AN ENHANCED MULTICATEGORY CLASSIFICATION USING SVM-ANP FOR MICROARRAY GENE EXPRESSION CANCER DIAGNOSIS

Mrs. S. SASIKALA
Research Scholar, Dravidian University, AndraPradesh
Head, Department of Computer Science
Sree Saraswathi Thyagaraja College, Pollachi,
Coimbatore, Tamil Nadu,
INDIA

Dr. S. SANTHOSH BABOO
Reader, PG and Research department of Computer Science,
Dwaraka Doss Goverdhan Doss Vaishnav College
Chennai, Tamil Nadu,
INDIA

Abstract—This paper deals with the advanced and developed methodology know for cancer multi classification using Support Vector Machine (SVM) for microarray gene expression cancer diagnosis, used for directing multicategory classification problems in the cancer diagnosis area. SVMs are an appropriate new technique for binary classification tasks, which is related to and contain elements of non-parametric applied statistics, neural networks and machine learning.This paper deals with a fast and efficient classification method called the SVM-ANP approach for a multicategory cancer diagnosis problem based on microarray data. ANOVA ranking has been used for ranking the input gene data. From the ANOVA ranking technique, top ranked genes are selected. This ranking technique would improve the overall performance of the classification approach. Moreover, Analytic Network Process (ANP) is integrated with the SVM classifier. ANP allows for complex interrelationships among decision levels and attributes. SVM-ANP integrated system provides better results in improved testing accuracy and less training time when compared with standard SVM.

Key-Words: -Cancer classification, SVM, ANOVA, ANP, Classifier, SVM-ANPand Gene Expression

---

1 Introduction

DNA micro-arrays facilitate scientists to analyze several genes at the same time and to find out whether the genes are active, hyperactive or silent in normal or cancerous tissue [8, 3]. There are several techniques to analyze the micro array data, but most of these techniques generate confusing amounts of raw data. Hence, a novel analytical technique must be established to determine whether cancer tissues have unique signatures of gene expression over normal tissues or other types of cancer tissues [4].Recent investigations and explorations have described several new characteristic features and the practical applicability of Support Vector Machines (SVMs) in

knowledge discovery and data mining. SVMs were widely used to discover informative patterns [5].The present research work illustrates that SVMs are very efficient in classifying and recognizing informative features or attributes (such as critically significant genes). Statistical ranking technique is used in the present research work for the better classification results.

2AnalyticNetworkProcess(ANP)

ANP is a common type of Analytical Hierarchical Analysis (AHP). Saaty [SAT80, STY80, SAA06] suggested the use of AHP to handle the problem of independence on alternatives or criteria, and the use of ANP to solve the problem of dependence among alternatives. AHP approaches form a structure of the decisions that employs a one-way hierarchical relation with regard to decision layers. ANP was also introduced by Saaty. It is a generalization of the AHP [7]. The main difference between AHP and ANP is that, AHP represents a framework with a unidirectional hierarchical relationship but ANP is developed for the subjective evaluation of a group of alternatives according to multiple criteria organized in a hierarchical structure.

The top element of the hierarchy in AHP is usually the overall goal for the decision model. The hierarchy decomposes the common to more particular properties until a level of manageable decision criteria is obtained. ANP does not need this hierarchical structure; it facilitates factors to 'control' and be 'controlled' by the varying levels or `clusters' of attributes. Some controlling factors are also present at the same level [2].This interdependency among factors and their levels is defined as a 'systems with feedback' technique. AHP does not consist of feedback loops among the factors that can regulate weightings and reduce the opportunity of the reverse ranking technique. The relative significance of the impacts on a given element is calculated on a ratio scale related to AHP. ANP facilitates for complex interrelationships among decision levels and properties. ANP feedback technique replaces hierarchies with networks in which the relationships between levels are not easily denoted as higher or lower, dominated or being dominated, directly or indirectly [6].

For example, not only does the significance of the criteria decide the importance of the alternatives as in a hierarchy, but also the importance of the alternatives may have an impact on the importance of the criteria [STY80]. Thus, a hierarchical structure with a linear top-to-down form is not appropriate for a complex system. ANP approach is competent to deal with interdependent relationships among the elements by acquiring the composite weights through the development of a supermatrix. The supermatrix concept contains parallels to the Markov chain process [STY80], where relative importance weights are altered by forming a supermatrix from the eigenvectors of these relative importance weights. The weights are, then, altered by deciding products of the supermatrix.

3 Proposed Cancer Classification Technique Using ANOVA Ranking With SVM –ANP Classifier

This proposed system mainly deals with cancer prediction by using SVM-ANP classification technique. SVM technique uses ANOVA test for grouping up the sample amount of sequential data. SVM technique overcomes the previous classification methodology by means of time consumption and by giving best accuracy rate. This projected method is comprised of two steps. In Step 1, all genes in the training data set are ranked using a scoring scheme. Then, the genes with high scores are retained. In Step 2, the classification capability of all simple combinations is

tested among the genes selected in Step 1 using a good classifier. A new method of ranking with ANOVA and classifying with SVM is proposed in this chapter. The mechanisms for Step 1 and Step 2 are described as follows.

Step 1: Gene Importance Ranking
In Step 1, the importance ranking of each gene is computed using a feature ranking measure, two of which are described below. Only the most important genes are retained for Step 2.

3.1ANOVA (ANalysis Of VAriance*)*
ANOVA is a technique, which is often used in analysis of data, and to draw interesting information based on P-values. The ANOVA is known to be robust and assumes that all the sample populations are normally distributed with equal variance and all observations (samples) are mutually independent [1]. The approach chosen is the one-way ANOVA which performs an analysis on comparing two or more groups (samples) which in turn returns a single p-value that is significant for groups that are different from others. The most significant varying information has the smallest p-values. Within groups estimate of

$$\sigma_y^2 = \frac{\sum_{ij}(y_{ij}-\bar{y}_j)^2}{\sum_j(n_j-1)} = \frac{SS_{WG}}{df_{WG}} = MS_{WG} \quad (1)$$

Between –group estimate of

$$\sigma_y^2 = \frac{\sum_j n_j(y_j-\bar{y}_{ij})^2}{(K-1)} = \frac{SS_{BG}}{df_{BG}} = MS_{BG} \quad (2)$$

$$\text{F}(df_{BG}, df_{WG}) = \text{b}\frac{BetweenGroupestimateof \quad \sigma_y^2}{WithinGroupestimateof \quad \sigma_y^2} = \frac{MS_{WG}}{MS_{BG}} \quad (3)$$

Of all the information existing in the ANOVA Table, if the p value for the F- ratio is less than the critical value (α), then the effect is said to be significant. In this proposed approach, the α value is set at 0.05, any value less than this will result in important effects, while any value greater than this value will result in non-significant effects. The very small p-value indicates that differences between the column means (group means) are highly significant. The probability of the F-value arising from two similar distributions gives us a measure of the significance of the between-sample variation as compared to the within-sample variation. Small p-values indicate a low probability of the between-group variation being due to sampling of the within-group distribution and small p-values indicate interesting features. This study uses the p-values to rank the important features with small values and the sorted numbers of features are used for further processing.
Fig 4.3 shows the feature selection method of the proposed approach.
Initially, all the features are ranked using a feature ranking measure and the most important features alone are retained for next the step. After selecting some top features from the importance ranking list, the data set is attempted to classify with only one feature. In this proposed approach, the Support Vector Machine (SVM) classifier is used to test n-feature combinations.

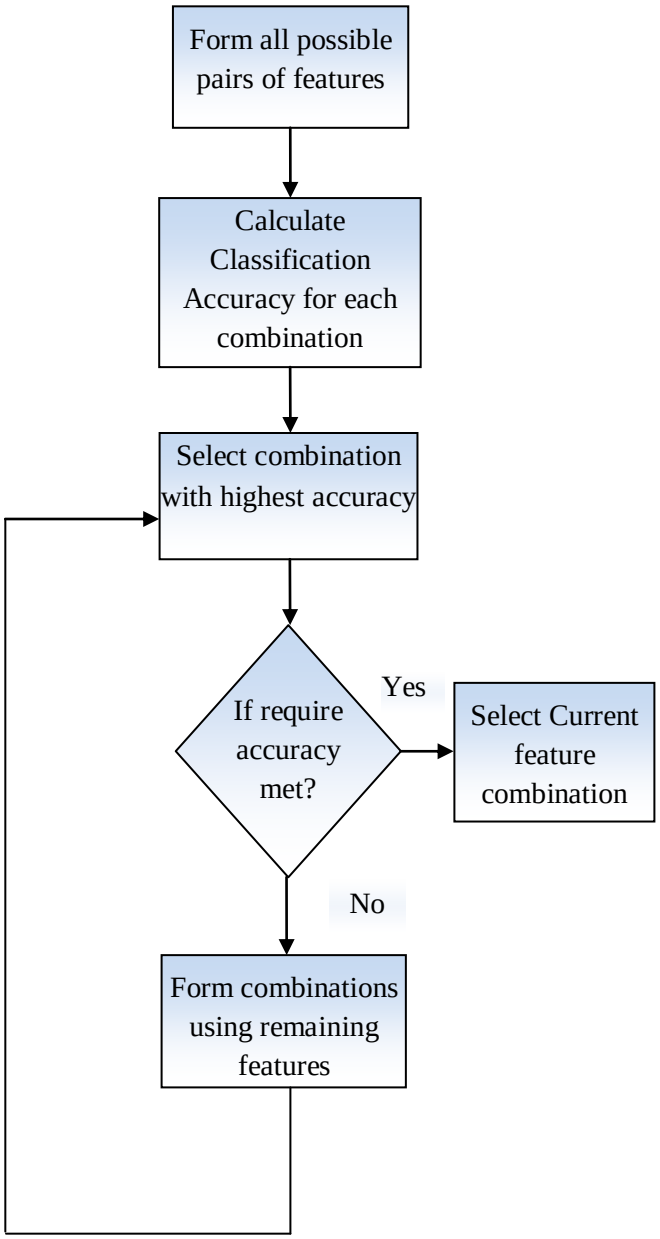

Fig 1 Feature Selection Method

Class Separability

Another frequently used method for gene importance ranking is the Class Separability (CS) . The CS of gene I is defined as

$$CS_i = SB_i / SW_i \quad (4)$$

$$SB_i = \sum_{k=1}^{K} (\bar{x}_{ik} - \bar{x}_i)^2 \quad (5)$$

$$SW_i = \sum_{k=1}^{K} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2 \quad (6)$$

For gene i, SB_i (the distances between samples of different classes) is the sum of squares of the interclass distances. SW_i (the distances of samples within the same class) is the sum of squares of the intraclass distances. A larger CS denotes a greater ratio of the interclass distance to the intraclass distance and, therefore, can be used to measure the capability of genes to separate different classes. In fact, the CS used here is similar to the F-statistic that is also widely used for ranking genes in literature. The difference between the CS and the F-statistic F is:

$$CS = F.(K-1)/(\sum_{k=1}^{K} n_k\text{-}1) \quad (7)$$

Because the term

$$F.(K-1)/(\sum_{k=1}^{K} n_k\text{-}1) \quad (8)$$

CS equation is a constant for a specific dataset; the CS can be regarded as a simplification of F-statistic. The two methods will guide to the same ranking results for the same data set.

Step 2: Finding the Minimum Gene Subset
After selecting some top genes from the importance ranking list, the data set is attempted to classify with only one gene. Each selected feature is given as input into the SVM-ANP classifier. If no good accuracy is obtained, continued classifying the data set with all the possible 2-feature combinations within the selected feature. If still no good accuracy is obtained, this procedure with 2-features combination is repeated and so on, until a good accuracy is obtained.

3.2 ANP Approach

The ANP is composed of four major steps [CLP06]:

Step 1:
Model Construction and Problem Structuring:The issue should be clearly described and decomposed into a rational system like a network. The structure can be obtained by the view of decision-makers through brainstorming or other suiTable techniques.

Step 2:
Pair-Wise Comparison Matrices and Priority Vectors:In ANP, decision elements at each component are compared pair-wise with regard to their significance towards their control criterion, and the components themselves are also compared pair-wise with respect to their contribution to the goal. Decision-makers are asked to react to a sequence of pairwise comparisons where two elements or two components at an instant will be compared the way in which they contribute to their specific upper level criterion [6]. Additionally, if there are interdependencies among elements of a component, pairwise comparisons also require to be generated, and an eigenvector can be acquired for each element to demonstrate the influence of other elements on it. The relative importance values are obtained with Saaty's 1-9 scale, where a score of 1 denotes equal significance between the two elements and a score of 9 denotes the intense significance of one element (row component in the matrix) compared to the other one (column component in the matrix) [6].

A reciprocal value is allotted to the inverse comparison; that is $a_{ij} = \frac{1}{a_{ji}}$, where a_{ij} (a_{ji}) represents the significance of the *i*th (*j*th) element. Like AHP, pairwise comparison in ANP is made in the framework of a matrix, and a local priority vector can be derived as an estimate of

relative significance related with the elements (or components) being compared by solving the following equation:

$$A \times w = \lambda_{max} \times w$$

where A denotes the matrix of pair-wise comparison, w denotes the eigenvector, and λ_{max} denotes the largest eigenvalue of A. Saaty [STY80] proposes various techniques for approximating w. In this approach, the following three-step procedure is used to synthesize the priorities [CLP06]:

- Sum the values in each column of the pair-wise comparison matrix;
- Partition each element in a column by the sum of its respective column. The resultant matrix is referred to as the normalized pair wise comparison matrix;
- Sum the elements in each row of the normalized pair-wise comparison matrix, and divide the sum by the n elements in the row. These final numbers provide an estimate of the relative priorities for the elements being compared with respect to its upper level criterion. Priority vectors must be derived for all comparison matrices.

Step 3:

Supermatrix Formation*:* The supermatrix concept is similar to the Markov chain process [7]. In order to acquire global priorities in a system with interdependent influences, the local priority vectors are entered in the resultant columns of a matrix. Thus, a supermatrix is actually a partitioned matrix, where each matrix segment denotes a relationship between two nodes (components or clusters) in a system [6].

$w^*_k =$

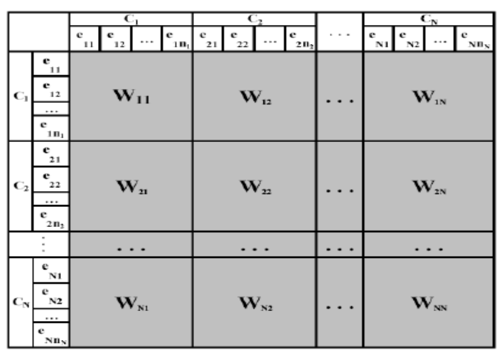

Fig2 : The Supermatrix of a Network

Let the components of a decision system be $C_k, k = 1,2, \ldots, n$, and each component k has m_k elements, denoted by $e_{k1}, e_{k2}, \ldots, e_{kmk}$. The local priority vectors obtained in Step 2 are grouped and located in suiTable positions in a supermatrix depending on the flow of influence from one component to another component, or from a component to itself as in the loop. A standard form of a supermatrix is shown in Fig2[7].

Step 4:

Selection of Best Alternatives:If the supermatrix produced in Step 3 covers the entire network, the priority weights of alternatives can be found in the column of alternatives in the normalized supermatrix. Alternatively, if a supermatrix only contains components that are interrelated, additional computation must be made to attain the overall priorities of the alternatives. The alternative with the largest overall priority should be the one selected.

3.3 Proposed SVM-ANP Classification Approach

This section describes the proposed cancer classification technique which uses ANP and SVM techniques. The raw input genes are ranked by using ANOVA ranking technique. The output of the ANOVA is the top ranked genes with good features. Then the ranked genes are classified using the SVM classifier which uses the ANP technique for updating the weight vector w. The overall algorithm of the proposed cancer classification approach which uses SVM and ANP is given below:

3.3.1 SVM-ANPClassification Algorithm

Step 1: The raw gene input data is given to the ANOVA ranking technique.

Step 2: Features are selected and the genes are ranked based on the ANOVA ranking.

Step 3: Top Ranked Genes are given as input to the SVM classifier.

Step 4: A two-class training set $(x_1, y_1), \ldots, (x_n, y_n)$ is considered, with $x_i \in \Re^m$ and $y_i = \{+1, -1\}$.

Step 5: A classifier consisting of a weighted sum of the form $f_{SVM}(x) = sign[w.\Phi(x)]$
Where w is the vector hyperplane, $\Phi: \Re^m \rightarrow \Re^M$ is a non-linear function that maps the input space to a *feature* space of a much higher, possibly infinite, dimension M>>m.

Step 6: The weights are expanded by the kernel function called Radial Basis Function.

$$f_{SVM}(x) = sign\left[\sum_{i=1}^{n} y_i\, \alpha_i \Phi(x_i).\Phi(x)\right] = sign\left[\sum_{i=1}^{n} y_i\, \alpha_i\, \mathrm{K}(x_i, x)\right]$$

$$\mathrm{K}(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right)$$

Step 7: The value w is also in the transformed space, with $w = \sum_i \alpha_i y_i \Phi(x_i)$.

Step 8: Choose the starting values of input weight vectors w with the help of ANP technique.

Step 9: Construct matrix $a_{kk'}$

Step 10: The normalized $a'_{k,k'}$ matrix is obtained by the following equation

$$a'_{k,k'} = \frac{a_{k,k'}}{\sum_{k=1}^{n} a_{k,k'}}$$

Step 11: Supermatrix is obtained by the following equation: $w^*_k \times a'''_{k,k'}$
This weight factor is updated to the SVM classifier for classification of the genes.

3.3.2 ANP Algorithm to Find the Weight Matrix (w)

Based on the pair-wise comparison of matrix *E* and after developing the hierarchy of ANP, the next stage creates matrices considering the interaction between binary items for the factors and sub factors, the feedback between factor groups, and the internal dependency between factor groups. The selection process is modified to a five–step hybrid procedure, as follows:

Step 1:

Note that in ANP, the pair-wise comparison matrix *A* on the diagonal has the rank of the element $a_{kk'}$reflecting the evaluation of *k* over gene k'. If $a_{kk'} < 1$, it means that unit *k* is evaluated less than gene K . $a_{kk'}$ is constructed from the paired ANOVA results; the sum of the evaluations are taken given to unit *k* by the models of both units and divide it by the sum of the evaluation given to unit by the models of both units $a_{kk'}$ and $a_{k'k}$. Obviously, $a_{k'k}$=$\frac{1}{a_{kk'}}$as performed by the ANP. The components of the binary comparative matrix are obtained via the following formula. (This formula shows the *k* decision gene efficiency, and the k'decision unit.)

$$a_{k,k'} = \frac{e_{k,k'}}{e_{k',k}}$$

Step 2:

At this step, each component of the matrix obtained at the second step is divided by that column's total
value. The matrix obtained here is a normalized

$$a'_{k,k'} = \frac{a_{k,k'}}{\sum_{k=1}^{n} a_{k,k'}}$$

Step 3:

Here, the column vector elements are found by summation over the rows.

$$a''_{k,k'} = \sum_{k=1}^{n} a'_{k,k'}$$

Step 4:

This step is the normalization of the column vector.

$$a'''_{k,k'} = \frac{a''_{k}}{\sum_{k=1}^{n} a''_{k}}$$

Step 5:

The last step, the $a'''_{k,k'}$and w^*_kmatrixes that are obtained in the first step are multiplied, and the relative dependant priorities of factors obtained. Fig 4.5 demonstrates the supermatrix of a network. A related process appears in the following formula

$$w^*_k \times a'''_{k,k'}$$

This weight factor is given as input to the SVM classifier.

4 Experimental Results and Discussion

An evaluation study of the proposed cancer classification algorithm is presented in this section. The results of an extensive set of simulation tests are shown, in which the proposed cancer classification technique is compared under a wide variety of different scenarios. For this research work MATLAB has been taken into consideration and the proposed technique has been

implemented using MATLAB. The platform used for this proposed approach is Windows XP. The processor used is Pentium IV. The experimentation needs a system RAM of 2 GB.

4.1 Experimental Process

The SVM classifier is applied to classify the lymphoma micro array data set. Initially, the selected 100 genes are added one by one to the network according to their ANOVA ranks. That is, only a two gene that is ranked 1 is used as the input to the network. Then the network is trained with the training data set and subsequently, tested the network with the test data set.

For a microarray data with n genes, each ANOVA classifier produces a hyperplane w, which is a vector of n elements, each corresponding to the expression of a particular gene [PJV99, AED00]. The absolute magnitude of each element in w can be considered as a measure of the importance of each corresponding gene [GST99]. Each ANOVA-SVM classifier is first trained with all of the genes, then the gene corresponding to the bottom 10 percent, w_{ij}, are removed. Each classifier is then again trained after the removal of genes. This process is repeated with iterations and a rank of all of the genes based on the statistical significance of each class can be obtained.

SVM is a machine classification technique that directly minimizes the classification error without requiring a statistical data model [MWD00]. This technique is popular since its implementation is very easy and achieves consistently high classification accuracy when applied to many real-world classification situations. The SVM algorithm can be used for both classification and regression (model fitting) problems. In classification, an SVM classifier can separate data that are not easily separable in the original data space by mapping data into a higher dimensional (transformed) space. Kernel functions are used by SVM to find a hyperplane that maximizes the distance (margin) between the two classes, while minimizing training error. The resultant model is sparse, depending only on a few training samples (the "support vectors"). The number of support vectors gets increase linearly with the available training data, requiring much higher computational complexity when classifying very large data sets (e.g., tens or hundreds of thousands of input variables).

4.2 Training and Testing Machine Learning Classifiers

The proposed technique uses Five and Ten-fold cross-validation to train and test SVM classifiers to avoid training and testing on the same data. In Ten fold CV, first, affected and healthy genes were randomly divided into 10 approximately equal, exhaustive, and mutually exclusive subsets. Next, classifiers were trained on 9 subsets and then tested on the 10th subset. This sequence was repeated 10 times, using each subset serving as the test set one time, so that each tested gene was never part of its training set and was tested only once.

4.2.1 PerformanceEvaluation

The performance of the proposed approaches are evaluated using the following parameters like

- Testing Accuracy
- Training Time

4.2.2 Datasets Used

The dataset used in this experimental analysis for the evaluation of the proposed approaches is Lymphoma Dataset.

Lymphoma

The lymphoma microarray data (http://llmpp.nih.gov/lymphoma), has three subtypes of cancer, i.e., CLL, FL, and DLCL. When applying the proposed method to this data set, the clustering result with two best partition eigenvectors is obtained. Seen from cluster results the three classes are correctly divided. Then two sets of l=20 genes are selected according to |Ri, 1| and |Ri, 2| respectively. (Here set have to be two.)

Table 1: GENE IDS (CLIDS) AND GENE NAMES IN THE TWO MICROARRAY DATASETS

Data Set	Gene ID/ CLID	Gene Name	Gene Rank	
			G1	G2
Lymphoma	GENE 1622X	*CD63 antigen (melanoma 1 antigen); Clone=769861	3	/
	GENE 2328X	*FGR tyrosine Kinase; Clone=728609	/	3
	GENE 3343X	*mosaic protein LR11=hybrid; Receptor gp250 precursor; Clone=1352833	/	4

From the two sets of 20 genes each, the two-gene combinations is chosen that can best divides the lymphoma data. Two pairs of genes have been found: 1) Gene 1622X and Gene 2328X, and 2) Gene 1622X and Gene 3343X, which perfectly divide the lymphoma data. Since the results are similar to each other, only the result of one group is shown. Gene ID and gene names of the selecting genes in the lymphoma data set are shown in Table 4.1, where the group and the rank of genes are also shown.

In Lymphoma dataset, there are

- 12 samples obtained from Diffuse Large B-Cell Lymphoma (DLBCL)
- 20 samples from Follicular Lymphoma (FL)
- 18 samples from Chronic Lymphocytic Leukemia (CLL)

Initially, the 62 samples are randomly partitioned into two parts:

- 31 samples for training
- 31 samples for testing

The entire set of 4,026 genes is ranked based on their ANOVA in the training data set. Next, the 100 genes with the highest ANOVA ranks are selected and evaluate one to one combinational examination and two to one combinational examination. The whole data set contains the expression data of 4,026 genes. Few samples are missing the dataset. For filling those missing, k-nearest neighbor algorithm is used in this work [9].

The SVM was implemented with the use of Platt's sequential minimal optimization algorithm in commercial software (Mat Lab, ver. 5.0; The Math Works, Natick, MA). For classification of the gene expression data, Gaussian (nonlinear) kernels of various widths were tested, and a Gaussian

kernel with width = √(2 × number of input variables) was chosen that gave the highest area under the receiver operating characteristic curve using 5-fold and 10-fold cross-validation.

Table2:A SAMPLE 2-GENE COMBINATION

Gene 1	Gene 2	Correct Rate	Error Rate	No. of Matches	Mismatches
1	2	0.9032	0.0968	4	1
1	3	0.9032	0.0968	5	0
1	4	0.8710	0.1290	6	0
1	5	0.9355	0.0645	6	0
1	6	1.0000	0	6	0
1	7	0.9032	0.0968	6	1
1	8	0.9355	0.0645	6	0
1	9	0.9355	0.0645	5	1
1	10	1.0000	0	5	0

Table2 shows a sample 2-gene combination. The parameters used in the above Table for the validation of the accurate combinations are Correct Rate, Error Rate, Number of Matches and Mismatches. In the above Table, (1, 6) gene combination provides highest correct rate, less error rate, less mismatches with significant number of matches.

Table3:Maximum accuracy achieved by the following 2-gene combinations

(1,6)	(1,10)	(4,6)	(6,11)	(6,12)	(6,14)	(6,15)	(6,16)
(6,17)	(6,19)	(21,23)	(21,25)	(21,26)	(21,32)	(21,65)	(21,73)
(21,81)	(32,35)	(32,45)	(32, 53)	(41,46)	(41, 51)	(51,63)	(77,77)

Table3 shows maximum accuracy achieved by the genes in the lymphoma dataset. Among the 100 genes, 24 gene combinations achieved very good accuracy which are listed in the above Table. From this gene combinations, the best gene which provides highest correct rate, less error rate, less number of mismatches and high number of matches is the (1,6) gene combination since this pair provides less error rate and less number of mismatches.

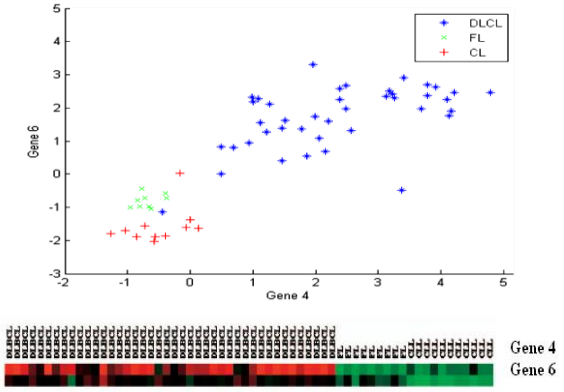

Fig3: Gene Expression Levels of 2-Gene Combination for (4, 6)

Then, smallest gene subsets that can ensure highly accurate classification for the entire data set is searched. Initially, it attempted to classify the data set using two genes tested for all possible combinations within the 100 genes.

The expression profiles of the 2-gene combinations in the lymphoma data for (4, 6) gene is presented in Fig3. This 2-Gene combination can separate DLBCL, FL, and CLL effectively.

Table4 shows a sample 3-gene combination. The parameters used in the above Table for the validation of the accurate combinations are Correct Rate, Error Rate, Number of Matches and Mismatches. In the above Table, (1, 5, 6) gene combination provides highest correct rate, less error rate, less mismatches with significant number of matches. Thus, it is the best 3-gene combination obtained from the experimental observation.

Table 4:A SAMPLE 3-GENE COMBINATION

Gene 1	Gene 2	Gene 3	Correct Rate	Error Rate	No. of Matches	Mismatches
1	2	3	0.8952	0.0881	4	1
1	3	4	0.8862	0.0910	5	1
1	4	5	0.9110	0.1001	5	0
1	5	6	1.0000	0	6	0
1	6	7	0.9256	0.0845	6	0
1	7	8	0.9534	0.0726	5	1
1	8	9	0.9355	0.0723	6	0
1	9	10	0.8953	0.0723	5	1
1	10	11	1.0000	0	6	0

Table 5:Maximum accuracy achieved by the following 3-gene combinations

(1,5,6)	(1,7, 10)	(2,4,5)	(2,4,6)
(6,8, 11)	(6,8, 12)	(6,21, 23)	(21,21, 28)
(21,23, 31)	(21,31,43)	(21, 36, 53)	(53,75, 81)
(52,64,78)	(43,48, 68)	(43, 87, 98)	(63, 73, 81)

Table5 shows maximum accuracy achieved by the genes in the lymphoma dataset. From the 100 genes, the gene combinations with very good accuracy are listed in the above Table. From this gene combinations, the best gene which provides highest correct rate, less error rate, less number of mismatches and high number of matches is the (1,5, 6) gene combination.

Table 6:ACCURACY COMPARISON USING ANOVA WITH NUMBER OF FOLDS=5

S. No	No. of Gene Combination	Accuracy (%)	
		SVM	SVM with ANP
1	100,2	87.12	91.25
2	100,3	89.45	92.14

Table 6 shows the testing accuracy of the SVM classifier with the SVM-ANP classification approach. Table provides the result for both the 2-Gene and 3-Gene combinations. It is clearly observed from the Table that the proposed cancer classification approach using SVM with ANP provides 91.25 % testing accuracy where as the standard SVM cancer classification provides 87.12% testing for 2-Gene combination.

Similarly, for the 3-Gene combination, the testing accuracy of the proposed cancer classification approach which uses SVM with ANP is 92.14% where as for the standard SVM, it is just 89.45%.

Thus the proposed cancer classification approach which uses SVM-ANP technique provides significant testing accuracy for both 2-Gene and 3-Gene combinations.

Fig 4 shows the graphical representation of the testing accuracy comparison for the proposed SVM-ANP cancer classification technique and the standard SVM cancer classification technique for the 5-fold CV. The Fig clearly shows that the proposed SVM-ANP approach provides better testing accuracy than the SVM cancer classification technique.

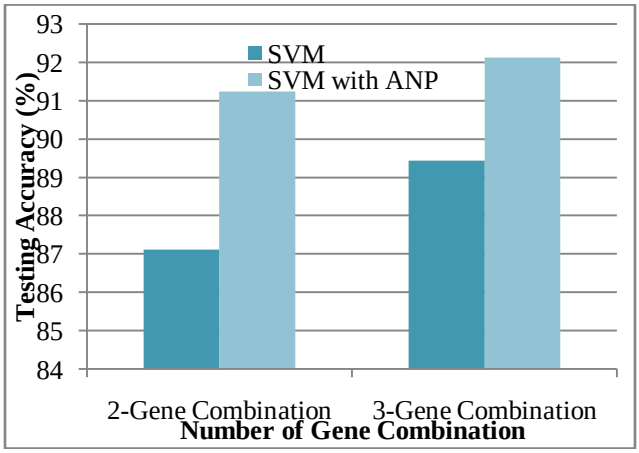

Fig 4 Testing Accuracy Comparison for 5-fold CV

Table 7:ACCURACY COMPARISON USING ANOVA WITH NUMBER OF FOLDS=10

S. No	No. of Gene Combination	Accuracy (%)	
		SVM	SVM with ANP
1	100,2	87.12	89.11
2	100,3	88.02	90.78

The average testing accuracy of the SVM and SVM with ANP with 10-fold CV is observed in Table 4.7. It is clearly observed from the Table that the proposed SVM with ANP provides higher accuracy when compared with the standard SVM for both 5 fold and 10 fold CV test.

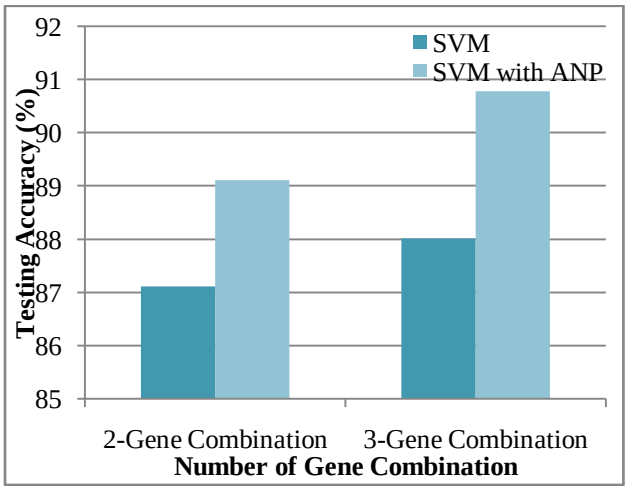

Fig5: Testing Accuracy Comparison for 10-fold CV

Fig5 shows the testing accuracy comparison of the cancer classification techniques for 2-Gene and 3-Gene combinations with 10-fold CV. For both the 2-Gene and 3-Gene combinations, the proposed SVM-ANP cancer classification approach is observed to provide better results.

Table 8:TRAINING TIME COMPARISON FOR 5-FOLD CV

S. No	No. of Gene Combination	Training Time	
		SVM	SVM with ANP
1	100,2	33.024	18.021
2	100,3	45.157	21.985

Training time taken by the cancer classification approach which uses SVM and SVM-ANP is tabulated in the above Table. The average training time taken by the proposed SVM-ANP classification technique with standard SVM for the lymphoma dataset is compared the Table 4.8. It clearly shows that the proposed SVM-ANP cancer classification approach is processed in very less time when comparing with the SVM classification approach.

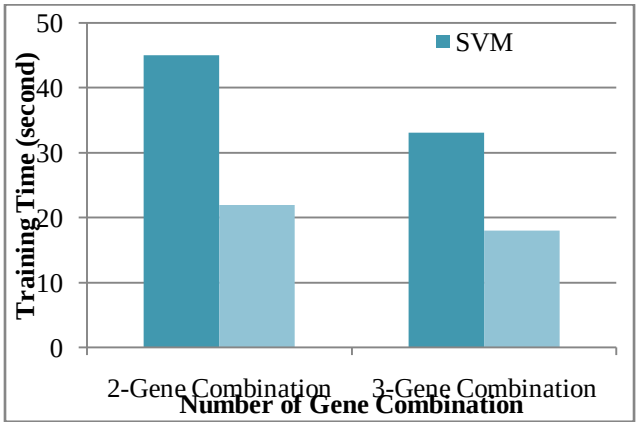

Fig6: Training Time Comparison

Fig6 shows the training time comparison of the cancer classification approaches. The Fig clearly shows that for both the 2-Gene and 3-Gene combinations, the proposed RVM-ANP approach takes less time for classification.

References:

[1]Armstrong. R. A and Hilton. A, "The use of analysis of variance (ANOVA) in applied microbiology", Microbiologist Vol. 5, No. 4, Pp. 18 – 21, Pp. 2097-2116, 2004.

[2] Babak Daneshvar Rouyendegh and Serpil Erol, "The DEA – FUZZY ANP Department Ranking Model Applied in Iran Amirkabir University", Acta Polytechnica Hungarica, Vol. 7, No. 4, 2010.

[3] Collins. F. S, Brooks. L. D and Chakravarti,A, "A DNA polymorphism discovery resource for research on human genetic variation", Genome Res., Vol. 8, Pp. 1229–1231, 1998.

[4] Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., and Fodor, S. P., "Accessing genetic information with high-density DNA arrays", Science, Vol. 274, Pp. 610-614, 1996.

[5] I. Guyon, N.Matic , V. Vapnik. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Discovering informative patterns and data cleaning", Editors, Advances in Knowledge Discovery and Data Mining, Pp. 181-203, 1996.

[6] L. M. Meade, and J. Sarkis, "Analyzing Organizational Project Alternatives for Agile Manufacturing Processes: An Analytical Network Approach", International Journal of Production Research, Vol. 37, Pp. 241-261, 1999.

[7] T.L Saaty, "Decision Making with Dependence and Feedback-The Analytic Network Process", RWS Publications, Pittsburgh, 1996.

[8] G. W. Wayt, "The unseen genome: beyond DNA. Scientific American", Vol 289, No. 6, Pp. 106-114, 2003.

[9] Olga G. Troyanskaya, Michael Cantor, Gavin Sherlock, Patrick O. Brown, Trevor Hastie, Robert Tibshirani, David Botstein and Russ B. Altman, "Missing Value Estimation Methods for DNA Microarrays,"Bioinformatics/computer Applications in The Biosciences, Vol. 17, No. 6, Pp. 520-525, 2001.