World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:5, 2007

A Hybrid Approach for Selection of Relevant Features for Microarray Datasets

R. K. Agrawal, and Rajni Bala

***Abstract*—**Developing an accurate classifier for high dimensional microarray datasets is a challenging task due to availability of small sample size. Therefore, it is important to determine a set of relevant genes that classify the data well. Traditionally, gene selection method often selects the top ranked genes according to their discriminatory power. Often these genes are correlated with each other resulting in redundancy. In this paper, we have proposed a hybrid method using feature ranking and wrapper method (Genetic Algorithm with multi-class SVM) to identify a set of relevant genes that classify the data more accurately. A new fitness function for genetic algorithm is defined that focuses on selecting the smallest set of genes that provides maximum accuracy. Experiments have been carried on four well-known datasets[1]. The proposed method provides better results in comparison to the results found in the literature in terms of both classification accuracy and number of genes selected.

***Keywords*—**Gene Selection, Genetic Algorithm, Microarray datasets, Multi-class SVM.

I. Introduction

DNA microarray offers the ability to measure levels of expressions of thousands of genes simultaneously. The hypothesis that many or all human diseases may be accompanied by specific changes in gene expression has generated much interest among the Bioinformatics community in classification of patient samples based on gene expression for disease diagnosis and treatment. Especially the classification of cancers from gene expression profiles is active research area in bioinformatics.

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 over fitting. Moreover high dimensional data requires inevitably large processing time. So for analyzing microarray data, it is necessary to reduce the data dimensionality by selecting a subset of genes (features) that are relevant for classification.

Feature Selection is often used as preprocessing technique in machine learning and data mining. It is often effective in reducing dimensionality, improving mining accuracy and enhancing accuracy of the classifier. There are two major approaches to gene selection: filter and wrapper approach [4, 5]. Most filter methods have adopted statistical feature selection, which needs less computation than the others do. It independently measures the importance of features to select good features. Since, the filter approach does not take into account the learning bias introduced by the final learning algorithm, it may not be able to select the most suitable set of features for the learning algorithm. The disadvantage of filter approach is that the features could be correlated among themselves [6], [7]. On the other hand, wrapper methods tend to find features better suited to the predetermined learning algorithm resulting in better performance. But, it also tends to be more computationally expensive since the classifier must be trained for each candidate subset. In literature, several strategies were considered to explore the space of possible subsets. Some of them are evolutionary algorithms used with a k-nearest neighbor classifier [8], parallel genetic algorithms using adaptive operators [9] and SVM Wrapper with standard GA [10]. The conventional wrapper methods using genetic algorithm have been applied to feature selection of small or middle scale feature datasets [4], [11]. But, it is hard to apply them directly to high dimensional datasets due to much processing time [12]. Reducing the search space for genetic algorithm will decrease the computation time. This can be achieved by selecting a reduced set of important genes from high dimensional genes without losing any informative gene.

Since last decade active research have been carried out in binary cancer classification with feature selection [13]-[15] however, only a small amount of work has been made on feature selection on multi-class datasets [16]-[19]. This paper focuses on selecting an optimal set of genes for multi-class datasets.

In this paper, a hybrid method that uses advantage of both filter and wrapper approach for gene selection is proposed. The filter method is used to select the top-ranked genes, say M [3]. The number of genes selected, M, is set by human intuition with trial-and-error. There are also studies on setting M based on certain assumption on data distributions [20]. These M genes may be correlated among themselves, which may lead to redundancy in the feature set. Also certain genes may be noisy which may decrease classification accuracy. So this set of selected genes is further reduced with the help of genetic algorithm combined with multi-class SVM. A new fitness function for GA is proposed which always selects the smallest set of genes that provides maximum accuracy. The proposed method is experimentally assessed on four well known datasets (Leukemia, Lymphoma, SRBCT and GCM). Comparisons with other state of art methods show competitive results.

This paper is organized as follows: Section II describes the theory of multi-class Support Vector Machine. In Section III Genetic algorithm is discussed briefly. Proposed method for gene selection for cancerous dataset is given in Section IV. Experimental results are shown in Section V and conclusions are drawn in Section VI.

[1] Leukemia, SRBCT, Lymphoma, GCM.

II. Multi-Class Support Vector Machine

Support Vector Machine (SVM) is based on statistical learning theory developed by Vapnik [29, 30]. It has been used extensively for classification of data. Given n training samples $\{(x_i, y_i), \forall i = 1,\ldots\ldots., n\}$, where x_i is the input feature vector for the i^{th} sample and y_i is the corresponding target class output value, the problem is to determine the optimal values of the weight vector w and bias b such that they satisfy the constraint

$$\mathbf{d}_i (w^T x_i + b) \geq 1 - \xi_i \quad \text{for } i = 1,2,3,\ldots, n$$

$$\xi_i \geq 0 \text{ for all } i$$

and such that the weight vector w and the slack variables ξ_i minimize the cost functional

$$\Phi(w, \xi) = \frac{1}{2} w^t w + C \sum_{i=1}^{n} \xi_i \quad (1)$$

where $C > 0$ is a user-specified regularization parameter.

The dual problem for the above can be formulated as: Given the training sample $\{(x_i, d_i)\}_{i=1}^{n}$, find the Lagrange multiplier $\{\alpha_i\}_{i=1}^{n}$ that maximize the objective function

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i=1}^{n} \alpha_i \alpha_j d_i d_j x_i^{\ t} x_j \quad (2)$$

subject to the constraints

(i) $\sum_{i=1}^{N} \alpha_i d_i = 0$

(ii) $0 \leq \alpha_i \leq C$ *for i = 1, 2, ….., n*

Having determined the Lagrange multipliers, denoted by $\alpha_{0,i}$, the optimal solution for the weight vector w is given by

$$w_0 = \sum_{i=1}^{n} \alpha_{0,i} d_i x_i \quad (3)$$

and the optimal bias b_0 can be obtained using the following equation

$$b_0 = 1 - w_o^{\ t} x \quad (4)$$

If in the current input space the patterns are not linearly separable then SVM can perform a nonlinear transformation via the inner-product kernel $K(x_i, x_j)$ to map the input space to a new high-order feature space where the patterns are more likely to be linearly separable. The use of such kernel function can lead to a decision function that is non-linear in the input space but its image is linearly separable in the high dimensional feature space. With this expansion the new dual form of the constrained optimization of a SVM can be stated as

Maximize

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i=1}^{n} \alpha_i \alpha_j d_i d_j K(x_i, x_j)$$

Subject to the constraints

(i) $\sum_{i=1}^{N} \alpha_i d_i = 0$

(ii) $0 \leq \alpha_i \leq C$ *for i = 1, 2, ….., n* (5)

SVM was originally designed for binary classification. How to effectively extend it for multi-class classification is still an ongoing research issue [31]. The most common way to build a k-class SVM is by constructing and combining several binary classifiers [32]. The representative ensemble schemes are One-Against-All and One-Versus-One. In One-Against-All k binary classifiers are trained, each of which separates one class from other $k-1$ classes. Given a test sample X to classifier, the binary classifier with the largest output determines the class label of X. One-Versus-One constructs $\frac{k*(k-1)}{2}$ binary classifiers. The outputs of the classifier are aggregated to make a final decision. Decision tree formulation is a variant of One-Against-All formulation based on decision tree. Error correcting output code is a general representation of One-Against-All or One-Versus-One formulation, which uses error-correcting codes for encoding outputs [25]. The One-Against-All approach, in combination with SVM, provides better classification accuracy in comparison to others [31]. Consequently we applied One-Against-All approach in our experiments.

III. Genetic Algorithms

A genetic algorithm (or GA) is a search technique used for computing true or approximate solution to optimization and search problems [33]. Genetic algorithms are categorized as global search heuristics. Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover.

Genetic algorithms are implemented as a computer simulation in which a population of abstract representations (called chromosomes or the genotype) of candidate solutions (called individuals) to an optimization problem evolves toward better solutions. Traditionally, solutions are represented as a finite sequence of 0's and 1's, but other encodings are also

possible. In general, the evolution starts from a population of randomly generated individuals and occurs in generations. In each generation, the fitness of every individual in the population is evaluated. Based on their fitness multiple individuals are stochastically selected from the current population, and modified (recombined and possibly randomly mutated) to form a new population. Next iteration is carried out with the newly obtained population. Generally, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached.

A typical genetic algorithm requires two things to be defined:

(i) a genetic representation of the solution domain
(ii) a fitness function to evaluate the solution domain.

In proposed method each chromosome represents a set of genes selected for classification. It is represented by a sequence of M 0's and 1's. In the chromosome 1 means that the corresponding gene is selected and 0 indicates the corresponding gene is not selected for the classification. The initial population is generated randomly. A new fitness function is proposed that selects the smallest subset of genes which provides maximum accuracy.

IV. Proposed Method for Gene Selection for Cancer Dataset

The proposed method is given in Fig. 1. It involves two phases. In the first phase, genes are ranked using any one of the ranking methods. These measures are determined using Rankgene [21] software developed at the computational Genomics Laboratory, Boston University. Experiments are performed to observe the variation of classification accuracy with the number of genes based on ranking for above measures. Top most M ranked genes are selected for the second phase. In the second phase, genetic algorithm in conjunction with multi-class SVM is used to select the smallest subset of genes from the above selected M genes that gives maximum accuracy.

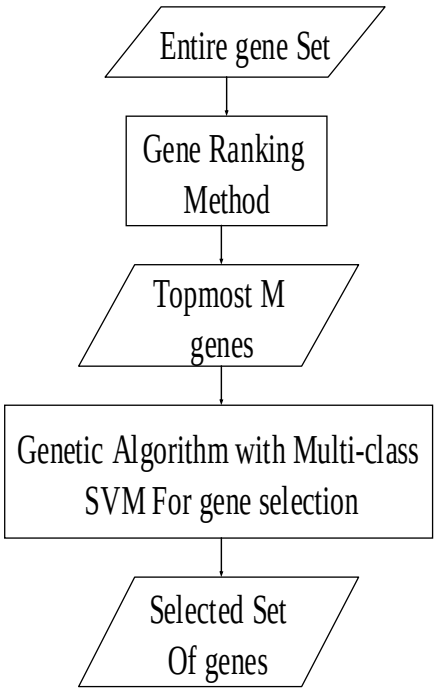

Fig. 1 Proposed Method

To achieve the above said objective, the fitness function for the genetic algorithm should focus on selecting a subset of genes that not only maximizes accuracy of a classifier but also minimizes the number of genes. This is multi-objective optimization problem. Here, one objective is to maximize the classification accuracy and second is to minimize the number of genes. This is achieved by defining the fitness function of chromosome x as

$$Fitness\ (x) = A(x) + {}^{P}\!/_{N(x)}$$

where for chromosome x,

$A(x)$ is the classification accuracy of multi-class classifier defined as ratio of number of correctly classified samples to total number of test samples.

$P = 100 / (M$ **Number of test sample used in classifier*)

$N(x)$ *is the size of gene set (number of 1's in chromosome x) used for classification*

The value of P chosen in fitness function will take care that number of genes are not minimized at the cost of accuracy.

V. Experimental Results

A. Datasets Used

Experiments are carried on four well-known gene expression data sets, which are the leukemia dataset [22], the small round blue cell tumors (SRBCT) [23], the lymphoma

dataset [24] and the GCM dataset [34]. Normalization is carried out so that every observed gene expression has mean equal to 0 and variance equal to 1. Table I gives the brief description of the datasets used in experiments. In the experiments the original partition of the datasets into training and test sets is used whenever information about the data split is available. In the absence of separate test set, 10 fold cross validation is used for calculating the classification accuracy.

TABLE I
DATASETS USED

Dataset	No. of genes	Classes	Train Sample	Test Sample
Leukemia	7129	3	38	34
SRBCT	2308	4	63	20
Lymphoma	4026	3	62 [a]	
GCM	16063	14	144	54

[a] 10 Fold cross validation used for calculating accuracy as test and training samples are not available separately.

B. Experimental Setup and Results

Attributes are ranked using Rankgene software. Ranking method used are Information Gain, Towing Rule, Gini Index and Sum of Variance. Multi-class SVM classifier (one against all) is implemented using MATLAB. The kernel chosen is RBF kernel ($K(\vec{x},\vec{y}) = \exp(-\gamma \| \vec{x} - \vec{y} \|^2)$ in the multi-class SVM classifier. The control parameter γ is taken as 0.01 and the regularization parameter C is fixed as 100. The variation of classification accuracy with the different number of genes is shown in Fig. 2a-2d. It is observed that approximately 100% accuracy is achieved on test dataset with first 50 genes in all the ranking methods for Leukemia, SRBCT and Lymphoma dataset. So for the above three datasets M is chosen as 50. However for GCM dataset accuracy of 100% on testing dataset is not achieved with first 50 genes. It is observed that there is no significant increase in classification accuracy with addition of gene beyond 100 genes. In fact, the behavior is similar for all the ranking methods used in the experiment. So M is chosen as 100 for GCM dataset.

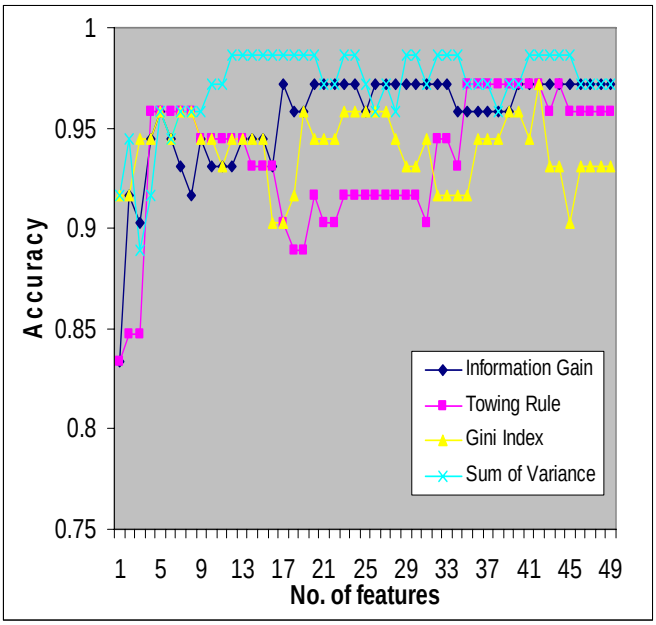

Fig. 2a Variation of classification accuracy with number of features (genes) for Leukemia dataset

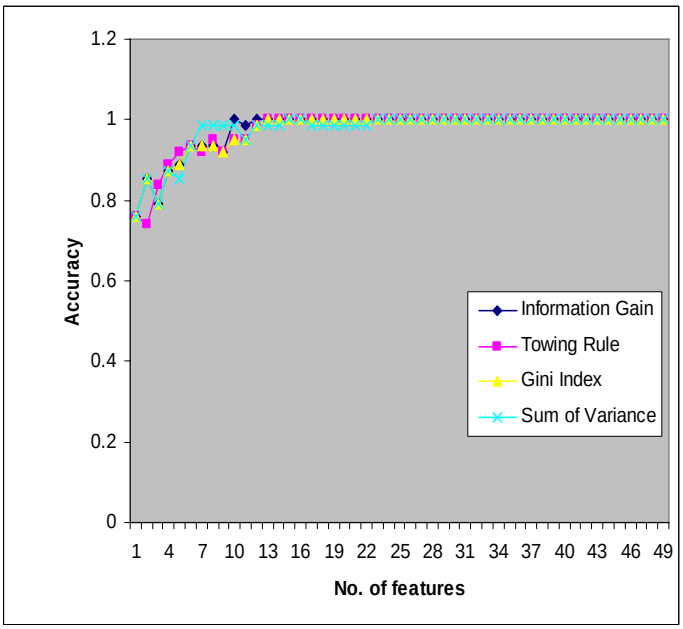

Fig. 2b Variation of classification accuracy with number of features (genes) for lymphoma dataset

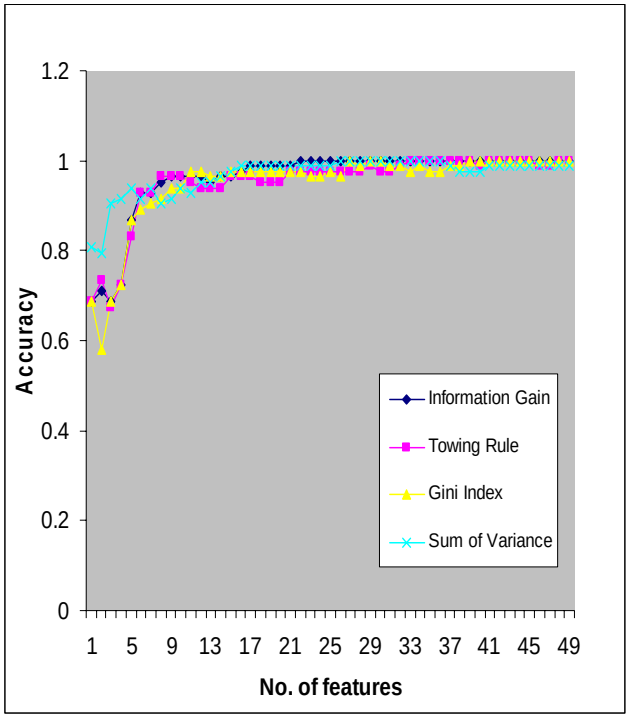

Fig. 2c Variation of classification accuracy with number of features (genes) for SRBCT dataset

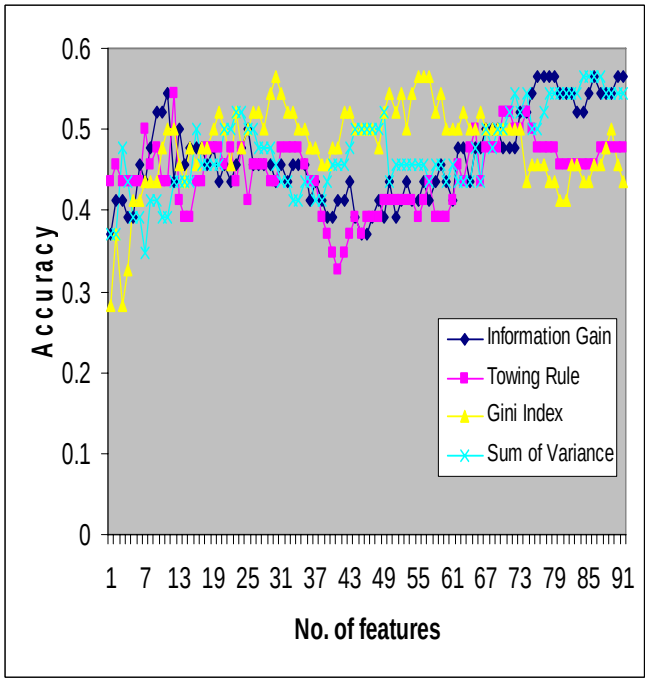

Fig. 2d Variation of classification accuracy with number of features (genes) for GCM dataset

GA with Multi-class SVM classifier (one against all) is implemented using MATLAB. The kernel function, the control parameter γ and the regularization parameter C in the multi-class SVM is same as used in phase I. GA parameters used for gene subset selection for all the datasets are given in Table II.

TABLE II
GA PARAMETERS USED FOR GENE SELECTION

Parameters	**Value**
Size of population	50
Length of Chromosome	100 for GCM 50 for rest
Number of generations	200 for GCM 500 for rest
Crossover rate	0.98
Mutation rate	0.02

Genetic algorithm in conjunction with SVM is run 10 times for all the ranking method on each dataset of Leukemia, SRBCT, Lymphoma and GCM. The results for each of the ranking method, in term of minimum number of genes selected, for all the datasets from 10 independent runs are summarized in Table III. It is observed that a classification accuracy of 100% is achieved with 2(two) genes for Leukemia, with 3(three) genes for SRBCT and with 3(three) genes for Lymphoma. However in case of GCM maximum accuracy of 78.26% is observed with 38 genes. The smallest genes set giving maximum accuracy for each of the dataset are listed in the Table IV.

TABLE III
THE MINIMUM NUMBER OF GENES SELECTED BY PROPOSED METHOD

Methods used for feature ranking	Leukemia		SRBCT		Lymphoma		GCM	
	#Genes	Accuracy (%)	#Genes	Accuracy (%)	#Genes	Accuracy (%)	#Genes	Accuracy (%)
Information Gain	2	100	3	100	3	100	38	76.08
Towing Rule	2	100	3	100	4	100	34	73.91
Gini Index	2	100	3	100	3	100	**38**	**78.26**
Sum of Variance	2	100	3	100	3	100%	31	71.73

TABLE IV
THE SMALLEST GENES SUBSETS THAT PRODUCE THE MAXIMUM CLASSIFICATION ACCURACY

Dataset	Smallest gene subsets with gene#
Leukemia	(6855_attr, 5543_attr) (758_attr, 4050_attr) (1834_attr, 5171_attr) (1834_attr, 2642_attr)
SRBCT	(417_att,509_attr,1613_attr) (255_attrr,1003_attr,1613_attr) (255_attr,554_attr,1613_attr) (255_attr,1613_attr,2046_attr) (187_attr,1601_attr,1613_attr) (255_attr,509_attr,1613_atrr) (187_attr,842_attr,1613_attr)
Lymphoma	(678_attr,3763_attr,3805_attr) (678_attr,758_attr,788_attr) (758_attr,3734_attr,3760_attr) (2683_attr,2736_attr,3760_attr) (734_attr,2841_attr,3763_attr) (768_attr,3735_attr,3763_attr)
GCM	(331_attr,665_attr,676_attr, 928_attr,1403_attr,1429_attr, 1685_attr,2156_attr,2161_attr, 2170_attr,2413_attr,3054_attr, 3087_attr,3361_attr,3537_attr, 3365_attr,3535_attr,3772_attr, 3919_attr,4142_attr,4351_attr, 4390_attr,4479_attr,4610_attr, 4882_attr,4984_attr,5119_attr, 5308_attr,5463_attr,5647_attr, 5772_attr,5998_attr,6027_attr, 6141_attr,6425_attr,6702_attr, 6723_attr,13869_attr)

Comparison of our results in terms of classification accuracy and number of genes with other four state of art method are available in Table V. It is observed from Table V that for leukemia and SRBCT dataset the proposed method gives better results in terms of both classification accuracy and the number of selected genes. The proposed method gives 100% classification accuracy with only 3(three) genes on lymphoma data. Due to paucity of results for lymphoma data as 3-class problem comparison with others is not shown for lymphoma. Even though an accuracy of 100% is achieved with all the ranking methods in case of Leukemia, SRBCT and Lymphoma, genes set selected are not same. This is so as GA is stochastic method. For GCM dataset an accuracy of 78.26 % is achieved with 38 genes.

TABLE V
COMPARISON OF CLASSIFICATION ACCURACY AND NUMBER OF GENES BETWEEN DIFFERENT METHODS

	Leukemia		SRBCT		GCM	
	Accuracy (%)	#genes	Accuracy (%)	#genes	Accuracy (%)	#genes
Our proposed method	**100**	**2**	**100**	**3**	**78.26**	**38**
Fu & Liu[16]	97.04	4	100	19	-	-
Guyon[26]	100	8	-	-	-	-
Tibsrani[27]	100	21	100	43	-	-
Khan[17]	-	-	100	96	-	
Ramaswamy [34]	-	-	-	-	78	16063
Ramaswamy [34]	-	-	-	-	70.8	30
Ramaswamy [34]	-	-	-	-	75.5	6400

Finally it is evident from experimental results that the proposed method provides better accuracy and identifies smaller set of important genes than by other methods for multi-class cancer datasets.

VI. CONCLUSION

In this paper, the selection of a set of important genes for cancer classification of four well known microarray datasets has been done using hybrid of feature ranking and GA with multi-class SVM. Selection of a subset of important genes for multi-class problem is a multi objective optimization problem. On one hand, we have to maximize the classification accuracy and on the other we have to minimize the number of genes. We have transformed these two objectives of the task in hand into a single one by introducing a new fitness function in terms of classification accuracy and dimensionality of gene subset. The proposed method determines smallest subset of genes with 100% classification accuracy for Leukemia, SRBCT and Lymphoma datasets. For GCM dataset an accuracy of 78.26% is achieved with 38 genes. The number of genes obtained by the proposed method is smaller in size in comparison to the results found in the literature. It is also observed that the genes subset selected might not be same in different runs as GA is stochastic method.

REFERENCES

[1] Alon U., Barkai N., Notterman DA., Gish K., Ybarra S., Mack D., Levine AJ., "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* ,96,1990.

[2] Ben-Dor A., Bruhn L., Friedman N., Nachman I., Schummer M., Yakhini Z., "Tissue classification with gene expression profiles", *Journal of Computational Biology*, 7(3-4),pp.559-583, 2000.

[3] Golub TR., Slonim DK. *et al*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, 286**,** pp.531-537, 1999.

[4] Kohavi R., John G., "Wrapper for feature subset selection", *Artificial Intelligence*, 97(1-2), pp.273-324, 1997.

[5] Langley P., "Selection of relevant features in machine learning", *In AAAI Fall Symposium on Relevance*, 1994.
[6] Ding C., Peng HC., "Minimum redundancy feature selection from microarray gene expression data", *In IEEE Computer Society Bioinformatics Conf*, pp. 523-528, 2003.
[7] Jaeger J., Sengupta R., Ruzzo WL., "Improved gene selection for classification of microarray", *In PSB*, pp. 53-64. 2003.
[8] Li L., Weinberg CR., Darden TA., Pedersen LG. "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)**,** pp.131-142, 2001.
[9] Jourdan L., "Meatheuristics for knowledge discovery: Application to genetic data", *PhD thesis*, University of Lille, 2003.
[10] Peng S., Xu Q., Ling XB., Peng X., Du W., Chen L., "Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines", *FEBS Letter*, 555(2), pp**.**358-362, 2003.
[11] Deb K., Goldberg DE., "An investigation of niche and species formation in genetic function optimization", *In Schaffer J. D. (Ed) Proc. 3rd Internat. Conf. Genetic Algorithm, Morgan Kaufmann, San Mateo*, pp. 42-50, 1989.
[12] Bins J., Draper B., "Feature selection from huge feature sets", *In Proc. Internat. Conf. Computer Vision*, 2**,** pp**.**159-165, 2001.
[13] Hong JH., Cho SB., "Efficient huge scale feature selection with speciated genetic algorithm", *Pattern Recognition letters*, 27**,** pp**.**143-150, 2006.
[14] Huerta EB., Duval B., Hao J., "A hybrid GA/SVM approach for Gene Selection and Classification of microarray data", *EvoWorkshops 2006, LNCS 3907* , pp.34-44,2006.
[15] Reddy AR., Deb K., "Classification of two-class cancer data reliably using evolutionary algorithms", *Technical Report KanGAL*, 2003.
[16] Fu L.M., Liu CSF., "Evaluation of gene importance in microarray data based upon probability of selection", *BMC Bioinformatics*, 6(67), 2005.
[17] Khan J., Wei JS., Ringer M., Saal LH., Ladanyin, Westermann F., Berthold F., Schwab M., Antonescu CR., Petterson C., Meltzer PS., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", *Nat. Med.*, 7, pp.673-679, 2001.
[18] Li T., Zhang C., Ogihara MA., "Comparative study of feature selection and multi class classification methods for tissue classification based on gene expression", *Bioinformatics*, 20**,** pp.2429-2437, 2004.
[19] Souza BF., Carvalho APLF., "Gene Selection based on multi-class support vector machines and Genetic algorithms", *Genetics and Molecular Research"*, 4(3), pp.599-607, 2005.
[20] Li W., Yang Y., "How many genes are needed for a discriminant microarray data analysis in Critical Assessment of Techniques for Microarray", *Data Mining Workshop*, pp.137-150, 2000.
[21] Su Y., Murali T.M., Pavlovic V., Kasif S. "RankGene: identification of diagnostic genes based on expression data", *Bioinformatics*, pp.1578-79, 2003.
[22] http://www-genome.wi.mit.edu/cgi-bin/cancer/publications
[23] http://research.nhgri.nih.gov/microarray/supplement/.
[24] http://llmpp.nih.gov/lymphoma
[25] Dietterich TG., Bakiri G., "Solving multi-class learning via error-correcting output codes", *General of Artificial Intelligence Research***,** 2**,** pp.263-86, 1995.
[26] Guyon I., Weston J., Barnhill S., Vapnik V. "Gene Selection for cancer classification using support vector machines", *Machine Learning***,** 46, pp.389-422, 2003.
[27] Tibshirani R., Hastie T., Narasimhan B., Chu G., "Diagnosis of multiple cancer types by shrunken centroids of gene expression", *In Proc. Natl Acad. Sci.*, U.S.A., 99, pp.6567-6572, 2002.
[28] Lee Y., Lee C., "Classification of multiple cancer types by multi category support vector machines using gene expression data", *Bioinformatics*, 19**,** pp.1132-1139, 2003.
[29] Corts C., Vapnik VN., "Support Vector Networks", *Machine Learning*, 2**,** pp.273-297, 1995.
[30] Vapnik VN., *The Nature of Statistical Learning Theory*. Springer, Berlin Heidelberg New York 1995.
[31] Rifkin R., Klautau A., "In Defence of One-Vs.-All Classification", *Journal of Machine Learning*, 5**,** pp.101-141, 2004.
[32] Hsu CW., Lin CJ., "A comparison of methods for Multi-class Support vector machine", *IEEE Transactions on Neural Networks*, 13(2), pp.415-425, 2002.
[33] Goldberg DE., *Genetic algorithm in search, optimization and machine learning.* Addison Wesley, 1989.
[34] Ramaswamy S.,Tamayo P. *et al* , "Multiclass cancer diagnosis using tumor gene expression signature", *Proc Natl. Acad Sci. USA*, 98(26), pp 15149-15154,2001.