A hybrid genetic algorithm and artificial immune system for informative gene selection

Mohammed Korayem[†], Waleed Abo Hamad [††] and Khaled Mostafa[††]

[†] Faculty of Computers and Information, Fayoum University, Fayoum, Egypt.
[††] Faculty of Computers and Information, Cairo University, Cairo, Egypt.

Summary
In this paper, we present a general approach for gene selection of high dimensional DNA Microarray data. The proposed approach represents a powerful new tool in the analysis and exploration of complex data. Very few genes are assumed to anticipate the pathological behavior of cancers. To this end, we proposed a hybrid between genetic algorithms and artificial immune system method; it takes into account the main immune aspects: selection and cloning of the most stimulated cells, death of non-stimulated cells, affinity maturation and reselection of the clones with higher affinity, generation and maintenance of diversity, hypermutation proportional to the cell affinity. The proposed approach is experimentally evaluated on the widely studied Colon, Leukemia and Lymphoma data sets. The results show that our approach is able to obtain very high classification accuracy which emphasizes the effectiveness of the selected genes and its ability of filtering the data from irrelevant genes. Also the criterion of the number of genes was integrated into the fitness function. Obtaining multimodal solutions is a major strength point of our method, only biologists and medical scientists can say which one of these solutions (gene subsets) is more biologically relevant to cancer diagnosis.

Key words:
Artificial immune system, Genetic algorithms, DNA microarray Data, Classification, Gene Selection.

1. Introduction

DNA microarray technology has greatly influenced the realms of biomedical research, with the hopes of significantly impacting the diagnosis and treatment of diseases. Microarray data has opened new possibilities and challenges in genetic studies. A basic assumption of the genetic studies is that the genome carries all the information about the characteristics and the development of an organism. Therefore an understanding of the genome would bring more objectivity in the problem under study. Gene expression microarrays allow measuring simultaneously the expression level of a great number of genes in tissue samples on a single microscope slide. Gene expression level indicates the amount of mRNA produced in a cell during protein synthesis and is thought to be correlated with the amount of corresponding protein. An important application of this technology is the prediction of disease state of a patient based on a signature of the gene expression levels. Such a diagnostic signature is typically derived from a dataset consisting of the gene expression measurements of a series of patients.

The primary objective is to build a classifier which classifies a new sample as accurately as possible into one of the diagnostic categories, for example tumor/normal tissue, or benign/malignant. Another objective is to find a small number of genes, i.e. a signature, often referred to as 'biomarkers' that may be useful in segregating patients in diagnosis, prognosis and for appropriate therapeutic selection in clinical management This process of identifying the genes relevant to the classification task is known as feature selection. Gene feature selection also allows the discovery of the genetic network structure or of the genetic mechanisms which are responsible for the onset and progress of a disease.

In general, since selected genetic markers contain the necessary "expression signatures" of important biological states (i.e., cancer, metastasis, etc.) they may provide guidance in experimental investigation of the pathogenesis of cancer. Researchers need to interpret results in the context of the inductive biases of each gene selection method before using these results to design expensive and labor-intensive experiments In effect, microarray studies provide geneticists with a short-list of genes worth investing hard-won funds into investigating.

Feature selection identifies the subset of differentially-expressed genes that are potentially relevant for distinguishing the classes of samples. The selected gene set should be small enough to allow diagnosis even in regular clinical laboratories and ideally identify genes involved in cancer-specific regulatory pathways.

Many selection techniques and methods, in particular Genetic Algorithms (GAs), have been developed to select

Manuscript received July 5, 2010
Manuscript revised July 20, 2010

informative genes in microarray data [21,20,16,8,5]. Section 2 gives a review of some of the most popular methods.

In this paper, gene selection and classification of DNA Microarray data is our major concern in order to distinguish tumor samples from normal ones. Therefore, we propose a novel approach for informative gene selection through adaptive search which is inspired from the artificial immune system and the genetic algorithms. The natural immune system uses a variety of evolutionary and adaptive mechanisms to protect organisms from foreign pathogens and misbehaving cells in the body. Artificial immune systems (AIS) seek to capture some aspects of the natural immune system in a computational framework, either for the purpose of modeling the natural immune system or for solving engineering problems [18]. The genetic algorithm is a probabilistic search algorithm that iteratively transforms a set of individuals (the population) into a new population of offspring individuals using the Darwinian principle of natural selection and using operations that are patterned after naturally occurring genetic operations, such as crossover and mutation. The proposed method is experimentally assessed on three well-known cancer datasets (Leukemia [12], Colon [2], and Lymphoma [1]). Comparisons with other selection methods show highly competitive results.

2. Review of Feature Selection Methods

In the literature there are three main approaches to solve this problem: the filter approach[9, 10, 12], the wrapper approach [16], and the embedded approach [13, 14]. In the filter approach, feature selection is performed without taking into account the classification algorithm that will be applied to the selected features. So a filter algorithm generally relies on a relevance measure that evaluates the importance of each feature for the classification task. A feasible approach to filter selection is to rank all the features according to their interestingness for the classification problem and to select the top ranked features. The drawback of such a method is to score each feature independently while ignoring the relations between the features. In contrast, wrapper approach, the gene subset selection algorithm conducts the search for a good subset by using the classifier itself as a part of evaluation function. The classification algorithm is used to evaluate each gene subset. Numerous search algorithms have been used to find an optimal gene subset. Evolutionary computation methods have been used to tackle this search problem which has advantage over ranking based gene selection method because different subsets of genes are evaluated in evolutionary computations through generation of different individuals of a population. In [17], a multi-objective evolutionary algorithm (MOEA) is used with the weighted voting classifier proposed by [12]. In [20], a probabilistic model building genetic algorithm (PMBGA) is presented as a gene selection algorithm.

Finally, in embedded methods, feature selection is performed as a part of the training process. An example of this approach is the method that uses support vector machines with recursive feature elimination (SVM/RFE) [13]. Another example is given in [14] which composed of a pre-selection phase according to a filtering criterion and a genetic search phase to determine the best gene subset for classification. In this sense, embedded methods are an extension of the wrapper models.

3. Proposed Gene Selection Method

Gene expression data is represented in an M by N matrix, where element xij represents the expression level of gene i under sample j. Such a matrix $X \in R^{M \times N}$, with M rows and N columns, is defined by its set of rows G = $\{g_1, \ldots\ldots\ldots\ldots g_M\}$ and its set of columns S = $\{s_1, \ldots\ldots\ldots\ldots s_M\}$.

We will use the term individual to mean a gene subset which may be a possible solution of the problem at hand. Possible solutions (individuals) in GA will be called chromosomes which consist of binary string of 0's and 1's of length M. When a bit i have a value of 1 this means the corresponding gene i is selected. Antibodies will be used to refer to individuals in AIS. Binary Hamming shape-space will be used, in which each antibody is represented by a bit string of length M.

Randomly fixed-length binary strings for L individuals were first generated to build up the initial population. Each string represents a gene subset and the values at each position in the string are coded as either presence or absence of a particular gene. Then, we calculate the fitness (i.e., how well a gene subset survives over the specified evaluation criteria) for each gene subset. A one-point crossover was then applied to form the new population. One crossover point (locus) is selected randomly, binary string from beginning of chromosome to the crossover point is copied from one parent and the rest is copied from the second parent. Parents are selected to crossover according to their fitness. The better the chromosomes are, the more chances to be selected. Roulette Wheel selection method was used to select parents. Crossover greatly accelerate search early in evolution of the population toward promising regions of the search space, also leads to effective combination of schemata(sub-solutions on different chromosomes) results in the propagation of the characteristics of the fittest individuals by exchanging genetic material [3].

Making the new population only by new offsprings can cause a loss of best chromosomes from the last population. Elitism was used to prevent losing best found solutions. Elitism first copies a few best chromosomes to the new population, the rest are replaced by the resulted offsprings form crossover process.

The clonal selection principle inspired from natural Immune system is then applied on the new resultant population. The concentration of antibodies (individuals) with high affinity (fitness) is increased in a process known as Cloning. The n highest affinity antibodies from the available antibody repertoire (population) were selected to be cloned (reproduced) independently. Because there may be multiple gene subsets (optima) that give high accuracy, the number of clones generated for each of the n antibodies is assumed to be the same. So that the number of clones generated for all these n selected antibodies is given by:

$$Lc = \sum_{i=1}^{n} \alpha L \tag{1}$$

Where LC is the total number of clones generated α is a multiplying factor, L is the total number of antibodies.

The reproduced (cloned) antibodies are then mutated with a rate that is inversely proportional to the affinity: the higher the affinity, the smaller the mutation rate. This process called somatic hyper-mutation. Somatic hyper-mutation allow the immune system to explore local areas around a specific antibody by making small steps towards an antibody with higher affinity. The affinity of the mutated clones is then calculated, and the n highest affinity mutated clones are selected and inserted in the new repertoire instead of the n lowest affinity antibodies. Hyper-mutation combined with clonal expansion is an adaptive process known as affinity maturation [4]. Maintaining multiple suitable solutions is desirable as multiple antibodies (gene subsets) can give us high accuracy. This can be accomplished by editing similar antibodies (self-reactive receptors) [11] [19].

The receptor editing process in our algorithm was accomplished by first creating a pool of distinct antibodies and then adding entirely newcomers to this pool in place of low affinity antibodies. The distinct antibodies in the pool are created such that the Hamming distance between any two antibodies is greater than a threshold ε.

The Hamming distance between any two antibodies is given by:

$$D(ab_i, ab_j) = \sum_{l=1}^{M} \omega \tag{2}$$

Where ω =1 when $ab_{il} \neq ab_{jl}$ and ω =0 otherwise.

Receptor editing offers the ability to escape from unsatisfactory local optima. Also adding a fraction of newcomer antibodies to the pool allows the diversity of the population and boarder search for the global optimum. Somatic hyper-mutation and receptor editing balance the exploitation of the best solutions with the exploration of the search space.

We have performed experiments on different microarray data sets but in each run, it terminates with many genes selected. In the research on microarray data, it is assumed that only a few genes anticipate the pathological behavior of cancers. Obtaining small subsets of selected genes with a high clustering accuracy is desirable. With this goal in mind:

1. We have designed our maturation mechanism in which the clones of n highest affinity antibodies will be muted with probability i.e., bits from the antibodies will be set to 0 with probability inversely proportional to its affinity.

2. In the final stage of the algorithm, we will add newcomers to the population with smaller subset of genes. The number of genes in these newcomer decrease from generation to the next one.

The detailed computational procedures are given in Fig. 1 as follows:

(1) P ← Generate L individuals (initial population) of different gene subsets.

(2) F ← Evaluate initial population P.

(3) Rm ← Produce m new offsprings from m parents using one-point crossover.

(4) Pd ← Retain d individuals from P with the highest evaluation (fitness),
where d =L-m.

(5) P ← Combine Pd with Rm.

(6) Pn ← Select the n highest affinity (fitness) antibodies (individuals) such that the Hamming distance between any two antibodies is greater than ε_1.

(7) CLc ← Reproduce (clone) individuals in Pn independently with the same clone number αL.

(8) C^*_{Lc} ← Submit antibodies in CLc to affinity maturation process.

(9) P^*_n ← Re-select the n highest affinity antibodies from C^*_{Lc}.

(10) P* ← Replace the lowest n affinity antibodies in P by P^*_n.

(11) P ← Select antibodies from P* such that the Hamming distance between any two antibodies is greater

than ε_2 (Receptor editing). Then add newcomers (repertoire diversity) such that the total number of antibodies is L.

Gene subset evaluation

When we evaluate a gene subset (chromosome or antibody), we don't take into consideration its accuracy on training data only but also the number of genes selected in it. In our method, the fitness (affinity) of an individual is given by:

$$F(Y) = w * ACC(Y) + (1- w) * (1- g(Y)/M) \quad (3)$$

where ACC(Y) is the accuracy on training data using only the expression levels of the selected genes in Y, g(Y) is the number of selected genes in Y, $w \in [0, 1]$ is the weight assigned to accuracy and M is the total number of genes. This type of weighted fitness was used in [17].

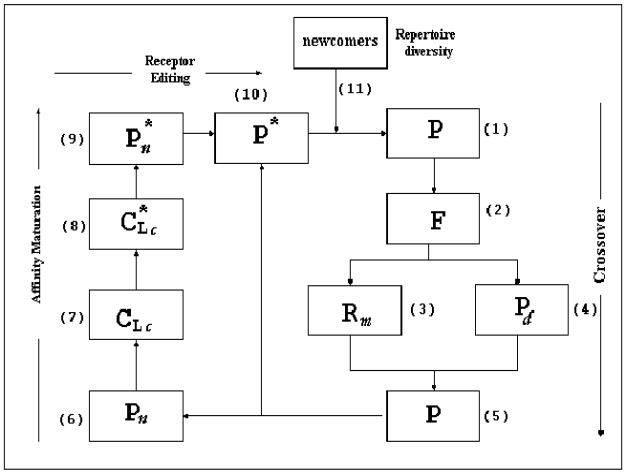

Fig. 1: Main steps in our hybrid GA-AIS algorithm

4. Accuracy Estimation

Since number of samples in training data set is small, cross-validation technique is used. In k-fold cross-validation, sometimes called rotation estimation, the data D is randomly partitioned into k mutually exclusive subsets, (D_1, D_2,…,D_k) of approximately equal size. The classifier is trained and tested k times; each time i (i = 1, 2,…, k), it is trained with Di excluded from D and tested on Di. When k is equal to the number of samples in the data set, it is called Leave-One-Out-Cross-validation (LOOCV) [15]. The cross-validation accuracy is the overall number of correctly classified samples, divided by the number of samples in the data. To avoid over-fitting, five-fold cross-validation is employed.

For the classifier, we used the GS method (sometimes called weighted voting classifier) proposed by Golub et al. [12], [23]:

$$\text{class}(x) = \text{sign}\left\{\sum\left[\left(\frac{\mu_1^g - \mu_2^g}{\sigma_1^g + \sigma_2^g}\right)\left(x_g - \frac{\mu_1^g + \mu_2^g}{2}\right)\right]\right\} \quad (4)$$

Where μ_1^g, σ_1^g and μ_2^g, σ_2^g are the mean and standard deviation for values of gene g in the classifier of class 1 and 2, respectively, and x_g is the expression values of gene g in sample x. If the computed value is positive, sample x belongs to class 1, negative value means x belongs to class 2. This classifier is only applicable to data sets with two classes. For multiclass problem, other sophisticated classifiers should be used, e.g. a SVM classifier [6].

5. Experimental Results

5.1 Data Sets

To evaluate the performance of our method, we applied it to three well-known gene expression data sets: the Colon [2], Leukemia [12], and Lymphoma [1]. The details of these data sets are summarized in table 1 after preprocessing.

Table 1: Summary of microarray data sets.

Title	# Genes	# Samples	Classes
Colon	2000	62	2
Leukemia	7129	72	2
Lymphoma	4026	96	2

Colon Data Set: The colon cancer data set contains 62 tissue samples, each with 2000 gene expression values. The tissue samples include 22 normal and 40 colon cancer cases. The data set is available at http://www.molbio.princeton.edu/colondata and was first studied in [2]. These gene expression values have been log transformed, and then normalized. The data is divided into training set and test set. The training set consists of 30 normal and 15 colon cancer cases, the test set consists of 10 normal and 7 cancer cases.

Leukemia Data Set: Leukemia Data Set is a collection of gene expressions of 7129 genes of 72 leukemia samples reported by Golub et al. [12]. The data set consists of 47 samples of Acute Lymphoblastic Leukemia (ALL) and 25 samples of Acute Myeloblastic Leukemia (AML). The data sets can be downloaded from *http://www.genome.wi.mit.edu/ MPR* .The affymetrix control-genes are first removed. Since many expression levels are

too low to be interpreted with confidence we further removed all genes where any of the gene expressions are below 20. Finally, we obtained 1762 genes. The logarithm of each value to the basis 2 is performed. This type of preprocessing has been used in [7]. The training set consists of 30 ALL and 15 AML, the test set consists of 17 ALL and 10 AML.

Lymphoma Data Set: The Diffused Large B-Cell Lymphoma (DLBCL) data set [1] contains gene expression levels of 96 normal and malignant lymphocyte samples, each measured using a specialized cDNA microarray, containing 4026 genes. The expression data in raw format are available at *http://llmpp.nih.gov/lymphoma/data/ figure1/figure1.cdt*. It contains 42 samples of DLBCL and 54 samples of other types. The training set consists of 30 DLBCLl and 36 other types, the test set consists of 12 DLBCL and 18 other types.

5.2 Experimental Setup

The parameters of gene selection algorithm are: Population size L = 100, number of parents selected for crossover m = γL, γ= 0.9, number of individuals chosen for reproducing (cloning) n = 10, number of clones for each antibody is αL, α was chosen to be 0.2, total run = 20, w = 0.9, w was chosen to give more emphasize on accuracy rather than on number of selected genes. n is crucial to the algorithms' capability of locating a large number of local optima, α is strongly related to the convergence speed and computational time required to run the algorithm. The number of generations is set to 50. At each run, the data set is split randomly into two subsets, a training set and a test set. The training set contains 2/3 of the samples and the test set contains 1/3 of the samples. After each run, instead of taking the best individual that has the highest fitness, we take all the gene subsets from the population that have the highest training accuracy and calculate test accuracy of each gene subset by the classifier.

5.3 Experimental Results and Comparisons

Here we present the experimental results of our GA/AIS method on the three datasets. In table 2, the best classification accuracy on training and test data and the number of genes selected are shown.

As shown, the highest training accuracy on Colon data is 97.78 % which is obtained with a gene subset having 8 genes, and the corresponding test accuracy is 97.78 %. Only 4 gene subsets get this high accuracy on training data as shown in table 3.

Table 2: Best results obtained by our hybrid GA/AIS method

Data set	Best training accuracy	Best test accuracy	Minimum number of selected genes
Colon	97.78 % (Test acc. =88.24%) (#Genes=8)	100 % (Train acc. =93.33%) (#Genes=12)	2 (Train acc. =95.56%) (Test acc. =64.70%)
Leukemia	100 % (Test acc. =100%) (#Genes=2)	100 % (Train acc. =100%) (#Genes=2)	2 (Train acc. =100%) (Test acc. = 100%)
Lymphoma	100 % (Test acc. =96.67%) (#Genes=10)	100% (Train acc. =98.48%) (#Genes=5)	3 (Train acc. =100%) (Test acc. =86.67%)

The maximum test accuracy on this data is 100%, the corresponding training accuracy is 93.33% and the number of selected genes is 12. The lowest number of genes in a subset is 2 which produce 95.56% and 64.70% training and test accuracy, respectively.

On Leukemia data set, the highest accuracy on both training and test (100%) is obtained by only two genes "M23197_at" and "M31523_at". Figure 2 shows that these two genes make the two classes of Leukemia data (ALL, AML) linearly separable.

On Lymphoma data set, the highest training accuracy obtained is 100 % with a gene subset having 10 genes, and the corresponding test accuracy is 96.67 %. The highest test accuracy on this data is 100%, the corresponding training accuracy is 98.84% and the number of selected genes is 5. These genes are "13978", "19292", "13071", "17791" and "16895". The minimum number of genes in a subset is 3 which produce 100% and 86.67% training and test accuracy, respectively.

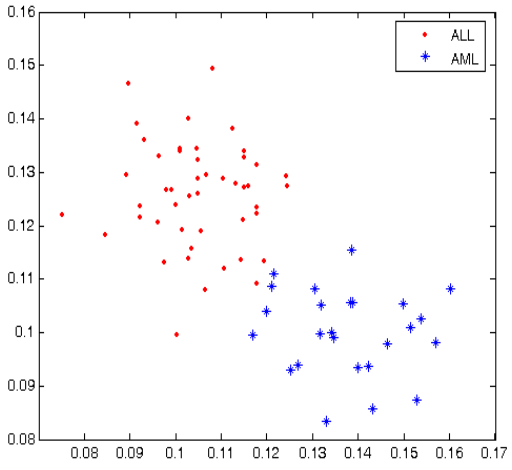

Figure 2: The two classes of Leukemia are linearly separable using the two genes selected by our method

Table 3: The four gene subsets that produce the highest training accuracy on Colon Data

Name	Gene subset 1	Gene subset 2	Gene subset 3	Gene subset 4
Hsa.2699	■			■
Hsa.27685	■	■	■	■
Hsa.19 D12765			■	■
Hsa.692 M76378	■	■	■	■
Hsa.2448	■	■	■	■
Hsa.9102		■		
Hsa.421 D16294	■			
Hsa.2291	■	■	■	■
Hsa.865 M84490	■	■	■	■
Hsa.43331		■	■	
Hsa.3952	■	■	■	■

the difference in error rate among classes and the number of selected genes. In [20], the authors present a probabilistic model building genetic algorithm (PMBGA) as a gene selection algorithm. In [14], a genetic embedded method for gene selection and classification of Microarray data is proposed. The proposed method is composed of a pre-selection phase according to a filtering criterion and a genetic search phase to determine the best gene subset for classification. Table 5 shows the average of our results on the three data sets together with those reported in [17], [20] and [14].

Table 4: Comparison of three ranking-based methods with our method.

Data set	BW ratio criteria		Correlation criteria		Fisher's Criterion		Our method (GA/AIS)	
	#genes	Acc.	#genes	Acc	#genes	Acc.	#genes	Acc
Colon	8.05±1.57	78.81	10.43±2.77	76.32	9.17±2.03	76.59	7.125±2.07	87.7
Leukemia	3.93±1.16	89.05	5.07±1.98	85.59	4.71±1.44	86.95	4.327±1.6	98.33
Lymphoma	5.96±1.31	88.27	8.01±1.94	84.47	7.13±1.86	86.02	5.829±1.464	96.6

Table 5: Comparison of other genetic approaches and our method.

Data set	[17]		[20]		[14]		Our method (GA/AIS)	
	#genes	Acc.	#genes	Acc.	#genes	Acc.	#genes	Acc.
Colon	11.4±4.27	80 ±8.3	4.44±1.74	81 ±8	7.05±1.07	84.6 ±6.6	7.125±2.07	87.7±5.06
Leukemia	15.2±4.54	90 ±7.0	3.16±1.00	90 ±6	3.17±1.16	91.5 ±5.9	4.327±1.6	98.33±1.87
Lymphoma	12.9±4.40	90 ±3.4	4.42±2.46	93 ±4	5.29±1.31	93.3 ±3.1	5.829±1.46	96.6±2.25

Comparison with ranking-based Selection Methods

Table 4 shows results of three ranking based method which are compared with results obtained by our method . these methods are The BW ratio introduced by Dudoit et al.[10] , the Correlation between a gene and a class distinction, proposed by Golub et al. [12] and The Fisher's discriminant criterion [9]. These methods were used in [14] on the same data sets . In each case, our method gets better accuracy than ranking- based methods. As shown in table 4, our method has a high accuracy on unseen data (test data).

In Table 4**:** Comparison of three ranking-based methods with our method.. Acc is the average classification rate (%) on test set. A value of the form $\mu \pm \sigma$ indicates mean value μ with standard deviation σ .

Comparison with Other Genetic Approaches

In [17], a multi-objective evolutionary algorithm (MOEA) is proposed, where the fitness function evaluates simultaneously the misclassification rate of the classifier,

Table 6 summarizes the best results obtained by our method for the Leukemia and Colon datasets together with the best results of five state-of-the-art methods from the literature. The conventional criteria are used to compare the results: the classification accuracy in terms of the rate of correct classification (first number) and the number of used genes (the number in parenthesis).

Table 6: Comparison of our GA/AIS method with five state of the art methods

Data set	Methods					
	[5]	[24]	[21]	[13]	[22]	Our method
Colon	99.83 (15)	91.9 (3)	93.55 (12)	97.0 (7)	98.0 (4)	100 (12)
Leukemia	100 (25)	100 (8)	100 (6)	100 (4)	100 (2)	100 (2)

For the Leukemia dataset, we obtain a classification rate of 100% using only 2 genes, which is much better than that reported in [5, 24, 21, 13]. However, [22] obtained the same result as our method. The most interesting results that we obtained with our model concern the Colon dataset since our approach offers the highest correct classification

rate (100%); the number of selected genes is greater than the one obtained by [24,13,22], but it is smaller than the one reported in [5] and same number of genes as in [21].

6 Conclusion

In this paper, we presented a general approach for gene selection of high dimensional DNA Microarray data. The proposed approach represents a powerful new tool in the analysis and exploration of complex data. Very few genes are assumed to anticipate the pathological behavior of cancers. To this end, we proposed a hybrid between genetic algorithms and artificial immune system method; it takes into account the main immune aspects: selection and cloning of the most stimulated cells, death of non-stimulated cells, affinity maturation and reselection of the clones with higher affinity, generation and maintenance of diversity, hypermutation proportional to the cell affinity. The proposed approach was experimentally evaluated on the widely studied Colon, Leukemia and Lymphoma data sets. The results show that our approach is able to obtain very high classification

Acknowledgments

We would like to thank Amr Badr for his helpful feedback .

References

[1] Alizadeh, A. et al., ”Distinct types of diffuse large Bcell lymphoma identified by gene expression profiling”, Nature, vol. 403, pp. 503–511, 2000.

[2] Alon U. et. al. , "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays" , Proc Natl Acad Sci USA, 96:6745–6750, 1999.

[3] Alonso J. et. al.." Mooring pattern optimization using genetic algorithms ". In 6th World Congresses of Structural and Multidisciplinary Optimization, Rio de Janeiro, Brazil. 2005.

[4] Berek C. and Ziegner M. " The maturation of the immune response". Immunology today, 14(10):400–404, Oct 1993.

[5] Bonilla Huerta E. et. al.. "A hybrid GA/SVM approach for gene
selection and classification of microarray data". Lecture Notes in Computer Science,3907:34–44, Springer, 2006.

[6] Boser B. E., et. al." A training algorithm for optimal margin classifiers". In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages 144– 152, ACM Press, 1992

[7] Busygin S., et. al.," Double conjugated clustering applied to leukemia microarray data. In Proceedings of the 2nd SIAM International Conference on Data Mining, Rio de Janeiro, Brazil. 2002.

[8] Deb K. and Reddy A. R.. "Reliable classification of two-class cancer data using evolutionary algorithms". Biosystems, 72(1-2):111–29, Nov 2003.

[9] Duda R. O. and Hart P. E.. "Pattern Classification and scene analysis". Wiley, 1973.

[10] Dudoit S., et. al." Comparison of discrimination methods for the classification of tumors using gene expression data". Journal of the American Statistical Association, 97(457):77–87, 2002.

[11] George AJ and Gray D, " Receptor editing during affinity maturation". Immunol Today, 20(4):196–196, April 1999.

[12] Golub TR, et. al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring". Science., 286(5439):531–7, Oct 1999.

[13] Guyon I., et. al. "Gene selection for cancer classification using support vector machines". Machine Learning, 46(1-3):389–422, 2002.

[14] Hernandez J. et. al. ,"A Genetic Embedded Approach for Gene Selection and Classification of Microarray Data", EvoBIO 2007.

[15] Kohavi, R., ”A study of cross-validation and bootstrap for accuracy estimation and model selection”, in Proceedings of the International Joint Conference on Artificial Intelligence, 1995.

[16] Kohavi R. and John G.H.. "Wrappers for feature subset selection". Artificial Intelligence, 97(1-2):273–324, 1997.

[17] Liu J. and Iba H. "Selecting informative genes using a multiobjective evolutionary Algorithm". In Proceedings of the 2002 Congress on Evolutionary Computation, pages 297–302, IEEE Press, 2002.

[18] Matthew Glickman, et. al.. "A machine learning evaluation of an artificial immune system". Evolutionary Computation, 13(2):179–212, June 2005.

[19] Nussenzweig MC. "Immune receptor editing; revise and select". Cell, 95(7):875–878,Dec 1998.

[20] Paul T.K. and Iba H.. "Selection of the most useful subset of genes for gene expression-based classification". Proceedings of the 2004 Congress on Evolutionary Computation, pages 2076–2083, IEEE Press, 2004.

[21] Peng S., et. al." Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines". FEBS Letters, 555(2):358–362, 2003.

[22] Reddy A. R. and Deb K.. "Classification of two-class cancer data reliably using evolutionary algorithms". Technical Report. KanGAL, 2003.

[23] Slonim D.K., et al.," Class Predication and Discovery Using Expression Data", Proc. of the 4th Annual International Conference on Computational Molecular Biology, 263-272,2000.

[24] Wang Y. et. al. "a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data'. Bioinformatics, 21(8):1530–1537, 2005

Mohammed Korayem received the B.S. and M.S. degrees in Computer Science from faculty of computers and information Cairo university, Egypt in 2002 and 2007, respectively. He is now a PhD student in computer science program at the school of informatics and computing, Indiana University, Bloomington, IN., USA.

Waleed Abo Hamad, M.Sc. is a researcher in the 3S group (A research unit in Dublin Institute of Technology (DIT) specialized in complex systems simulation and optimization). He joined the 3S group in 2008 having spent four years as a senior researcher in Cairo University-Egypt where he received his B.Sc. and M.Sc. degrees in Computer Science. Waleed has been an active member of the support team of Avicenna Knowledge Center (AKC) project funded by the European Commission and UNISCO-Paris. Currently, he is a PhD student at DIT. His research interests include Modeling and Simulation, Optimization, Computational Intelligence, Machine Learning and Cooperative Intelligent Systems.

Khaled Mostafa received the B.S. degree in Communication and M.S. and Ph.D. in Computer Science from faculty of Engineering, Cairo University, Egypt In 1986, 1993 and 1998 respectively. He is now Assoc. Prof, in Faculty of Computers and Information, Cairo University, Egypt.