A Hybrid GA/SVM Approach for Gene Selection and Classification of Microarray Data

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. We propose a Genetic Algorithm (GA) approach combined with Support Vector Machines (SVM) for the classification of high dimensional Microarray data. This approach is associated to a fuzzy logic based pre-filtering technique. The GA is used to evolve gene subsets whose fitness is evaluated by a SVM classifier. Using archive records of "good" gene subsets, a frequency based technique is introduced to identify the most informative genes. Our approach is assessed on two well-known cancer datasets and shows competitive results with six existing methods.

Keywords: Genetic algorithms, Fuzzy logic, Support vector machines, Feature selection, Classification, Microarray data.

1 Introduction

The DNA Microarray technology allows measuring simultaneously the expression level of a great number of genes in tissue samples. A number of works have studied classification methods in order to recognize cancerous and normal tissues by analyzing Microarray data [1, 8, 2]. The Microarray technology typically produces large datasets with expression values for thousands of genes (2000∼20000) in a cell mixture, but only few samples are available (20∼80).

From the classification point of view, it is well known that, when the number of samples is much smaller than the number of features, classification methods may lead to data overfitting, meaning that one can easily find a decision function that correctly classifies the training data but this function may behave very poorly on the test data. Moreover, data with a high number of features require inevitably large processing time. So, for analyzing Microarray data, it is necessary to reduce the data dimensionality by selecting a subset of genes that are relevant for classification.

In the last years, many approaches, in particular various Genetic Algorithms (GAs) and Support Vector Machines (SVMs), have been successfully applied to Microarray data analysis [6, 19, 16, 10, 15, 17, 18, 13]. In Section 3, we review some of the most popular approaches.

F. Rothlauf et al. (Eds.): EvoWorkshops 2006, LNCS 3907, pp. 34–44, 2006.

In this paper, 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 a hybrid model that uses several complementary techniques: fuzzy logic, a Genetic algorithm (GA) combined with a Support Vector Machine (SVM) and an archive-based gene selection technique. Comparing with previous studies, our approach has several particular features. First, to cope with the difficulty related to high dimensional data, we introduce a fuzzy logic based pre-processing tool which allows to reduce largely the data dimensionality by grouping similar genes. Second, our GA uses archives to record high quality solutions. These archives are then analyzed to identify the most frequently appearing genes which would correspond to the most predictive genes. Third, the GA combined with a SVM classifier is used both for selecting predictive genes and for final gene selection and classification.

The proposed approach is experimentally assessed on two well-known cancer datasets (Leukemia [8] and Colon [1]). Comparisons with six state-of-the-art methods show competitive results according to the conventional criteria.

The remainder of this paper is organized as follows. In Section 2, we describe briefly the two Microarray datasets used in this study. In Section 3, we review some popular gene selection approaches for the classification of Microarray data. In Section 4, we introduce the general scheme of our hybrid model. In Section 5, we describe our GA/SVM approach. Experimental results are presented in Section 6. Finally conclusions are given in Section 7.

2 Datasets

In this study, we use two well-known public datasets, the Leukemia dataset and the Colon cancer dataset. All samples were measured using high-density oligonucleotide arrays [2].

The Leukemia dataset[1] consists of 72 Microarray experiments (samples) with 7129 gene expression levels. The problem is to distinguish between two types of Leukemia, Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). The complete dataset contains 25 AML samples of and 47 ALL samples. As in other experiments [8], 38 out of 72 samples are used as training data (27 ALL samples and 11 AML samples) and the remaining samples (20 ALL samples and 14 AML samples) are used as test data.

The Colon cancer dataset[2] contains the expression of 6000 genes with 62 cell samples taken from colon cancer patients, but only 2000 genes were selected based on the confidence in the measured expression levels [1]. 40 of 62 samples are tumor samples and the remaining samples (22 of 62) are normal ones. In this paper, the first 31 out of 62 samples were used as training data and the remainder samples as test data.

[1] Available at: http://www.broad.mit.edu/cgi-bin/cancer/publications/.
[2] Available at: http://microarray.princeton.edu/oncology/affydata/index.html.

3 Review of Feature Selection Approaches

Feature selection for classification is a very active research topic since many application areas involve data with tens of thousands of variables [9]. This section concerns more specifically a literature review of previous studies on feature selection and classification of Microarray Data, with a special focus on the Leukemia and the Colon datasets presented in Section 2.

Feature selection can be seen as a typical combinatorial problem. Informally, given a dataset described by a large number of features, the aim is to find out, within the space of feature subsets, the smallest subset that leads to the highest rate of correct classification. Given the importance of feature selection, many solution methods have been developed. Roughly speaking, existing methods for feature selection belong to three main families [9]: the filter approach, the wrapper approach and the embedded approach.

The filter methods separate the feature selection process from the classification process. These methods select feature subsets independently of the learning algorithm that is used for classification. In most cases, the selection relies on an individual evaluation of each feature [8, 6], therefore the interactions between features are not taken into account.

In contrast, the wrapper approach relies on a classification algorithm that is used as a black box to evaluate each candidate subset of features; the quality of a candidate subset is given by the performance of the classifier obtained on the training data. Wrapper 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 possible subsets. In particular, in [14], evolutionary algorithms are used with a k-nearest neighbor classifier. In [12], the author develops parallel genetic algorithms using adaptive operators. In [18], one finds a SVM wrapper with a standard GA. In [20], the selection-classification problem is treated as a multi-objective optimization problem, minimizing simultaneously the number of genes (features) and the number of misclassified examples.

Finally, in embedded methods, the process of selection is performed during the training of a specific learning machine. A representative work of this approach is the method that uses support vector machines with recursive feature elimination (SVM/RFE) [10]. The selection is based on a ranking of the genes and, at each step, the gene with the smallest ranking criterion is eliminated. The ranking criterion is obtained from the weights of a SVM trained on the current set of genes. In this sense, embedded methods are an extension of the wrapper models. There are other variants of these approaches, see [21, 7] for two examples.

4 General Model for Gene Selection and Classification

The work reported in this paper is based on a hybrid approach combining fuzzy logic, GA and SVM. Our general model may be characterized as a three-stage sequential process, using complementary techniques to shrink (or reduce) grad-

ually the search space. The rest of this section gives a brief description of these three stages.

Stage 1 ***Pre-processing by fuzzy logic.*** This stage aims to reduce the dimension of the initial problem by eliminating gene redundancy. This stage is basically composed of four steps. First, the gene expression levels are transformed into fuzzy subsets with Gaussian representations. Second, the Cosine amplitude method is employed to assess fuzzy similarities between genes. We build a similarity matrix that is then transformed to a matrix of fuzzy equivalence relations by different compositions. Third, using $\alpha-$cuts [23] with decreasing values of α, we obtain groups of similar genes that correspond to fuzzy equivalence classes of genes. Fourth, for each group, one gene is randomly taken as the representative of the group and other genes of the group are ignored. Applying this dimension reduction technique to the datasets presented in Section 2, the set of 7129 genes for Leukemia (2000 genes for Colon respectively) is reduced to 1360 genes (943 genes respectively). Therefore, the search space is dramatically reduced. As we show later in Section 6, with this reduced set of genes, we will be able to obtain high quality classification results. A detailed description of this stage goes beyond the scope of this paper and can be found in [3].

Stage 2 ***Gene subset selection by GA/SVM.*** From the reduced set of genes obtained in the previous pre-processing stage, this second stage uses a wrapper approach that combines a GA and a SVM to accomplish the feature (gene) subset selection. The basic idea here consists in using a GA to discover "good" subsets of genes, the goodness of a subset being evaluated by a SVM classifier

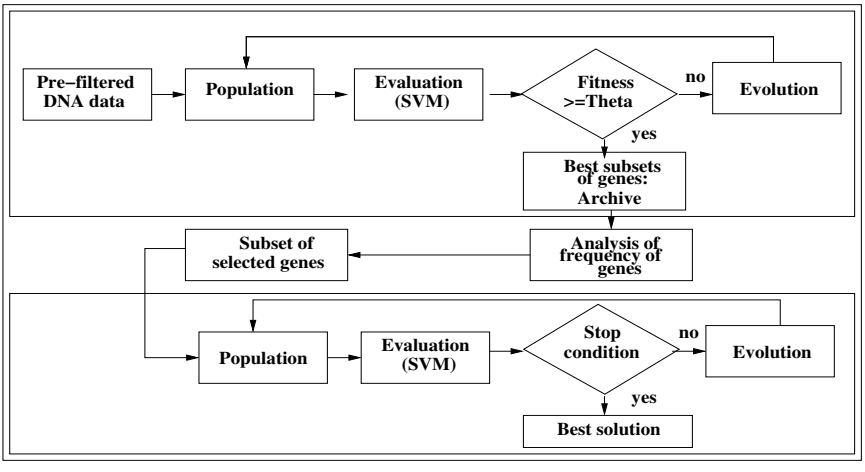

Fig. 1. The general process for gene subset selection and classification using GA/SVM: Gene subset selection (Stage 2 - top); Gene selection and classification (Stage 3 - bottom)

on a set of *training data* (see Section 2). During this stage, high quality gene subsets are recorded to an archive in order to be further analyzed.

At the end of the GA, the analysis of the archived gene subsets is performed: gene subsets are compared among them and the most frequently appearing genes are identified. This process typically leads to a further reduced set of genes (<100 genes for the Leukemia and Colon dataset). Fig.1 (top) shows a general picture of this stage.

Stage 3 *Classification.* Stage 2 has identified a reduced set of relevant genes which is now used in the final step of gene selection and classification. From this set of genes, a new round of search is carried out using the previous GA/SVM, this time to classify the *test data* (see Section 2). This stage will thus select the most predictive genes to classify the test data. Fig.1 (bottom) shows a general picture of this stage.

5 Gene Selection and Classification by GA/SVM

We describe now the hybrid GA/SVM algorithm for carrying out Stages 2 and 3 of the general model for gene selection and classification. As explained previously, the GA is designed both for discovering good gene subsets and for final gene selection and classification. The SVM-based classifier is used to ensure the fitness evaluation of each candidate gene subset. One important feature of the GA developed in this work is the use of an archive to record quality gene subsets discovered during the gene subset selection stage. This archive is then analyzed to identify a small number of highly frequently appearing genes that are used in the final classification stage. Notice that the idea of archiving good solutions is not really a new one because it is already used in some multiobjective evolutionary algorithms [26]. However, as we will see later in Section 5.3, our way of exploiting the information of the archive to identify predictive genes is original and useful.

From these retained genes obtained from archive analysis, the same GA/-SVM algorithm is applied to the test data to perform the final gene selection and classification tasks.

5.1 The Genetic Algorithm

General Schema. The basic components of our GA are presented later in this section. Here we show the general algorithm. The GA follows a generational schema with a form of elitism. To obtain a new population from the current population P, the top E% of the population P are recorded, E being fixed to 10% or 15% in our experiments (see Section 6). Then, the following two actions are taken: 1) select two parents and apply (with a given probability) the crossover to create two new solutions which are muted (with a given probability), and 2) replace the parents by their offspring. These two actions are repeated for a pre-fixed number of times. Finally, the recorded elite chromosomes are copied backed to the population P to replace the worst rated chromosomes. At this point, one generation is accomplished.

Chromosome and initial population. The chromosomes are binary-encoded, each allele (bit) of the chromosome represents a gene. If an allele is '1' it means that this gene is kept in the gene subset and '0' indicates that the gene is not included in the subset. Each chromosome represents thus a gene subset. For Stage 2 of the general model, the chromosome length is equal to the number of genes pre-selected by the fuzzy pre-processing (i.e. 1360 for the Leukemia dataset and 943 genes for the Colon dataset). For Stage 3, the chromosome length depends on the size of the gene subset retained after analyzing the solution archive (see section 5.3). In both cases, the initial population of the GA is randomly generated according to a uniform distribution.

Fitness function. The fitness of a chromosome, i.e. a subset of genes, is assessed by the classification rate on the initial datasets. In other words, a subset of genes leading to a high classification rate is considered to be better than a subset leading to a low classification rate. In our case, a SVM classifier (see Section 5.2) ensures this classification task.

Selection, crossover, mutation, and replacement. We use the roulette wheel selection and random one-point crossover and multi-uniform mutation operators. Offspring replaces always their parents. An elitism mechanism is also applied to conserve the top 10% or 15% chromosomes of the population between two successive generations.

Archives of high quality gene subsets. Given a chromosome (a candidate subset of genes), the SVM classifier gives its fitness in terms of classification rate on the training data set. If the classification rate is high enough (defined by a threshold theta, see Fig. 1.a), the subset of genes is recorded in an archive. In this paper, the threshold theta is set to 0.90 and 0.91 respectively for the Leukemia and Colon dataset.

Stopping criterion. The evolution process ends when a pre-defined number of generations is reached or a fitness value of 100% is obtained.

5.2 The SVM Classifier

Support Vector Machines [24] are basically binary classification algorithms. When the data are linearly separable, SVM computes the hyperplane that maximizes the margin between the training examples and the class boundary. When the data are not linearly separable, the examples are mapped to a high dimensional space where such a separating hyperplane can be found. The mechanism that defines this mapping process is called the kernel function. SVM are powerful classifiers with good performance in the domain of Microarray data [10, 17]. They can be applied to data with a great number of genes, but it has been showed that their performance is increased by reducing the number of genes [6, 2].

In our wrapper GA/SVM algorithm, we use a SVM classifier to assess the quality of a gene subset. For a chromosome x that represents a gene subset, we apply a Leave-One-Out Cross-Validation (LOOCV) method to calculate the

average accuracy (rate of correct classification) of a SVM trained with this gene subset [11]. The LOOCV procedure means that one sample from the dataset is considered as a test case while a SVM is trained on all the other samples, and this evaluation is repeated for each sample. So for each chromosome x, $Fitness(x) = accuracy_{SVM}(x)$.

One of the key elements of a SVM classifier concerns the choice of its kernel. In our study, we have chosen to use the RBF kernel. We also experimented Gaussian and polynomial kernels. For polynomial kernels, the main difficulty is to determine an appropriate polynomial degree while the results we obtained with the Gaussian kernel are not satisfactory. Notice that RBF has been used in several previous studies for Microarray data classification [4, 18, 5].

5.3 Archive Analysis

At the end of stage 2 and prior to the final classification (Stage 3), the archive is analyzed and the most frequently appearing genes in the archive are retained for the final gene selection and classification (stage 3). Typically, this analysis will lead to a limited number of genes (between 50 to 100). From these genes, the GA/SVM algorithm will then determine the final set of genes relevant to classify the data.

6 Experimental Results and Comparisons

6.1 Parameters Settings

For our GA/SVM algorithm, the GA is implemented in Matlab (Version 5.3.1 for Windows). The SVM classifier is based on the SVM Toolbox developed by Gavin Cawley[3].

Table 1. GA parameters for the stage of gene subset selection (Stage 2)

Parameters	Leukemia	Colon
Size of population	500	500
Length of chromosome	1360	943
Number of generations	2500	2500
Crossover rate	0.95	0.98
Mutation rate	0.02	0.01
Elitism rate E	10%	15%

The GA parameters used in our model of gene subset selection for the Leukemia and Colon datasets are shown in Tables 1 and 2. For the SVM classifier, the same parameters settings are used in the two stages of gene subset selection and classification. The normalization parameter C is fixed at 100 and the control parameter γ for the RBF kernel of SVM is fixed to 0.5. Notice that

[3] http://theoval.sys.eua.uk/~gcc/svm/toolbox

Table 2. GA parameters for the stage of classification (Stage 3)

Parameters	Leukemia	Colon
Size of population	50	50
Length of chromosome	100	50
Number of generations	500	500
Crossover rate	0.985	0.985
Mutation rate	0.02	0.01
Elitism rate E	15%	15%

given the input data used by the GA/SVM are already normalized during the Fuzzy Logic pre-processing, the normalization parameter C has in fact little influence in our case.

6.2 Results and Comparisons

To carry out our experiments, our GA/SVM algorithm is run 5 times on each of the Leukemia and Colon datasets. To calculate the average classification rate of a given gene subset, the LOOCV procedure [11] is employed.

Table 3 summarizes our results (Column 2) for the Leukemia and Colon datasets together with the results of six state-of-the-art methods from the literature (Columns 3-8). 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, "-" indicating that the number of genes is not available). For AG/SVM, the classification rate that we present is the average classification rate obtained from the 5 independent runs and the number of selected genes is the minimum number obtained from these runs. Detailed results can be found in Table 4.

As it can be observed, for the Leukemia dataset, we obtain a classification rate of 100% using 25 gens, which is much better than that reported in [6, 5]. This same performance is achieved by [25, 18, 20, 10], with fewer genes selected. [20] and [10] reports the minimal number of genes. However, in [20] the evolutionary method begins with a largely reduced set of 50 genes, published in [8] as interesting genes.

The most interesting results that we obtained with our model concern the Colon dataset since our approach offers the highest (averaged) correct classification rate (99.41%); the number of selected genes is greater than the one obtained by [20] or by [25, 10], but it is smaller than the one reported in [18]. An analysis

Table 3. Comparison of GA/SVM with six state of the art methods

	Methods						
Dataset	GA&SVM	[6]	[25]	[18]	[5]	[20]	[10]
Leukemia	100(25)	94.10(-)	100(8)	100(6)	95.0(-)	100(4)	100(2)
Colon	99.41(10)	90.30(-)	91.9(3)	93.55(12)	91.0(-)	97.0(7)	98.0(4)

Table 4. GA/SVM performance on 5 runs

Runs	Run 1	Run 2	Run 3	Run 4	Run5	Average class. rate
Leukemia	100(25)	100(28)	100(30)	100(46)	100(35)	100
Colon	99.64(10)	99.83(15)	97.88(10)	99.83(15)	99.83(15)	99.41

of our results shows that several biologically significant genes reported in [8] are found by our approach.

Table 4 shows the detailed results of 5 independent runs of our GA/SVM algorithm. As it can be observed, these results are quite stable. For the Leukemia dataset, each of the 5 runs obtains a classification rate of 100% while for the Colon dataset, the best run gives a classification rate of 99.64. Even the worst obtains a classification rate of 97.88.

7 Conclusions

In this paper, we presented a general approach for gene selection and classification of high dimensional DNA Microarray data. This approach begins with a fuzzy logic based pre-processing technique that aims to cope with the imprecise nature of the expression levels and to reduce the initial dimension of the input dataset. Following this pre-processing stage, a hybrid wrapper system combining a Genetic Algorithm with a SVM classifier is used to identify potentially predictive gene subsets that are then used to carry out the final gene selection and classification tasks. Another important feature of our approach concerns the introduction of an archive of high quality solutions, which allows limiting the GA/SVM exploration to a set of frequently appearing genes.

This approach was experimentally evaluated on the widely studied Leukemia and Colon cancer datasets and compared with six previous methods. The results show that our approach is able to obtain very high classification accuracy. In particular, to our knowledge, this is the first time that a averaged correct classification rate of 99.41% (with 10 genes) is reached for the Colon dataset.

This approach can be further improved on several aspects. First, we notice that our method does not provide the smallest number of genes on the Leukemia data. This is due to the fact that the GA is only guided by the criterion of classification accuracy. Therefore, the criterion of the number of genes should be integrated into the fitness function. This can be achieved by an aggregated fitness function or a bi-criteria evaluation. Second, the high computation time required in stage 2 can be reduced by the use of a faster classifier (or an approximate fitness function). For example, the m-features operator reported in [22] may be considered. Also, a fine-tuning of SVM parameters in stage 3 may lead to improved results. Finally, we intend to apply our approach to other DNA chip data and to study the behavior of our model.

Acknowledgments. E. Bonilla Huerta is supported by a Mexican CoSNET research scholarship. This work is partially carried out within the French Ouest

Genopole Program. We thank the reviewers of the paper for their very helpful comments.

References

1. U. Alon, N. Barkai, D. A. 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. In *Proc. Natnl. Acad. Sci. USA*, volume 96, 1999.
2. 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.
3. E. Bonilla Huerta, B. Duval, and J.K. Hao. Feature space reduction of large scale gene expression data using Fuzzy Logic. Technical Report, LERIA, University of Angers, January 2006.
4. M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, S.W. Sugnet, T.S. Furey, M. Ares Jr., and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines *Proc. Natl. Acad. Sci. U S A.*, 97(1): 262–267, 2000.
5. S. Chao and C. Lihui Feature dimension reduction for microarray data analysis using locally linear embedding. In *APBC*, pages 211–217, 2005.
6. T. S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.
7. L. Goh, Q. Song, and N. Kasabov. A novel feature selection method to improve classification of gene expression data. In Proceedings of the Second Asia-Pacific Conference on Bioinformatics, pages 161–166, Australian Computer Society, Darlinghurst, Australia, 2004.
8. T. 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*, 286:531–537, 1999.
9. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
10. I. 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.
11. T. Joachims Estimating the Generalization Performance of a SVM Efficiently. *Proceedings of the International Conference on Machine Learning (ICML)*, Morgan Kaufman, 2000.
12. L. Jourdan. Metaheuristics for knowledge discovery : Application to genetic data (in French). PhD thesis, University of Lille, 2003.
13. K-J. Kim and S-B. Cho. Prediction of colon cancer using an evolutionary neural network. *Neurocomputing (Special Issue on Bioinformatics)*, 61:361–379, 2004.
14. 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.
15. J. Liu and H. Iba. Selecting informative genes using a multiobjective evolutionary algorithm. In *Proc. of Congress on Evolutionary Computation (CEC'02)*, pages 297–302, 2002.

16. F. Markowetz, L. Edler, and M. Vingron. Support vector machines for protein fold class prediction. *Biometrical Journal*, 45(3):377–389, 2003.
17. S. Mukherjee. *Classifying Microarray Data Using Support Vector Machines.* Springer-Verlag, Heidelberg, 2003.
18. S. Peng, Q. Xu, X.B. Ling, X. Peng, W. Du, and L. Chen. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Letter*, 555(2):358–362, 2003.
19. S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, and T.R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U S A.*, 98(26):15149–15154, 2001.
20. A. R. Reddy and K. Deb. Classification of two-class cancer data reliably using evolutionary algorithms. Technical Report. KanGAL, 2003.
21. Y. Saeys, S. Aeyels Degroeve, D. Rouze, and Y. P. Van de Peer. Feature selection for splice site prediction: A new method using eda-based feature ranking. *BMC Bioinformatics*, 5-64, 2004.
22. S. Salcedo-Sanz, F. Prez-Cruz, G. Campsand, and C. Bousoo-Calzn. Enhancing genetic feature selection through restricted search and Walsh analysis. *IEEE Transactions on Systems, Man and Cybernetics, Part C*, 34:398–406, 2004.
23. T.J. Ross. *Fuzzy Logic with Engineering Applications.* McGraw-Hill, 1997.
24. V. N. Vapnik. *Statistical Learning Theory.* Wiley N.Y., 1998.
25. Y. Wang, F. Makedon, J.C. Ford, and J.D. Pearlman. Hykgene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics*, 21(8):1530–1537, 2005.
26. E. Zitzlere, K. Deb, and L. Thiele. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary Computation* 8(2): 173-195, 2000.