begins with the B/W filter that pre-selects the first 150 top-ranked genes. To further explore the combinations of these genes, we rely on a hybrid Genetic Algorithm combined with Fisher's Linear Discriminant Analysis. In this LDA-GA, LDA is used not only to assess the fitness of a candidate gene subset, but also to inform the crossover and mutation operators. This GA and LDA hybridization makes the genetic search highly efficient for identifying small and informative gene subsets.

We use a double function fitness that provides a interesting way for the LDAGA to explore the gene subset space either for the minimization of the selected genes or for the maximization of the prediction accuracy.

We have extensively evaluated our embedded approach on seven public datasets using a rigorous 10-fold cross validation process. A large comparison was carried out with 10 state-of-art algorithms that are based on a variety of methods. The results clearly show the competitiveness of our filter-embedded approach. For all the datasets, our approach is able to select small gene subsets while ensuring the best or the second best classification rate. 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 supported by the PROMEP project ITAPIEXB-000.

References

1. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3):389–422, 2002.
2. 2. T. Golub, D. Slonim, et al. M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286:531–537, 1999.

3. A. Alizadeh, M.B. Eisen, et al. Distinct types of diffuse large (b)–cell lymphoma identified by gene expression profiling. Nature, 403:503–511, February 2000.
4. U. Alon, N. Barkai, 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.
5. 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.
6. C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. Bioinformatics and Computational Biology, 3(2):185–206, 2005.
7. 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.
8. G.J. Gordon, R.V. Jensen, et al. 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 Research, 17(62):4963–4967, 2002.
9. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. Machine Learning Research, 3:1157–1182, 2003.
10. K-J. Kim and S-B. Cho. Prediction of colon cancer using an evolutionary neural network. Neurocomputing, 61:361–379, 2004.
11. R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, pages 1137–1143, 1995.
12. L. Li, C.R. Weinberg, T.A. Darden and L.G. Pedersen. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method.. Bioinformatics, 17(12):1131–1142, 2001.
13. S. Lee. Mistakes in validating the accuracy of a prediction classifier in highdimensional but small-sample microarray data. Statistical Methods in Medical Research, 17:635–642, 2008.
14. 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 Proceedings of IEEE 7th International Symposium on Bioinformatics and Bioengineering, pages 1439–1444, 2007.
15. S. Li, X. Wu, and X. Hu. Gene selection using genetic algorithm and support vectors machines. Soft Comput., 12(7):693–698, 2008.
16. S. Pang, I. Havukkala, Y. Hu, and N. Kasabov. Classification consistency analysis for bootstrapping gene selection. Neural Computing and Applications, 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, 2:301–311, 2006.
19. 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. Lancet, 359:572–577, 2002.
20. S. L. Pomeroy, P. Tamayo, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature, 415:436–442, 2002.
21. 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, 1:203–209, 2002.
22. F. Tan, X. Fu, et al., Improving Feature Subset Selection Using a Genetic Algorithm for Microarray Gene Expression Data, CEC-IEEE, 2529–2534, 2006.

23. 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.
24. P. Yang and Z. Zhang. Hybrid methods to select informative gene sets in microarray data classification. In Australian Conference on Artificial Intelligence, pages 810– 814, 2007.
25. W-H. Yang, D-Q. Dai, and H. Yan. Generalized discriminant analysis for tumor classification with gene expression data. Machine Learning and Cybernetics., 1:4322–4327, 2006.
26. J. Ye. Characterization of a family of algorithms for generalized discriminant analysis on under sampled 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 ESCAPE'07, volume 4614 of Lecture Notes in Computer Science, pages 435–446. Springer, 2007.
29. L. Zhang, Z. Li, and H. Chen. An effective gene selection method based on relevance analysis and discernibility matrix. In PAKDD, volume 4426 of Lecture Notes in Computer Science, pages 1088–1095, 2007.