Evolutionary Selection of Minimum Number of Features for Classification of Gene Expression Data Using Genetic Algorithms

Alper Küçükural
Sabancı University
Comp. Biology Lab
MDBF Tuzla
34956 Istanbul,Turkey
90-216-4832129
kucukural@su.sabanciuniv.edu

Reyyan Yeniterzi
Sabancı University
Comp. Biology Lab
MDBF Tuzla
34956 Istanbul,Turkey
90-216-4832129
reyyany@su.sabanciuniv.edu

Süveyda Yeniterzi
Sabancı University
Comp. Biology Lab
MDBF Tuzla
34956 Istanbul,Turkey
90-216-4832129
suveyday@su.sabanciuniv.edu

O. Uğur Sezerman
Sabancı University
Comp. Biology Lab
MDBF Tuzla
34956 Istanbul, Turkey
90-216-4839513
ugur@sabanciuniv.edu

ABSTRACT

Selecting the most relevant factors from genetic profiles that can optimally characterize cellular states is of crucial importance in identifying complex disease genes and biomarkers for disease diagnosis and assessing drug efficiency. In this paper, we present an approach using a genetic algorithm for a feature subset selection problem that can be used in selecting the near optimum set of genes for classification of cancer data. In substantial improvement over existing methods, we classified cancer data with high accuracy with less features.

Categories and Subject Descriptors

I.2 [**Computing Methodologies**]: PATTERN RECOGNITION—Design Metodology—*Feature Evolution and Selection*; J.3 [**Computer Applications**]: LIFE AND MEDICAL SCIENCES—*Biology and genetics*

General Terms

Algorithms, Verification

Keywords

Biomarkers, colon cancer, prostate cancer, ovarian cancer, feature selection, classification, genetic algorithms.

1. INTRODUCTION

Microarray technology allows monitoring expression levels of thousands of genes simultaneously [1-4]. Through comparison of disease and control data sets, it is possible to obtain a significant set of genes that signals the existence of a disease. This set of genes can be used for the development of diagnostic kits that would enable early diagnosis of the disease.

GECCO'07, July 7–11, 2007, London, England, United Kingdom.

The feature subset selection problem refers to the task of selecting a useful set of attributes. For practical purposes it is crucial to determine the minimum set of genes that can classify the disease data with highest accuracy. Since every measurement has its own financial cost and diagnostic value keeping these costs at minimal levels while maintaining accuracy is of significant practical interest.

Pattern Classification problems require selection of a subset of features from a larger data set to represent the patterns to be classified. If the features do not capture the information for classification, the accuracy of the classification method will be limited by the lack of this information regardless of methodology. The abundance of irrelevant attributes would unnecessarily increase the search space while decreasing the accuracy of the classification algorithm. Using a large set of attributes in a classification problem would require a larger number of samples to learn the search space at high accuracy.

2. RELATED WORK

A number of approaches have been proposed in the literature for the feature subset selection problem. Early works used breadth first search and branch and bound algorithms which performed well with conventional statistical classifiers but poorly with non-linear classifiers [5-8]. Others used heuristic search and randomized population based search techniques such as genetic algorithms [9-11].

Some of the most recent works on feature selection methods are using genetic algorithms. In order to optimize classification for feature selection, Punch *et. al.* applied genetic algorithm and K-nearest neighbor algorithm to large biological data sets [12].

In recent years several classification methods are applied to different types of data sets mostly including several disease gene expression data. For instance, Guyon *et. al.* points out another approach that uses Support Vector Machine methods based on recursive feature elimination to classify cancer patients and normal patients [13]. Support vector machines (SVMs) are linear classifiers that use supervised learning methods for classification and regression. They map input vectors to a higher dimensional space and construct a hyper plane that separates the data which at

the same time leads to minimum empirical classification error and maximum geometric margin. Moreover, Jirapech-Umpai and Aitken used evolutionary algorithm, a stochastic search and optimization technique, to find the near optimal set of predictive genes in order to classify a leukemia dataset [14].

Another method was claimed by Fuyer *et. al.* for classification and verification of microarray expression data. They discovered mis-labeled tissue sample in the dataset with SVMs. After they corrected this mislabeled data and removed outliers, the microarray expression data could be classified perfectly with SVM. [15]

In the colon data set that is used in this paper, Alon *et. al.* used clustering algorithms to determine highly correlated expression levels of genes that can be used for diagnosis. They found a set of genes that could be used to cluster the disease data and the control data with 90% accuracy [16].

Fröhlich *et. al.* worked on the same data set and used Genetic algorithms in combination with Support Vector classification to determine the minimum set of genes with the highest classification accuracy. They found a set of 30 genes that can distinguish cancer data from the control set with 85 % accuracy [19]. His graduation thesis has a section with detailed overview of the state of the art feature selection methods.

3. METHODOLOGY

Our algorithm involves a basic genetic algorithm with roulette wheel based selection strategy (Figure 1). We attempted several combinations of parameters and used the combination that yielded optimum results for all data. The following parameter settings are used at each run:

Population Size: 20
Number of generation: 160
Crossover Rate: 0.9
Mutation Rate: 0.05

Each individual in the population represents a candidate solution to the feature subset selection problem. The vector representing the solution is m long, where m represents number of attributes. We used a binary encoding 1s representing that a feature is selected in that parent to be used in classification. There are 2^m possible individuals. In our data sets, m ranges from 2000 to 15154, thus making exhaustive search impossible.

The number of selected features (1's) in each individual is initially fixed to 20 to determine a subset of m features that has the highest classification accuracy. The features are randomly generated for the initial population.

The fitness function uses the classification accuracy of the features in an individual realized by the Support Vector Classification tool in a toolbox called PRTools in Matlab [20] which used 50% of the data as a training set and 50% of the data as a test set. We used other classification methods including linear discriminant analysis (LDA) and neural networks but SVM yielded the highest classification accuracy. Basic cross over and mutation operations are applied to generate offspring from selected parents. After each generation, the worst scoring n offspring were replaced by best scoring n parents if the parent's fitness scores are better than the scores of children's. N is chosen as 10% of the population size.

We implemented a dynamic parent generation procedure that enables selection of smaller number of good features while creating a new population. During the search, each feature present in an individual was assigned the fitness value of the parent they occupied. After initial 30 generations, the average fitness score for each feature is obtained by dividing the total fitness score of the feature by the number of times that feature was chosen in an individual. This value is high if the feature was mostly present in individuals with high prediction accuracy. A new population of individuals is created according to this selection scheme. Basic GA is run for 10 generations; after every 10 generations, a new population of individuals is generated from the updates fitness scores of the genes. The algorithm of the method is given in Figure 2.

Each feature is assigned a probability of selection in the generation of new individuals depending on its fitness score from the previous 10 runs. Individuals are generated via a roulette wheel selection scheme. Each feature is assigned a numerical region depending on its relative fitness score amongst m features. While generating an individual 20 random numbers are selected and corresponding features are chosen to be included in the individual. If a feature has a high fitness score, it may be selected several times thus decreasing the number of 1s in the individual. This novel parent generation scheme enabled us to achieve a dynamic selection of near optimal number of features (genes).

After the generation of a new population, individuals are randomly selected to become parents according to their fitness scores and generate offspring. The cross over operator is devised to keep the number of features (1's) in an individual constant. The number of features that will be involved in cross over is equals to a random proportion of the number of 1s that the parent with fewer features possesses (rounded to the nearest integer). Features participating in the cross over are also randomly selected and exchanged between the partners.

1. First generation G_1 of n parents $\{P_1, ..., P_n\}$ is created using binary encoding. Each parent P_i is randomly generated and represents a feature subset of m features.
2. Fitness function of each P_i, $F(P_i)$ is calculated using the Support Vector Classification.
3. Until the termination condition met
   - **a-** Randomly pair all the parents $\{P_1, ..., P_n\}$ with each other.
   - **b-** Basic cross over and mutation operations are applied to generate a new generation $G_{(i+1)}$ of n offspring $\{P_1', ..., P_n'\}$
   - **c-** Fitness score of each P_i', $F(P_i')$ is calculated using the Support Vector Classification.
   - **d-** Roulette Wheel Selection is used and worst scoring n offspring were replaced by best scoring n parents if the parent's fitness scores $F(P_i)$ are better than the scores of children's $F(P_i')$.

Figure 1. Basic Genetic Algorithm

Other feature selection methods find a fixed number selected features and usually decrease this number by leave one out strategy to determine the minimum set of features. The strength of our procedure relies on our dynamic parent generation scheme

based on a fitness score of each feature (gene). This approach resembles the selfish gene idea where individuals are mere transporters of genes and the parent's function is to carry on the strong genes to next generations.

The ultimate goal is the survival of the gene. In our dynamic selection approach strong genes are selected more by individuals and passed on to succeeding generations. In dynamic individual generation, the number of genes in an individual can be decreased if the individual has a good fitness value with fewer genes. This procedure mimics the efficiency of nature as well. In nature, as the species evolve, similar functions can be achieved with a fewer number of genes.

1. The first generation G_1 of n parents $\{P_1, ..., P_n\}$ is created using binary encoding. Each parent P_i is randomly generated and represents a feature subset of m features.
2. Fitness function of each P_i, $F(P_i)$ is calculated using the Support Vector Classification.
3. Until the termination condition is met
   - **a-** Randomly pair all the parents $\{P_1, ..., P_n\}$ with each other.
   - **b-** Standard cross over and mutation operations are applied to generate a new generation $G_{(i+1)}$ of n offspring $\{P_1', ..., P_n'\}$
     OR

     At each n generation, G(k) where k=0 (mod n), new offspring are generated by using population.
     - The average fitness score for each feature f_i is obtained by dividing the total fitness score of the feature by the number of times that feature was chosen in a parent.
     - Each feature f_i is assigned a probability to be selected in generation of new individuals depending on its fitness score from the previous n runs.
     - Roulette wheel selection is used to generate new parents.
   - **c-** The fitness score of each P_i', $F(P_i')$ is calculated using the Support Vector Classification.
   - **d-** Each feature that is present in a parent P_i is assigned the fitness value of that parent, $F(P_i)$.
   - **e-** Roulette Wheel Selection is used and worst scoring n offspring were replaced by best scoring n parents if the parent's fitness scores $F(P_i)$ are better than the scores of children using $F(P_i')$.

Figure 2. Population Genetic Algorithm

Several other researchers worked on the same data sets to determine the set of genes that can be used to differentiate the cancer patients from the control group. So far the best results were obtained by Bing et al. for prostate and colon cancer data sets in 2004 [21]. They used a combination of feature selection methods such as ranksum test, Principle Component Analysis and t test along with clustering algorithms in their work. These researchers chose top ranked 30 genes that are differentially expressed and verified their classification accuracy using 3 fold, 10 fold, and leave-one-out cross validation. The best result for ovarian dataset was obtained from Liu *et. al*. in 2002 [22]. The correlation based feature selection method was used in their work. They also ranked subsets of features rather than ranking individual features and determined the most discriminative features. They employed several classification methods with 10-fold cross validation to confirm their results using kNN, SVM, Naïve Bayes, and others [22].

4. RESULTS

4.1 Data Set

The experiments given here are real world data sets obtained from three different sources. They all show expression levels of certain number of genes in cancer and control patients. The expression data were obtained from micro array studies.

In the first data set we used a colon cancer data obtained by Alon *et. al*. using Affymetrics oligonucleotide arrays. Data showed gene expression levels of 2000 genes for 40 tumor and 22 normal colon tissue samples [16].

Second data set consisted of gene expression levels of 15154 genes in 162 ovarian cancer and 91 control patients are obtained from Petricoin *et. al.* [17].

The third data set consisted of gene expression levels of 12,600 genes taken from 52 prostate cancer and 50 control samples are obtained from Singh *et. al.* [18].

4.2 Experiments

The features obtained from GA that yield to the highest classification accuracy are selected for each data set. These features are given to BSVM[1] package for 10 fold cross validation of each data set with selected features. We could not run n fold cross validation in the SVM classification step due to time considerations during the GA; for this reason the accuracy values of our algorithm and the BSVM differ.

Table 1. Classification Accuracy using ten fold cross validation and BSVM tool

	Colon		Ovarian		Prostate	
	Acc.	**#Feat.**	**Acc.**	**#Feat.**	**Acc.**	**#Feat.**
All Features	96.77	2000	100	15154	88.83	12600
Our Features	98.38	12	100	12	96.07	19
Others Results	91.94*	30*	100**	17**	97.06*	30*

* Results from Bing *et. al*[21], ** Results from Liu *et.al.*[22]

The results of the experiments are summarized in Table1. The colon data can be classified with 98.38 % accuracy using 12 features only. In the literature highest classification accuracy for this data set was 91.94% using 30 features. Our accuracy was higher than the accuracy (96.77%) we would obtain if we had used all the features in ten fold cross validation.

We compared performance of our algorithm with a basic GA in Colon dataset, without a dynamic population generation scheme, run for 160 steps while keeping all the other parameters (20 parents and 20 features) the same (see Figure 3). Our algorithm

[1] The current implementation of BSVM can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/bsvm/

(see Figure 4) finds a better solution and converges faster than the basic GA (93.6% accuracy). As seen in Figure 4, a number of features used in classification steadily decreases to around 12 features.

The ovarian data can be classified with 100% ten fold cross validation accuracy using the 12 features determined by our algorithm.

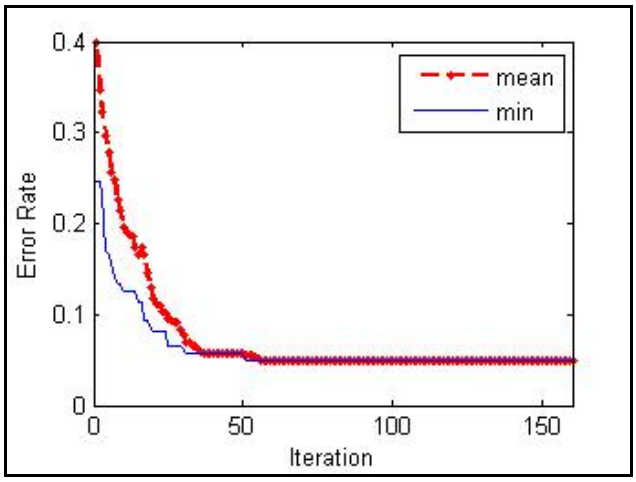

Figure 3. Average error rates of the population and the best individual scores of 10 Basic GA runs for Colon Data

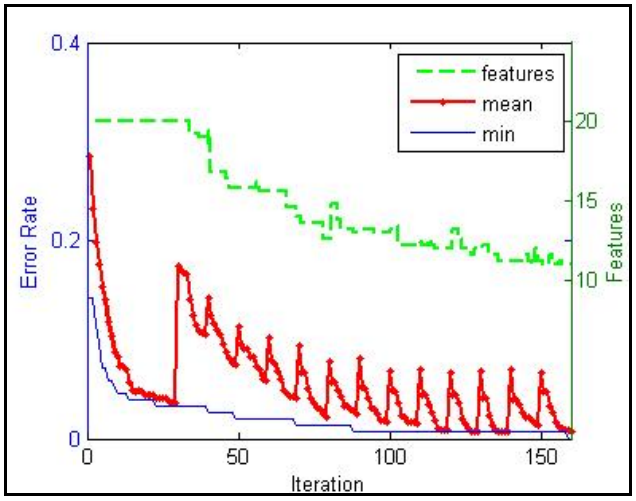

Figure 4. Average error rates of 10 runs of our algorithm for Colon Data. Green lines indicate the average number of minimum features in each iteration.

In the literature, the highest classification accuracy for the ovarian data set was also 100% by Liu *et. al.*[22]. However, they could only bring down the number of features to be used to 17 without lowering the classification accuracy.

The accuracies for both approaches are the same; however, we accurately classified the data using a smaller number of features. One hundred percent classification accuracy was achieved when we used all the features in ten fold cross validation. (Table 1.)

We compared our results for ovarian data with a basic GA to see the improvement over basic GA. The results are summarized in Figure 5 and Figure 6. The basic GA obtained 99% accuracy comparable to our results but our algorithm converged faster and used only 12 features to achieve the same accuracy while the basic GA used 20 features. The number of features gradually decreased to 12 gradually as seen in Figure 6.

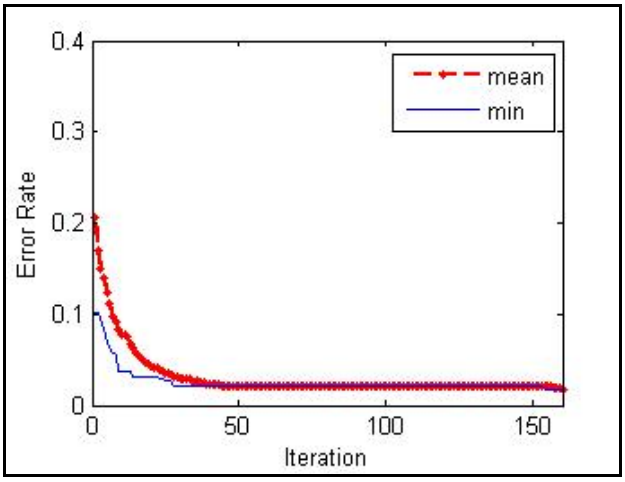

Figure 5. Average error rates of the population and the best individual scores of 10 Basic GA runs for Ovarian Data.

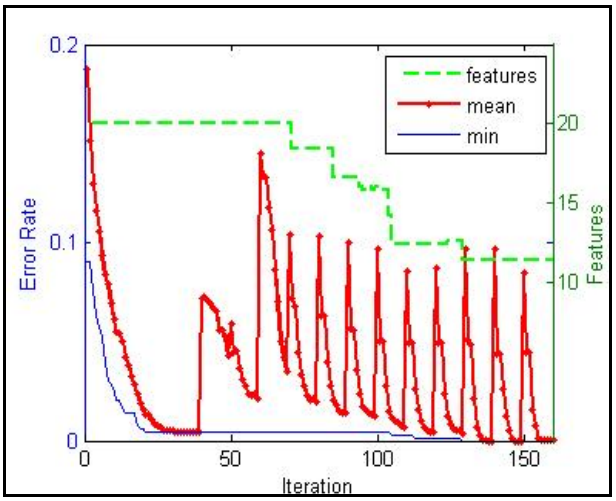

Figure 6. Average error rates of 10 runs of our algorithm for Ovarian Data. Green lines indicate the average number of minimum features in each iteration.

Prostate cancer data can be classified with a 96.07% ten fold cross validation accuracy using the 19 features that were determined by our algorithm. In the literature, highest classification accuracy for this data set was also 97.06% by Bing *et. al.*[22]. Although they could decrease the number of features to be used, as many as 30 features were employed without lowering the classification accuracy. The accuracies for both approaches are comparable; however, we classified the data using smaller number of features.

88.83% classification accuracy is achieved when we used all the features in ten fold cross validation. (Table 1.)

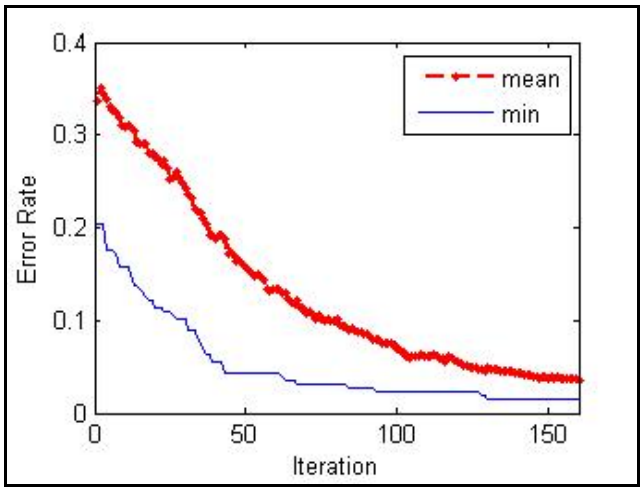

Figure 7. Average error rates of the population and the best individual scores of 10 Basic GA runs for Prostate Cancer Data.

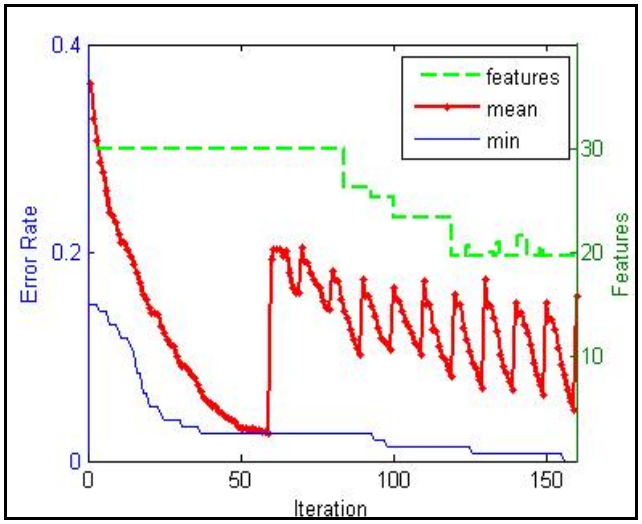

Figure 8. Average error rates of 10 runs of our algorithm for Prostate Cancer Data. Green lines indicate the average number of minimum features in each iteration.

In the prostate cancer data set, there are large number of features (12600) and a relatively less amount of cancer and control data. Therefore, our algorithm could not converge and efficiently search the space with existing parameters. To this end, we searched the parameter space for several combinations of parameters and discovered that starting the search with 30 parents and 30 features yielded the best prediction accuracy. Figure 5 summarizes the basic GA algorithm run with 30 parents and 30 features for prostate cancer. The best classification accuracy was 98.4%. Our algorithm converges faster to a better solution (100%). (Figure 8.)

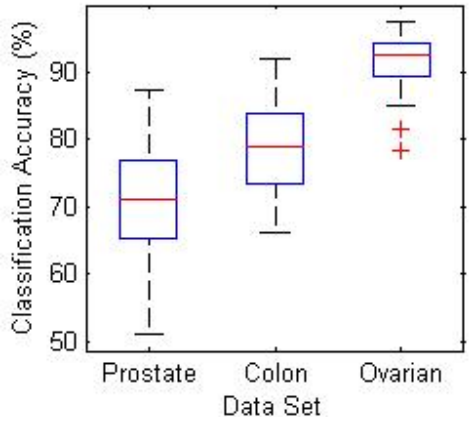

Figure 9. Classification accuracies of 40 experiments of randomly chosen features from each dataset

The classification power of the features selected by our method could not be achieved randomly. To prove this claim, we randomly selected 12 features from the ovarian and colon dataset and 19 features from the prostate dataset and classified the data using BSVM ten fold cross validation. This experiment was repeated 40 times. The results are summarized in Figure 9. The results were much worse than our results for prostate and colon data sets. Randomly selected genes could achieve on average 91.6% classification accuracy, an amount a little worse than our results in ovarian data set.

The consistency of the selected features by our algorithm was checked by running the algorithm ten times on the same training and the test colon data set. Each run ended up with different subset of features, however, the correlation between the features coming from different subsets were higher than 76%.

5. CONCLUSION AND DISCUSSION

Feature subset selection problem is of significant practical interest especially in determining number features to be used for diagnostic purposes from real disease data. As the number of attributes to be tested increase, the cost of diagnostics increases proportionally. Also determining the most relevant genes involved in a disease pathway may lead to development of novel therapeutic measures.

In this study, we have tried to determine the minimum set of genes that can be used to differentiate the cancer patients from the control group. The selection of the subset of features is a combinatorial optimization problem which can not be solved in polynomial time; furthermore the complexity of the problem increases exponentially with the total number of features.

Several researchers have used a combination of GA with several different classification algorithms as we did in our algorithm with SVM but in most of these cases, they fix the number of features to be used and solve for a fixed number of features. Then they use a heuristics to decrease the number of features selected while keeping the classification accuracy. GA was only used to search the space for different combinations of a fixed number of features. In our approach while initially searching the space for optimum combination of features, we also determine which features are

involved in good solutions. Using this information we could generate fitter parents and converge faster than do existing approaches. During the parent generation step, each gene has different survival probability proportional to their fitness value. We choose a fixed number of features and if the same gene is chosen several times, this selection would automatically decrease the number of genes in that parent. If selected features decrease the fitness score of that individual it would be eliminated by the survival of fittest approach of GA. If selected features improves the score this parent will be selected and passed onto the next generations.

Dynamic parent generation step is inspired by nature. The idea of a fitter and fewer genes (features) make-up for fitter and more evolved efficient parents enabled us to dynamically reduce number of genes. In this way we could obtain a smaller number of features with the highest classification accuracy for each data set. If selection of features decreases the fitness score of that parent, this parent would be eliminated by the survival of the fittest approach of GA.

The selected set of genes for the colon cancer data includes oncogenes, cell adhesion molecules and collagens which were shown to be involved in colon cancer by experimental studies. Similar set of features were also found with Guyon *et. al.* as well however these genes were amongst 16 genes that were selected by their method with the same accuracy. Our method selected 11 genes for this data set and 5 of them were known colon cancer related genes.

6. REFERENCES

[1] David W, Galbraith, Global analysis of cell type-specific gene expression. *Comp Funct Genom* 2003, 4:208-215.

[2] Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE, Davis RW, Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci*. USA 1997, 94:2150-2155.

[3] Eisen MB, Spellman PT, Brown PO, Botstein D, Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci* USA 1998, 95:14863-14868.

[4] Van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH, Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002, 415:530-536.

[5] B. Schölkopf, A. J. Smola, Learning with Kernels, *MIT Press*, Cambridge, MA, 2002

[6] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern classification, *2nd ed.John Wiley and Sons*, New York, 2001.

[7] A. Webb, Statistical Pattern Recognition, *Wiley*, New York, 2002.

[8] V. Vapnik, Statistical Learning Theory*, John Wiley and Sons*, New York, 1998

[9] M. Raymer, W. Punch, E. Goodman, L.Kuhn, A. Jain, Dimensionality Reduction UsingGenetic Algorithms*, IEEE Transactions on Evolutionary computing*, 2000

[10] F. J. Ferri, V. Kadirkamanathan, J. Kittler, Feature Subset Search using Genetic Algorithms, *IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, Essex, 1993

[11] M. Richeldi, P. Lanzi, A Tool for Performing efective feature selection by investigating the deep structure of the data, *Proceedings of the International Conference on Tools with Artifcial Intelligence*, pp. 102 - 105, 1996

[12] Punch W F, Goodman E D, Pei M, Chia-Shun L, Hovland P, Enbody R. "Further Research on Feature Selection and Classification Using Genetic Algorithms", *Proceedings of the Fifth International Conference on Genetic Algorithms*, ICGA, 557-564, 1993.

[13] Guyon I, Weston J, Barnhill S, Vapnik V. "Gene Selection for Cancer Classification using Support Vector Machines", *Machine Learning*, vol.46, issue 1-3, pp. 389-422, 2004.

[14] Jirapech-Umpai T, Aitken S. "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes", *BMC Bioinformatics*, vol.6, 2005.

[15] Terrence S. Furey, Nigel Duffy, Nello Cristianini, David Bednarski, Michel Schummer, and David Haussler, Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data, *Bioinformatics*. 2000, 16(10):906-914

[16] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, A. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues probed by oligonucleotide arrays, *Cell Biology*, 96:6745-6750, 1999

[17] Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA: Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002, 359:572-577.

[18] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR: Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 2002, 1:203-209.

[19] Holger Fröhlich, Feature Selection for Support Vector Machines by Means of Genetic Algorithms, *Diploma Thesis in Computer Science*, University Marburg, 2002

[20] Ferdi van der Heijden, Robert P.W. Duin, Dick de Ridder and David M.J. Tax,:Classification, parameter estimation and state estimation - an engineering approach using Matlab. *John Wiley & Sons*, ISBN 0470090138 (2004)

[21] Liu, Bing; Cui, Qinghua; Jiang, Tianzi; Ma, Songde (2004) " A combinational feature selection and ensemble neural network method for classification of gene expression data" *BMC Bioinformatics* 5 136

[22] Liu H, Li J, Wong L: A comparative study on feature selection and classification methods using gene expression profiles and proteomic Patterns. *Genome Inform Ser Workshop Geonome Inform* 2002, 13:51-6