A MODIFIED FRAMEWORK FOR FEATURE SUBSET SELECTION IN MICROARRAY DATA

Burak Aksu[1], Nihan Özşamlı[1], Kemal Kılıç[1], O. Uğur Sezerman[2]*

[1]Manufacturing Sys. Eng. Lab, Sabancı University
MDBF Tuzla 34956 Istanbul, Turkey
[2]Comp. Biology Lab, Sabancı University
MDBF Tuzla 34956 Istanbul, Turkey
*Phone: 90-216-4839596, email: kkilic@sabanciuniv.edu

ABSTRACT

In this paper a modified framework for feature subset selection problem is proposed. The framework suggests the cleaning of the correlated genes from the data set, forming the initial population based on an index that assess the discriminative ability of the features and application of a GA, which utilizes a fast classifier algorithm while calculating the fitness function scores. Different replications of the framework yield different feature subsets withclose classification accuracy. Since most of the highly correlated genes are eliminated during the preprocessing phase, these feature subsets have less correlation among their members. Therefore, convergence to different feature subsets might be an indicator of different pathways. The experimental analysis demonstrate that the proposed modifications improve the classification accuracies

1. INTRODUCTION

Microarray technology enabled scientists to discover gene expression data and study genetic mechanisms. The structure of the gene expression data is in a sense peculiar. There are too many features (i.e., genes) and relatively fewer number of samples, since data collection is pretty expensive. However, most of the features are not relevant within the context. Therefore, in order to gather relevant information from the data, feature subset selection has been widely applied to microarray data sets.

Feature subset selection problem is basically finding a subset of features (i.e. subset of genes) that has minimum elements, while maximizing the predictive accuracy that can perform correct classification of samples (i.e. tissue samples). The solution methods to this problem can lead us to exact prediction of classes or subclasses of tissue samples (e.g. healthy and diseased). In addition, these methods can be used to discover disease mechanisms [2].

Reaching a good subset of features that has high prediction accuracy is computationally costly. Since there are thousands of genes to start with, it is impossible to search every feature subset in reasonable time. Therefore, an intelligent algorithm is required that searches mostly the promising parts of the search space. Genetic Algorithms (GA) [3] is a global optimization metaheuristic, which has many successful applications in the literature.

Kucukural *et al.* proposed a GA for the feature subset selection algorithm in gene expression data. In this paper, a framework is proposed that utilizes a modified version of Kucukural *et al.*'s algorithm. The proposed framework consists of four steps. First step is preprocessing the data and cleaning most of the features that are highly correlated with at least one feature that remains in the data set. Next, initial population is generated based on the probabilities associated with the features. The associated probabilities are estimated with the Ranked Gene List method, which basically estimates the discriminative ability of a feature based on an index. Later, the initial gene pool is created and a modified version of Kucukural *et al.*'s GA based algorithm is utilized in order to determine the feature subsets that yields the highest classification accuracy.

In section 2, the relevant literature will be reviewed. We will also address to one of the important issues currently being discussed by the academic community, namely, ending up with distinct feature subsets with similar predictive accuracy, in the same section. Next we will provide the details of the proposed framework. The experimental analysis and results will be presented in section 4. Finally, we will provide some conclusion remarks and list some of the planned future research topics.

2. LITERATURE REVIEW

Feature subset selection problem has been receiving a lot of attention from the researchers of various fields because of its crucial role in learning from data.There are numerous different approaches proposed in the literature. These approaches are mainly classified as filter, wrapper and embedded methods[4]. Filter algorithms are efficient algorithms and widely applied to microarray data. Within the context of microarray data, the filter approach determines the ability of each gene to classify the labels accurately and selects the genes based on this computed ability score.

However, Xiong [5] states that the classification power of a set of genes may be more than the sum of their individual classification abilities. Therefore, wrapper approaches, which search the set of genes rather than individual genes, are more appropriate. Nevertheless, the combinatorial nature of the wrapper approaches results in computationally demanding algorithms. In order to search the set of all possible solutions (all possible subsets), efficiently various stochastic search techniques are proposed. Among the most applied stochastic search methods, genetic algorithms have been widely used for feature subset selection problems [6], [7], [8].

One of the common problems of feature subset selection is the selection of the classification method that will be utilized in order to estimate the classification ability of the feature subset or the individual feature. In the literature there is an abundance of different approaches for classification of the data using the given set of features. Among these approaches, Support Vector Machines (SVM) [9] has been accepted to be the prominent method, especially in microarray-based expression data [10], [11].

Liu *et al.* [12] asserted that the feature subsets that have close and high classification accuracy, turns out to share relatively few number of common genes. This raised certain issues. They investigated by biological pathway analysis, if these distinct subsets of genes could be used as cancer biomarkers and might lead to the identification of cancer subtypes which are yet unknown. According to their analysis, the pathways they attained were mostly known for their relations with cancer. Therefore, they concluded that their approach potentially could discover optimal predictor gene sets that have no dependency among each other.

Similarly, Gormley *et. al.* [13] pointed to the lack of common genes in different gene sets (expression profiles) that defined the same biological state. They explained their findings with the connection between large number of good feature sets and previously defined multiple, mostly exclusive biomarker sets. That is to say, each subset was referring to a different pathway.

Kucukural *et al* also end up with distinct feature subsets for each replication of the GA approach that they proposed [1]. They conducted correlation analysis among these distinct feature subsets. Based on correlation analysis, they concluded that identified distinct feature subsets actually were highly correlated with each other in terms of their expression level.

One of the main motivations of this paper is to investigate the reasons that lead to distinct feature subsets. For this purpose, the initial data set is pre-processed so that the number of genes that are highly correlated with each other is reduced. The pre-processing stage itself is a combinatorial problem. One must eliminate as much genes as possible while maintaining the information hidden in the expression data. Furthermore, if correctly done, even this pre-processing stage potentially improves the predictive accuracy of the feature subset reduction algorithm, since it will reduce redundancy in the search space.

The modified framework is applied to colon cancer data set[1]. There are several studies based on colon cancer data set. Furey *et al.* [11] performed a filtering and SVM classifier on colon cancer data set and found 92.4% accuracy. Using clustering based deterministic annealing algorithms, Alon *et. al.* [14] reached 90% accuracy. Wang *et al.* [15] found 12% and 11% test error for the models that included normal selection and automatic feature selection as the last step respectively; prior to this step, self organizing maps and fuzzy c-means clustering was used in colon cancer data set. Liu *et al.* [16] used entropy based greedy approach which reduced the data-set to 9 genes with 91% accuracy. Another approach presented was entropy based simulated annealing with 26 genes and 87% accuracy. A hybrid GA that included SVM was used by Huerta *et al.*[17] and %99.4 accuracy with 10 features was their best solution.

[1] Colon cancer dataset can be downloaded from http://microarray.princeton.edu/oncology/affydata/index.html

3. MODEL STATEMENT AND METHODOLOGY

The feature subset selection problem has the objective of maximizing the predictive accuracy of the classification while minimizing the number of elements in gene subset. In the model formulation X represents the set of features to be selected, the function $g(x)$ is used to measure the classification accuracy of subset x and $n(x)$ stands for the number of elements in the subset of x. In this case, we have a multi objective optimization problem which can be stated as follows:

$$\max g(x) \quad x \in X$$
$$\min n(x) \quad x \in X$$

Such a multi objective problem searches for the best subset of features in a large search space where each subset represents a possible discriminative set of features. In this paper we propose the following framework in order to obtain the most discriminative set:

A. Elimination of the correlated genes in the experi- mental data set. (*Pre-processing)*
B. Identify the probabilities associated with each gene, which will be utilized as a selection index while generating the initial population by the Ranked Gene List method [18]. (*Determination of the Probabilities)*
C. Initial population generation.
D. Application of modified version of *Kucukural et al.'s* algorithm.

A. Preprocessing

Containing large number of genes that are highly correlated with each other in the analyzed data set, yields a search space that has redundant information. This makes the identification of discriminative gene subset harder and computationally more costly. Therefore, in order to obtain a subset of relevant genes, highly correlated genes are eliminated from data set.

The colon cancer data set consists of 2000 genes, a 2000x2000 correlation matrix is constructed and we picked the genes having correlation value greater than 80%. Among those highly correlated genes a total of 1528 genes are directly eliminated from the set. The elimination process is combinatorial; one should eliminate as much genes as possible but maintain the information in the data. For this purpose we developed the algorithm depicted in Fig. 1.

In the algorithm, the colon cancer data set is formed as a graph G(N,A) where genes are represented as nodes and the arcs represent the connection between two genes that have a correlation higher than ρ (in this paper we used ρ=0.80). The degree of each node is the number of genes connected to that gene. Elimination starts with sorting the degrees of each node in descending order. For each node, each branch node's

degree is compared with the source node's degree. The node with the highest degree is kept for further iterations and the other nodes are moved to the eliminated list. If compared nodes' degrees are equal, algorithm takes one node randomly and puts the other node in a dummy list. This dummy list is used to decrease the effect of randomization in case of having equal degree. Therefore, when a node in dummy list happens to be chosen once more randomly in further iterations, that node in dummy list can be used for comparison again. If the *dummy list* is not empty at the end of iterations, those nodes in the list are moved to the *eliminated list*. With this method, repetition of highly correlated genes can be prevented; because by choosing the highest source or branch node, it is guaranteed that the eliminated genes are represented by the selected gene. This method takes 472 genes as the initial search space instead of 2000 genes for Genetic Algorithm.

1) Form graph, *G(N,A)*, for the data set
2) Compute the degrees of each node; *deg(N_i), i*
3) Sort the nodes according to degrees, *deg(N_i)*, in de- scending order.
4) For all *i,*
   - Degree of each node, *deg (N_i)*, is com- pared to degree of connected node N_j : *deg(Nj)*
   - If *deg(N_i)* > *deg(N_j)*, take N_i to the *selected list* and add N_j to the *eliminated list*.
   - If *deg(N_i)* = *deg(N_j)*, add one of the nodes randomly to the *dummy list* and the other to the *selected list* .
   - A node in the *dummy list* can be chosen randomly and used again in equal degree comparison in further iterations.
5) Delete the genes that are in the *eliminated list* from data set.

Figure 1 – Pseudo-code of preprocessing

B. Determination of the probabilities

The overall objective of this step is to associate probabilities to the genes that will specify their likeliness to be in the initial population pool. The associated probabilities are actually indicators of the discriminative capability of the genes, i.e., how promising is they are in terms of the classification accuracy. In this paper, the *Ranke Gene List* (RGL) approach [18] is utilized to determine the discriminative abilities of each gene. RGL evaluates each one of the genes discrimination ability as follows:

Let N_0 denotes the set of cancer class, and N_1denotes the set of healthy class

N_0= {$x_{0,1}$,$x_{0,2}$,...., $x_{0,n0}$}
N_1={$x_{1,1}$,$x_{1,2}$,...., $x_{1,n1}$}

where $x^{1,j}$ ($x^{0,j}$) denotes the vector of the gene expression levels for the j^{th} sample of the healthy class (cancer). The i^{th} element (referring to the i^{th} gene) of the vector is referred to as $x_i^{1,j}$. Weight of each gene (w_i) is evaluated as follows;

$$dc_i = \left| \frac{\sum_{j=1}^{n_0} x_i^{0,j}}{n_0} - \frac{\sum_{j=1}^{n_1} x_i^{1,j}}{n_1} \right| \quad (1)$$

$$md_i^0 = \sum_{j=1}^{n_0} \sum_{l \neq j} \left| x_i^{0,j} - x_i^{0,l} \right| \quad (2)$$

$$md_i^1 = \sum_{j=1}^{n_1} \sum_{l \neq j} \left| x_i^{1,j} - x_i^{1,l} \right| \quad (3)$$

$$md_i = \frac{md_i^0 + md_i^1}{\binom{n_0}{2} + \binom{n_1}{2}} \quad (4)$$

$$w_i = \frac{dc_i}{md_i + \alpha} \quad (5)$$

The distance between the mean expressions of healthy and cancer classes is dc_i; md_i^0 and md_i^1 are the distances between each class for all pairs, where the mean distance between expressions of pairs in same class is md_i. Given these variables, the weight for each gene i is calculated as w_i. The small constant α prevents dividing by zero in weight calculation. We determined α by subtracting minimum of md_i from the maximum of dc_i. This measure will guarantee α being greater than zero. From the given formulation above, the selection probability of each gene is calculated by dividing each gene weight to the sum of all genes' weights.

C. Initial population generation

As opposed to Kucukural *et al.*'s algorithm which creates initial population pool randomly, in the proposed approach, roulette wheeling is utilized based on the probabilities associated with each gene during the prior step (Step B).

D. Apply modified version of Kucukural et al.'s algorithm

Kucukural *et al.* developed a genetic algorithm framework for the feature subset selection problem. The developed GA utilizes a chromosome structure in which each chromosome refers to a feature subset. Their algorithm has a novel approach on determining the number of genes selected for classification (size of the feature subset). They propose to utilize the fitness function scores of previously created chromosomes (candidate feature subsets) while reducing the number of genes selected at certain iterations. The fitness function value of each chromosome is basically the classification accuracy of the regarding feature subset which is attained by utilizing Support Vector Machines (SVM) as the classifier. Each fitness function score of the chromosomes is assigned to the genes which construct that chromosome, and for each gene these values are summed across each chromosome of the population for each iteration. At certain iterations (it is proposed to set this parameter to be 30 for the first time – referred to as ***first point***-, and 10 for the rest –referred to as

repeat number-), these summed values for each gene is summed across the iterations, so that for each gene an *overall summation* is obtained. That is to say, the process is double summation of the fitness values over chromosomes (if the gene is included in it) and over iterations for each gene. The *overall summation* is next divided by the frequency of the occurrence of the particular gene in the chromosomes throughout those iterations, in order to attain an *average score* for each gene.

If a gene is included in many good subsets (with high classification accuracy), the score of the gene is higher than other genes. After the *average scores* are calculated at certain iteration, a totally new initial population is generated using the roulette wheeling technique based on the *average scores.* Genes with higher total score, are more likely to be selected, therefore these genes can be selected multiple times. If this occurs, number of features in a chromosome decrease. Consequently, the number of genes in a subset is decreased continuingly as the iterations are executed. For further details of the Kucukural *et al.*'s algorithm we kindly refer the reader to [1].

In this paper we propose some modifications to Kucukural *et al.*'s approach. First of all the SVM technique utilizes as the classifier is quite time consuming. On the other hand in the feature subset selection problem computational time is invaluable, hence one should be very careful where it is allocated. Rather than spending the time with the classifier while identifying the fitness function score, one must search more of the solution space. Therefore, Fisher's classifier [19] is utilized during the first iterations until first point in the algorithm, with some Parameter Sets (*during the diversification phase*) and SVM is used for the rest of the iterations (*during the intensification phase*). Furthermore, in order to guarantee the coverage of the genes particularly during the diversification phase, we ensured that every gene is selected at least four times in the population.

The initial population creation and selected parameters in Kucukural *et al.*'s are also modified, so that a better coverage of the genes is guaranteed during the initial stages where diversification is crucial. These modifications will be discussed in the next section in more detail.

4. EXPERIMENTAL ANALYSIS AND RESULTS

In this paper, the colon cancer data set is utilized to evaluate the performance of the modified algorithm. The colon cancer data consists of expression levels of 2000 genes and collected from 40 tumor and 22 healthy tissue samples.

In order to evaluate the performance of the proposed modifications we conducted experimental analysis. The ex- periments were designed based on two aspects of the re- search. First of all, different replications of Kucukural *et al.*'s GA, yield results that were converging to feature subsets consisted from highly correlated genes. Therefore, we propose to utilize preprocessing, in one sense, in order to prevent this convergence. In order to check its performance, we conducted six replications of the framework with same parameters and evaluate their correlations with each other.

Secondly, preprocessing and initial population generation algorithms, as well as the modifications we proposed for Kucukural's *e al.* GA, such as utilizing a faster classifier and choosing parameters in order to ensure broader coverage of the genes during the diversification phase, were introduced in order to improve the classification accuracy. We choose four parameter sets in order to check the performance of the proposed modifications in terms of the classification accuracy.

The first parameter set (which is also used for the six replications experiment mentioned above) have the parameters used by Kucukural *et al.* Second parameter set has relatively very high initial population size during the diversification phase (until *first point*). On the other hand, third parameter set has very high number of iterations. The parameters used in experimental design are given in Table 1.

Table 1 – Three parameter sets used in the experimental analysis

Parameter Set	1	2	3
Size of Initial Population	30	1000	50
First Point	30	30	1000
Size of Population After First Point	30	30	30
Number of Iterations After First Point	130	130	130

Throughout the experiments the GA parameters such as the crossover rate and mutation rate were set to be the values identified by Kucukural *et al.*'s study. Therefore, crossover rate is set to be 0.9 and mutation rate is set to be 0.05. The initial chromosome size (size of the feature subset) is chosen to be 20 as suggested by Kucukural *et al.,* which is due to a change after the *first point* based on the methodology discussed earlier. For parameter sets 2 and 3, Fisher's Method was utilized as the classifier during the diversification phase. After the *firs point*, fitness function values are calculated by SVM. Finally for the third parameter set, in order to overcome the local trapment due to long number of iterations, we generated a new population based on the probabilities that are identified with Rank List Gene method. This approach ensured a more diverse search of the solution space.

From the six replications with parameter set 1, we obtained the following gene sets to yield the highest classification accuracies; **(1)** 2, 5 15, 28, 43, 45, 54, 99, 111, 138; **(2)** 11, 18, 49, 80, 89, 111, 136, 231, 883, 1864, 1903; **(3)** 2, 3, 12, 44, 45, 49, 53, 286, 296, 331, 453; **(4)** 1, 3, 10, 18, 32, 38, 49, 54, 89, 221, 296; **(5)** 1, 8, 13, 39, 43, 111, 120, 164, 244, 298, 610; **(6)** 8, 9, 19, 39, 43, 50, 54, 115, 211, 316, 437, 554. The genes with highest frequency and their number of occurrences are depicted in Table 2**.**

Table 2 –The genes with highest frequencies selected from six replications with parameters set 1

Gene	1	3	8	39	43	45	49	54	111
Frequency	2	2	2	2	2	2	2	3	3

The convergence to different feature subsets might be due to several reasons. First of all these different subsets might be different representations of the same pathway,

secondly they might be referring to a different pathway or sample size of the data set is not sufficient enough to eliminate the misleading ones. Kucukural *et. al.* [1] stated that 76% correlation was detected between different elements of feature subsets. The proposed framework eliminates most of the genes that are highly correlated during the preprocessing phase. Results from the six replications also yield different feature subsets each with very high classification accuracy. The correlation of the expression levels of the genes that are member of the feature subsets are illustrated in Figure 2. Note that out of the 6*5=30 possible illustrations randomly chosen 6 of them are presented due to the limited space. Our analysis shows that the resulting feature subsets are not consist of genes that are highly correlated with each other. This finding supports the explanation of Gormley's and Liu's, i.e., the different feature subsets might be indicator of different signaling pathways (if not small sample size).

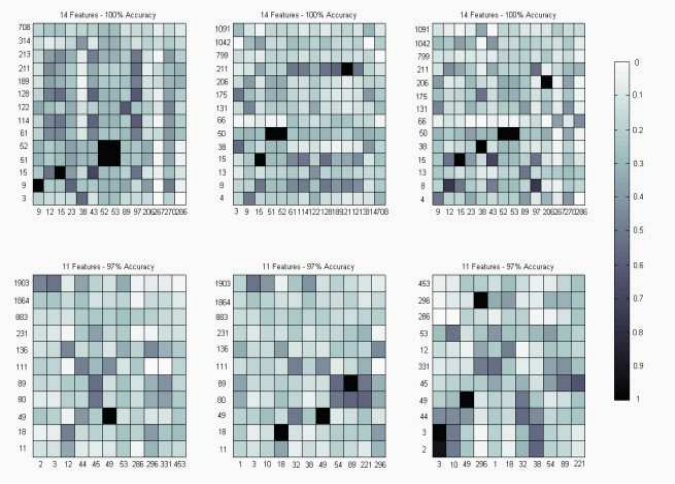

Figure 2. The correlations among the genes of resulting feature subsets.

The results of the experiments and their comparison with earlier research in terms of number of features and classification accuracy is depicted in Table 3. The results suggest that the proposed modifications are valuable and improve the classification accuracy. Particularly, parameter set 2 and 3, which searches the solution space more than parameter set 1, by utilizing a much faster classifier, namely Fisher's method as opposed to SVM, yields the best classification accuracy. However, despite searching the solution space more and yielding better accuracies, parameter set 2 and 3, did not decrease the number of features as much as set 1.

5. CONCLUSIONS

In this paper we proposed a modified framework for feature subset selection problem. The framework suggests the cleaning of the correlated genes from the data set, forming the initial population based on an index that assess the discriminative ability of the features and application of a genetic algorithm that utilizes a fast classifier while calculating the fitness function scores. The results reveal that elimination of most of the highly correlated genes during the preprocessing yield different feature subsets with less correlation among their members. This suggests that convergence to different feature subsets might be an indicator of different pathways. The results also demonstrate that the proposed modifications yields better classification accuracies.

A better experimental analysis to identify which modifications improve the results most, a better algorithm to eliminate the correlated genes and application of the proposed framework to other data sets is left as future research.

Table 3 – The results of the analysis and comparison with earlier research. (PS refers to *Parameter Set*)

	PS 1	PS 2	PS 3	Kucukural *et al.*	Liu *et al.*	Bonilla *et al.*
Accuracy (%)	96.8	100	100	98.4	91	99.4
# of features	11	13	14	12	9	10

REFERENCES

1. Küçükural, A., Yeniterzi, R., Yeniterzi, Sezerman, O. U.: Evolutionary Selection of Minimum Number of Features for Classification of Gene Expression Data Using Genetic Algorithms. In: Proceedings of the 9th annual conference on Genetic and evolutionary computation, pp. 401- -406. ACM, New York (2007)
2. Bø, T.H., Jonassen, I.: New Feature Subset Selection Procedures for Classification of Expression Profiles. Genome Biology. 3(4), 0017.1- -0017 (2002)
3. Holland, J.: Genetic Algorithms Computer programs that "evolve" in ways that resemble natural selection can solve complex problems even their creators do not fully understand, http://wwwlia.deis.unibo.it/Courses/AI/fundamentalsAI 2005-06/lucidi/seminari/roli/holland.GAIntro.pdf
4. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. J. Machine Learning Res. 3, 1157- - 1182 (2003)
5. Xiong, M., Jin, L., Li, W., Boerwinkle, E.:Computational Methods for Gene Expression Based Tumor Classification. BioTechniques. 29, 1264- -1270 (2000)
6. Li, L., Darden, T., A., Weinberg C., R., Levine A., J., Pedersen L.,G.: Gene Assesment and Sample Classification for Gene Expression Data Using a Genetic Algorithm/k-nearest Neighbor Method. Comb. Chem. High Throughput Screen. 4, 727- -739 (2001)
7. Ooi, C., H., and Tan, P.: Genetic Algorithms Applied to Multi-class Prediction for the Analysis of Gene Expression Data. Bioinformatics. 19, 37- -44 (2003)
8. Peng, S., *et. al.*: Molecular Classification of Cancer Types from Microarray Data Using the

Combination of Genetic Algorithms and Support Vector Machines. FEBS Lett. 555, 358- -362 (2003)
9. Vapnik, V., Golowich, S., E., Smola, A.: Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. In: M. Mozer, M. Jor- dan, and T. Petsche (eds.) Neural Inf. Processing Sys., vol. 9. MIT Press, Cambridge, MA (1997)
10. Brown, M., P., *et. al.*: Knowledge Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. Proc. Natl. Acad. Sci. 97, 262- -267 (2000)
11. Furey, T., S., *et. al.* :Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. Bioinformatics. 16, 906- -914 (2000)
12. Liu, J., J., Cutler, G., Li, W., Pan, Z., Peng, S., Hoey, T., Chen, L, Ling, X.: Multiclass Cancer Classification and Biomarker Discovery Using GA-based Algorithms. Bioinformatics. 21, 2691- -2697 (2005)
13. Gormley, M., Dampier, W., Ertel, A., Karacali, B., Tozeren, A.: Prediction Potential of Candidate Biomarker Sets Identified and Validated on Gene Expression Data from Multiple Datasets. 8:415 (2007)
14. Alon, U., Barkai, N., Notterman, D., A., Gish, K., Ybarra, S., Mack, D., Levine, A.,J.: 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)
15. Wang, J., Bo, T., H., Jonassen, I., Myklebost, O., Hovig, E.:Tumor Classification and Marker Gene Prediction by Feature Selection and Fuzzy C-means Clustering Using Microarray Data. BMC Bioinformatics. 4:60 (2003)
16. Liu, X., Krisham, A., Mondry, A.: An Entropy-based gene selection method for cancer classification using microarray data. BMC Bioinformatics. 6:76 (2005)
17. Huerta, E., B., Duval, B., Hao, J., K.: A Hybrid GA/SVM Approach for Gene Selection and Classification of Microarray Data. In: Rothlauf, F. (eds.) Evo Workshops 2006. LNCS, vol. 3907,pp. 34- -44. Springer-Verlag Berlin, Heidelberg (2006)
18. Hedenfalk, I., *et. al.*: Gene Expression Profiles in Hereditary Breast Cancer. New England J. Med. 344(8), 539- -548 (2001)
19. Fisher, R., A.: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics. 7, 179- -188 (1936)