Two Step Feature Extraction Method for Microarray Cancer Data using Support Vector Machines

C.Arunkumar
Assistant Professor (Sr.Grade)
Department of Information Technology
Amrita School of Engineering, Coimbatore,
Tamilnadu, India

S.Ramakrishnan, Ph.D
Professor and Head
Department of Information Technology
Dr.Mahalingam College of Engineering and
Technology, Pollachi, Tamilnadu, India

ABSTRACT
Diagnosis of cancer is one of the most emerging clinical applications in microarray gene expression data. However, cancer classification on microarray gene expression data still remains a difficult problem. The main reason for this is the significantly large number of genes present relatively compared to the number of available training samples. In this paper, a novel approach to feature extraction combining the statistical t-test and absolute scoring is proposed for achieving better classification rate. Suitable classification approaches using the linear Support Vector Machines, the Proximal Support Vector Machines and the Newton Support Vector Machines is also discussed. A comparative analysis on the different techniques for feature extraction is also presented. Microarray cancer data based on Adenoma and Carcinoma with 7086 and 7457 genes of 4 and 18 patients respectively is used for this study. Increase in the classification rate of the proposed new method is clearly demonstrated in the results.

General Terms
Neural Networks, Support Vector Machines.

Keywords
Normalization, Linear SVM, Proximal SVM, Newton SVM, absolute scoring

1. INTRODUCTION
Genetic information of cells is stored in DNA and all cells in an organism have exactly the same genome. However, due to different tissue types, different development stages, and even different environmental conditions, genes that are present in cells in the same organism can be expressed in different combinations and/or different quantities during transcription from DNA to messenger RNA (mRNA). These different gene expression patterns, including both the combination and quantity, account for the huge variety of states and types of cells in an organism. Different organisms have different genomes and different gene expression patterns. Recently, gene expression microarrays (including the cDNA microarray and the GeneChip) have been developed as a powerful technology for functional genetics studies, which simultaneously measure the mRNA expression levels of thousands to tens of thousands of genes. A typical microarray expression experiment monitors the expression level of each gene multiple times under different conditions or in different phenotypes. For example, a comparison can be made between healthy tissue and cancerous tissue or one kind of cancerous tissue versus another. By collecting such huge gene expression data sets, it opens the possibility of distinguishing phenotypes and identifying disease-related genes whose expression patterns are excellent diagnostic indicators. A typical gene expression data set is extremely sparse compared to a traditional classification data set. The gene expression data usually comes with only dozens of tissue samples but with thousands or even tens of thousands of gene features. This extreme sparseness is believed to deteriorate the performance of a classifier significantly [29]. As a result, the ability to extract a subset of informative genes while removing irrelevant or redundant genes is crucial for accurate classification. For example, the Adenoma data has only 8 samples (tissues) with 7,086 features (gene expression measurements). If the process of gene selection is ignored, the researcher or biologist would need to discriminate and classify very few samples in a very high-dimensional space. It is unnecessary or even harmful for classification because it is believed that no more than 10 percent of these 7,086 genes are relevant to Adenoma classification. Furthermore, it is also helpful for biologists to find the inherent cancer mechanisms to develop better diagnostic methods and find better therapeutic treatments. From the viewpoint of data mining, this problem of gene selection is essentially a feature selection or dimensionality reduction problem. The ultimate goal of a good dimensionality reduction method is to remove irrelevant or redundant features while keeping informative or important features for classification. A lower dimensional feature spaced model is expected to capture the inherent data distribution better and, thus, produces a better performance. SVMs are known to be suitable for high dimensional microarray data and are able to classify non-linear relationships in the data through the use of kernel functions specific to the datasets. There are many pattern classification algorithms available, but not all provide a ranking by significance of genes. Ranking of the genes allows for selection of a small manageable subset of important genes. The support vector machine is well suited to this problem due to its ability to generalize classifications from high-dimensional, small sample size datasets. Datasets usually require some level of preprocessing to produce optimal gene ranking results. The raw microarray data can be normalized using a number of common methods which may affect significant gene identification. In the case of the SVM, the choice of kernel function can also affect classifier prediction results.

2. PROPOSED METHOD
This paper proposes a new method in feature extraction that combines the t-test statistics and the absolute scoring method. This proposed method produces higher accuracy in the classification rate when linear, proximal or Newton SVM is used for classification. The steps involved are given as under:

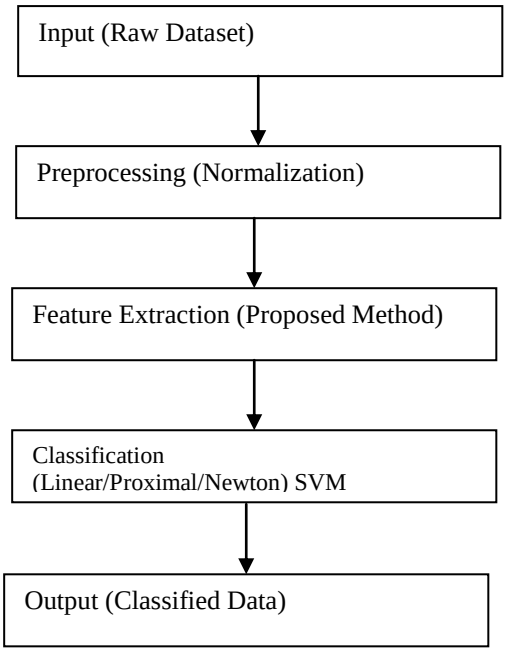


Figure 1. Architecture of the proposed method

2.1 Normalization

Microarray data normalization adjusts the experimental and control fluorescence intensities to account for any biases that arise from the microarray experimental methodology. Normalization procedures attempt to remove non-biological variances within microarray datasets [4], [27], [28]. Normalization is the process of reducing unwanted variation either within or between arrays. Multiple chips might serve as a source of information. Typical assumptions of most major normalization methods are (one or both of the following): Only a minority of genes are expected to be differentially expressed between conditions and any differential expression is as likely to be up-regulation as down-regulation (i.e., about as many genes going up in expression as are going down between conditions). Several types of normalization exist. Our paper uses the min-max normalization procedure as shown in the Table below:

Table 1. Computation of the normalized values using min-max norm

C1	C2	C3	C4	max(C1-C4)	min(C1-C4)	Norm1	Norm2	Norm3	Norm4
29.84	46.64	53.59	31.87	53.59	29.84	0.00	0.71	1.00	0.09
0.93	2.17	11.03	-0.85	11.03	-0.85	0.15	0.25	1.00	0.00
3.26	-1.63	1.58	5.52	5.52	-1.63	0.68	0.00	0.45	1.00
10.72	20.61	14.71	-5.52	20.61	-5.52	0.62	1.00	0.77	0.00
3.26	4.34	8.93	4.25	8.93	3.26	0.00	0.19	1.00	0.17

The formula to compute the min-max normalization is given as under in Eq. 1

$$Min - Max\ norm = (C1 - \min(C1 - C4)/ (\max(C1 - C4) - \min(C1 - C4)) \quad \text{(Eq. 1)}$$

From the above table and Eq. 2, C1-C4 represents the un-normalized gene expression values. The maximum and the minimum values are determined. Norm 1 is calculated by taking the ratio of the difference between the un-normalized value (C1) and the minimum value in the group (min (c1-c4)) to the difference between the maximum and the minimum value among the group. Similarly the other norm values are computed. The data is normalized to a scale [0 1]. The process of normalization has not improved the accuracy of the classifier as the values are not extracted from the microarray slide. Instead our analysis uses the extracted microarray gene expression values stored in Microsoft Excel worksheets.

2.2 Dataset used

The dataset is downloaded from the Princeton University Gene Expression Project. It consists of adenoma and carcinoma data. Two types of datasets are available. The first type is the raw dataset and the second type is the housekeeping genes. We use Type I (raw dataset). The features are extracted from the raw dataset. The adenoma data consists of 7086 genes of four patients each suffering from cancer. It also has 7086 gene information of the same four patients under the controlled environment (say after the inducement of drugs). The carcinoma data consists of 7457 genes of 18 patients suffering from cancer. It also has 7457

gene information of the same 18 patients under the controlled environment (probably after the inducement of drugs) [5].

2.3 t-test computation

The t-statistics method is used commonly for a ranking, selection and a two-class prediction [6]. The t-test computation is a six step process. The sequences of steps to compute the t-test value are:

1. Compute the average of the gene expression values of the cancerous samples for the first gene.
2. Compute the average of the gene expression values of the non-cancerous samples for the first gene.
3. Calculate the absolute difference between the average of the cancerous and non-cancerous sample.
4. Compute the standard deviation of the gene expression values of the cancerous samples for the first gene.
5. Compute the standard deviation of the gene expression values of the non-cancerous samples for the first gene.
6. The t-test value is calculated using the formula as shown in Eq. 2

$$t = (Mx - My) / \sqrt{((Sx2/Nx) + (Sy2/Ny))} \quad \text{(Eq. 2)}$$

7. Table 2 shows the calculation performed from steps 1 through 6 for the first 5 genes. The above series of steps are performed for the remaining genes in the sample.
8. Based on the t-test values the genes are ranked in the decreasing order. The top 1000 genes are selected and are given as the input to the absolute scoring method as described in the next part.

Table 2. Computation of the t-test value

C1	C2	C3	C4	N1	N2	N3	N4	A1-Average(C1-C4)	A2-Average(N1-N4)	Abs Diff	Std Dev(C1-C4)	Std Dev(N1-N4)	t-test value
3179.24	3102.06	3093.47	3410.58	742.01	1464.82	917.84	979.14	3196.34	1025.95	2170.3	147.95	309.36	4.75
148.26	113.89	133.45	120.27	24.28	36.87	29.40	34.88	128.97	31.36	97.61	15.22	5.68	4.67
187.43	179.51	145.53	184.02	41.76	17.38	13.45	31.01	174.12	25.90	148.22	19.33	12.98	4.59
2707.88	2264.18	2416.24	2683.42	569.13	932.30	679.66	925.36	2517.93	776.61	1741.3	214.63	181.49	4.40
161.78	144.26	151.84	146.62	69.44	85.33	68.76	54.75	151.12	69.57	81.55	7.78	12.50	4.02

2.4 Feature Extraction based on Absolute Scoring

The data that is taken from the Princeton microarray database is raw data. This dataset contains thousands of gene information. Out of the thousands of different genes available for analysis, only a few hundred genes contain relevant information to identify whether a tissue is cancerous or not [7]. In order to extract the features, we use a specific algorithm as in [8]. The sample results are tabulated as under in Table 2. The five different steps followed to extract the relevant genes are:

1. Obtain the mean of the expression values for each gene of cancerous samples and mean of the expression values for each gene of normal samples.
2. Obtain absolute difference between the mean of cancerous samples and the mean of normal samples.
3. Arrange the genes based on absolute difference in decreasing order.
4. Select Top 250 genes.
5. Apply the following formula on selected 250 genes.

$$F(xi) = (\mu(\text{cancerous}) - \mu(\text{normal})) / (S(\text{cancerous}) + S(\text{normal}))$$

where μ is the mean and S is the standard deviation. Select 200 genes with highest absolute F (xi) scores as our top features.

Table 3. Sample Dataset of Feature Extraction

Normal(1)	Normal(2)	Normal(3)	Normal(4)	Mean	STDEV	Abs Diff of Mean	Add STDEV	Abs Diff/ADD STDEV
361.29	592.04	471.38	444.27	467.2443	95.47043	361.9600576	3.3983986	106.5090065
705.59	771.65	749.92	656.48	720.9075	50.99847	663.1135372	11.070606	59.89857464
227.26	436.65	306.94	239.33	302.5489	96.03305	220.0679359	5.0084692	43.93916138

2.5 Proposed Two step Feature Extraction Method

Extracting a subset of informative genes from microarray expression data is a critical step in the diagnosis and treatment of cancer[20]. Microarray data consists of thousands of genes that are used to evaluate the expression levels. Most of the genes are not related to cancer. The challenge on microarray data is feature selection that searches for a subset of genes that are responsible for the cause of cancer[21]. This paper proposes a two-step feature extraction method. In the first step, the raw data is processed using the t-test statistics method. The genes are ranked in descending order based on their top t-test values. Then the second method of feature

extraction based on absolute scoring is used. The top 300 genes are extracted from both the classes. This extracted data is fed as the input to the Support Vector Machines classifier. The linear SVM, proximal and Newton SVM classifier classifies the input data and the percentage of accuracy of the classifier is predicted accordingly.

3. CLASSIFICATION USING SVM

The Support Vector Machines (SVM) is a supervised learning technique. A Support Vector Machine (SVM) performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories. SVM models closely resemble the neural networks. SVM's that makes use of a kernel function are an alternative training method for polynomial, radial basis function(RBF) and multi-layer perceptron (MLP) classifiers in which the weights of the network are found by solving a quadratic programming problem with linear constraints. On the other hand, a standard neural network finds a solution by solving a non-convex, unconstrained minimization problem. In fact, a SVM model using a sigmoid kernel function is equivalent to a two-layer, perceptron neural network. Identification of genetic markers is a crucial step in the diagnosis, prognosis and treatment of cancer[23]. The classification process takes place based on the reference vectors that are known in advance. Gene expression vectors are mapped from the expression space to a higher level feature space as depicted in Fig 2. The distance measurement is based on the mathematically based kernel function which then performs the process of classification and clustering. If the kernel function is not chosen properly, SVM will not be able to find an optimal separating hyperplane in feature space. SVM is linear since it makes use of the hyperplane to separate the two distinct classes. The separating hyperplane is selected so that the margin between the separating surfaces that split the positive and negative feature vector space is maximum. This is done to avoid overfitting. After the separating hyperplane is selected, then the computational burden gets reduced greatly as the computation of the decision function involves the inner dot product of the points in the feature space[26].

There are four main advantages: Firstly it has a regularization parameter, which makes the researcher think about the problem of over-fitting and the method to avoid it. Secondly it uses the kernel function, so the researcher can build expert knowledge by analyzing the kernel function. Thirdly an SVM is defined by a convex optimization problems (no local minima) for which there are more efficient methods (e.g. SMO). Lastly, there is substantial evidence to the fact that it is an approximation to a bound on the test error rate that suggests that it would be a good idea to use the same. The advantage of SVM is that it is possible to train a non-linear, generalizable set with a small training data set. It exhibits robust performance even under noisy conditions for multiple biological analysis data. The disadvantages of SVM are the computational complexity involved in the training, selection of the kernel function and other parameters [9],[10]. The disadvantages are that the theory only really covers the determination of the parameters for a given value of the regularization and kernel parameters and choice of kernel. SVM moves the problem of over-fitting from optimizing the parameters to model selection.

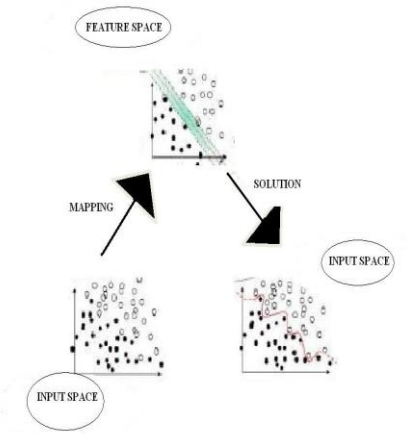


Figure 2.Architecture of Support Vector Machine

If the number of features is large, it is not needed to map the data to a higher dimensional space. Non linear mapping does not improve the performance. Many microarray data in bioinformatics are of this type. Hence it is sufficient to use a linear kernel SVM for this purpose. The general working of SVM is shown in Fig 3 below:

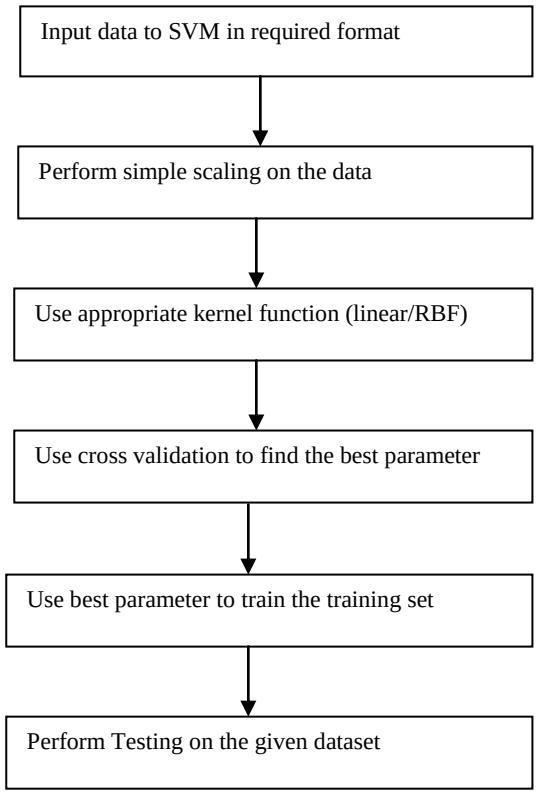


Figure 3. Working of Support Vector Machine

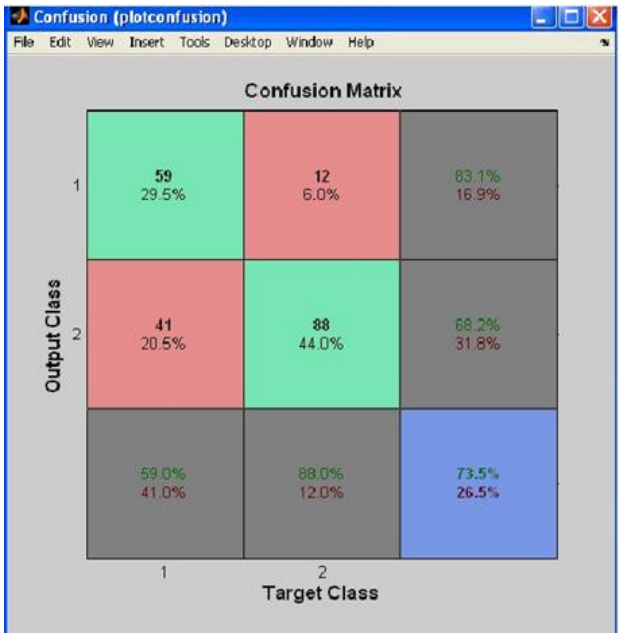
In this paper, three classification methods are analyzed. The first classifier used is the linear classifier. The second is the proximal SVM and the third is a Newton SVM. The linear classifier is a standard classifier that uses a linear kernel function. The Proximal SVM and Newton SVM differ from the linear classifier in the sense that the linear classifier takes a long time to converge. It also consumes a lot of execution time and the number of iterations is also huge. The Proximal

SVM and Newton SVM are well suited for microarray gene expression data. They are based on the method of regularized least square error. The Newton SVM converges in a maximum of 7 to 8 iterations.

3.1 Linear SVM

Linear SVM is the newest extremely fast machine learning algorithm that is used to solve multiclass classification problems from very large data sets. It implements an original proprietary version of a cutting plane algorithm for designing a linear support vector machine. Linear SVM is categorized as a linearly scalable routine because it creates an SVM model in a CPU time that scales linearly with the size of the training data set. A linear SVM is a machine learning algorithm that is most suitable for solving multiclass classification problems. It does not require high computing resources. For classification with a large number of features, as in the case of microarray data, linear kernel SVMs are said to outperform the complicated forms. It is linear in the sense that it creates an SVM model in a CPU time that is linearly scalable with respect to the size of the training dataset.

The genes selected from the proposed two step Feature Extraction are fed as the input to the Linear SVM. In solving a supervised classification task one uses a set of input-output training data pairs to design a decision function. The linear SVM takes the Adenoma and Carcinoma dataset as input. 50% of the genes are used for training and the rest is used for testing. As the given problem is to identify the cancerous and non-cancerous samples, the decision function is [0 1] for the cancerous samples and [1 0] for the non-cancerous samples. The accuracy of the classifier is shown in figure 4 and figure 5 below:



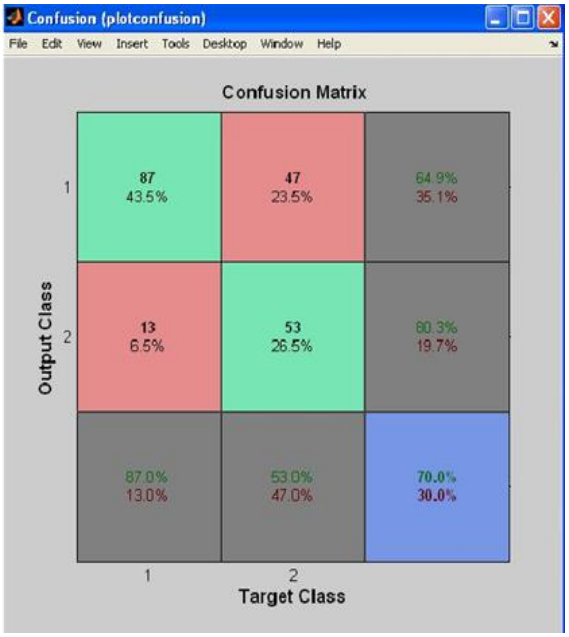


Figure 4. Linear SVM on Adenoma(Proposed Method) **Figure 5. Linear SVM on Carcinoma(Proposed Method)**

The Linear SVM showcases an accuracy of 73.5% for adenoma dataset and 70% for carcinoma dataset. This accuracy is obtained by using the proposed two step Feature Extraction Method.

3.2 Proximal SVM

A proximal support vector machine (PSVM), is used to solve simple nonsingular system of linear equations, for either a linear or nonlinear classifier. In contrast, standard support vector machine classifier requires a more costly solution of a linear or quadratic program to solve a non singular system of linear equations. For a linear classifier with millions of data points to classify, all that is needed by PSVM is the inversion of a small matrix of the order of the input space dimension typically of the order of 100 or less. For a nonlinear classifier, a linear system consists of equations of the order of the number of data points needed to be solved by the classifier. This allows the researcher to easily classify datasets with as many as a few thousands of points. Computational results on publicly available datasets indicate that the proposed proximal SVM classifier has comparable test set correctness to that of standard SVM classifiers. The main distinction between them is that the computational time is considerably faster by a certain magnitude in the case of proximal SVM.

The genes selected from the proposed two step Feature Extraction are fed as the input to the Proximal SVM. A standard support vector machine performs classification by assigning the given input sequence to one of the two disjoint half spaces and the points are classified by assigning them to the closer of the two planes in feature space. Also a standard SVM consumes a large amount of computational time to solve a linear or quadratic equation. In Proximal Support Vector Machine (PSVM), also termed as regularized least squares is a simple and efficient algorithm to perform classification on larger datasets [10]. The proximal SVM takes a matrix A as the input, the variable d which takes one of the two values 1 or -1. The positive sign indicates cancerous genes and a negative sign indicates non-cancerous genes. The variable "k" represents the folding factor. The variable "nu" is the weighting factor that takes as input any value as -1, 0 or any other number. The value -1 indicates easy estimation, 0 indicates hard estimation. The default value of "nu" is 0. The mandatory variables in this algorithm are A, d and k. [11],[12]

In order to validate the proposed two step feature extraction method, the first experiment is carried out on all the 14172 genes of adenoma dataset as shown in Fig 6 and 14914 genes of carcinoma dataset as shown in Fig 7. The entire gene

dataset is fed into the Proximal SVM classifier and the cross validation is performed for different values of k namely 1,3,5,8 and 10. In the second experiment, the absolute scoring method of ranking genes is applied for the same values of k as in the previous experiment. In the third experiment, the proposed t-test statistics combined with the absolute scoring method is applied for the same values of k where "k" is the cross validation parameter. The results are marked as c1 and c2. It is clearly evident from the figures below that the training and testing accuracy of the classifier is higher in the proposed two step process when compared to the absolute scoring method or the usage of the entire dataset.

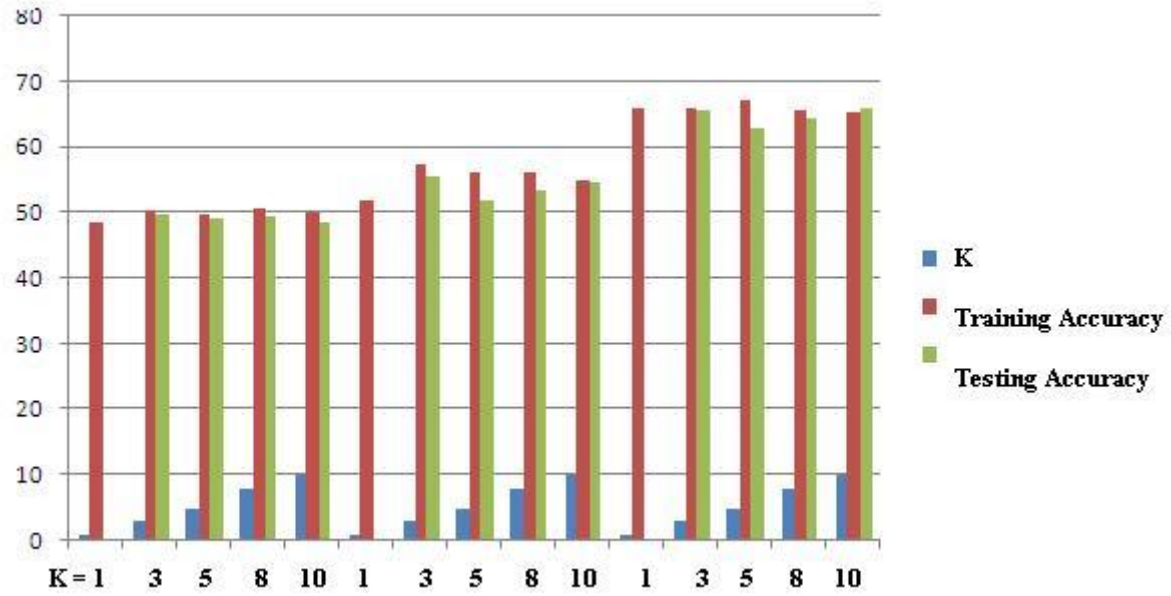


Figure 6. PSVM – Adenoma

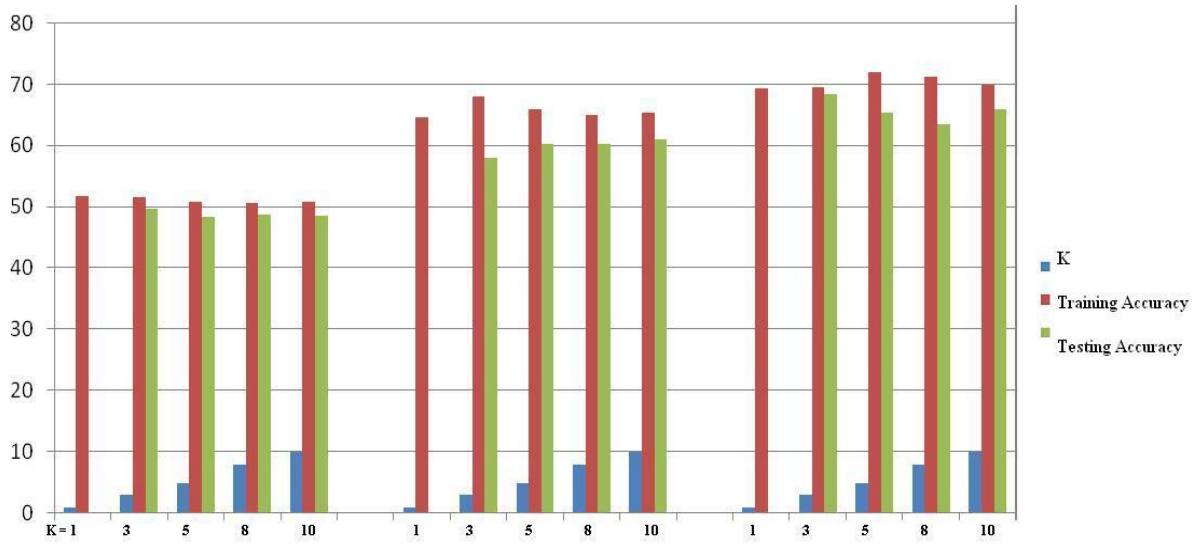


Figure 7. PSVM – Carcinoma

3.3 Newton SVM

The genes selected from the proposed two step Feature Extraction are fed as the input to the Proximal SVM. A standard support vector machine performs classification by assigning the given input sequence to one of the two disjoint halfspaces and the points are classified by assigning them to the closer of the two planes in feature space. Also a standard SVM consumes a large amount of computational time to solve a linear or quadratic equation. In Newton Support Vector Machine (NSVM), is a simple and efficient algorithm to perform classification on larger datasets [13][14][15]. This algorithm converges in a maximum of 6 to 7 iterations. The Newton SVM takes a matrix A as the input, the variable d that takes one of the two values 1 or -1. The positive sign indicates cancerous genes and a negative sign indicates non-cancerous genes. The variable „k" represents the folding factor. The variable „nu" is the weighting factor that takes as input any value as -1, 0 or any other number. The value -1 indicates easy estimation, 0 indicates hard estimation. The default value of „nu" is 0. The mandatory variables in this algorithm are A, d and k.

In order to validate the proposed two step feature extraction method, the first experiment is carried out on all the 14172 genes of adenoma dataset as shown in Fig 8 and 14914 genes

of carcinoma dataset as shown in Fig 9. The entire gene dataset is fed into the Newton SVM classifier and the cross validation is performed for different values of k namely 1,3,5,8 and 10. In the second experiment, the absolute scoring method of ranking genes is applied for the same values of k as in the previous experiment. In the third experiment, the proposed t-test statistics combined with the absolute scoring method is applied for the same values of k where „k" is the cross validation parameter. The results are marked as f1 and f2. It is clearly evident from the figures below that the training and testing accuracy of the classifier is higher in the proposed two step process when compared to the absolute scoring method or the usage of the entire dataset.

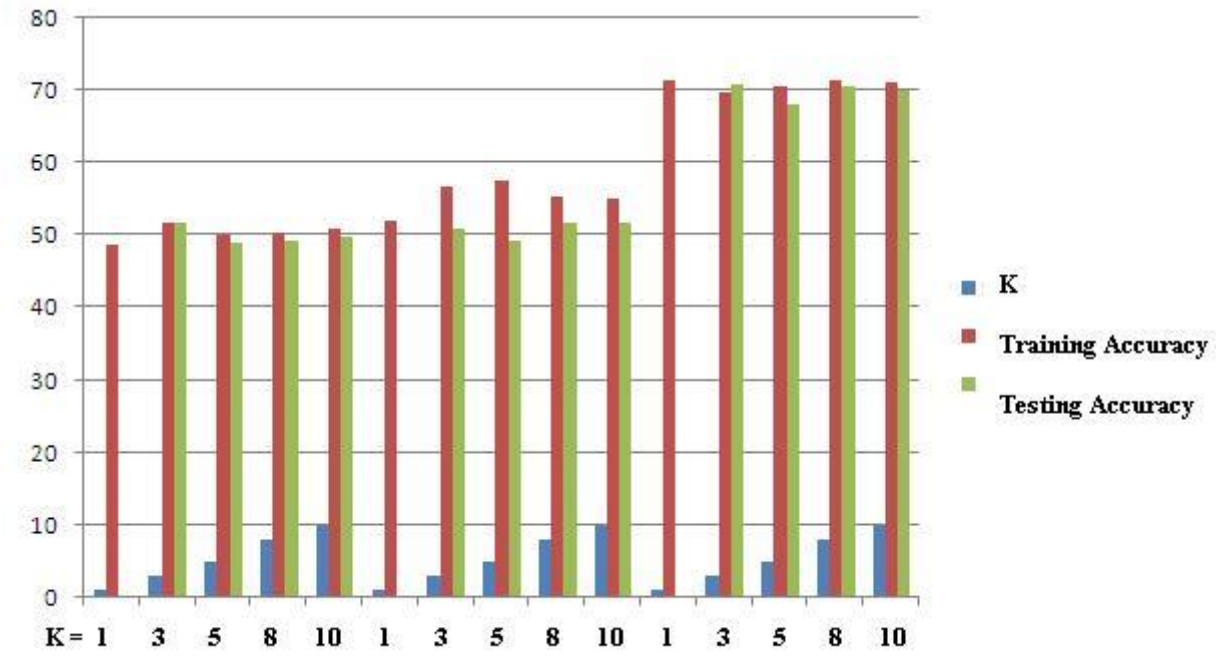


Figure 8. NSVM – Adenoma

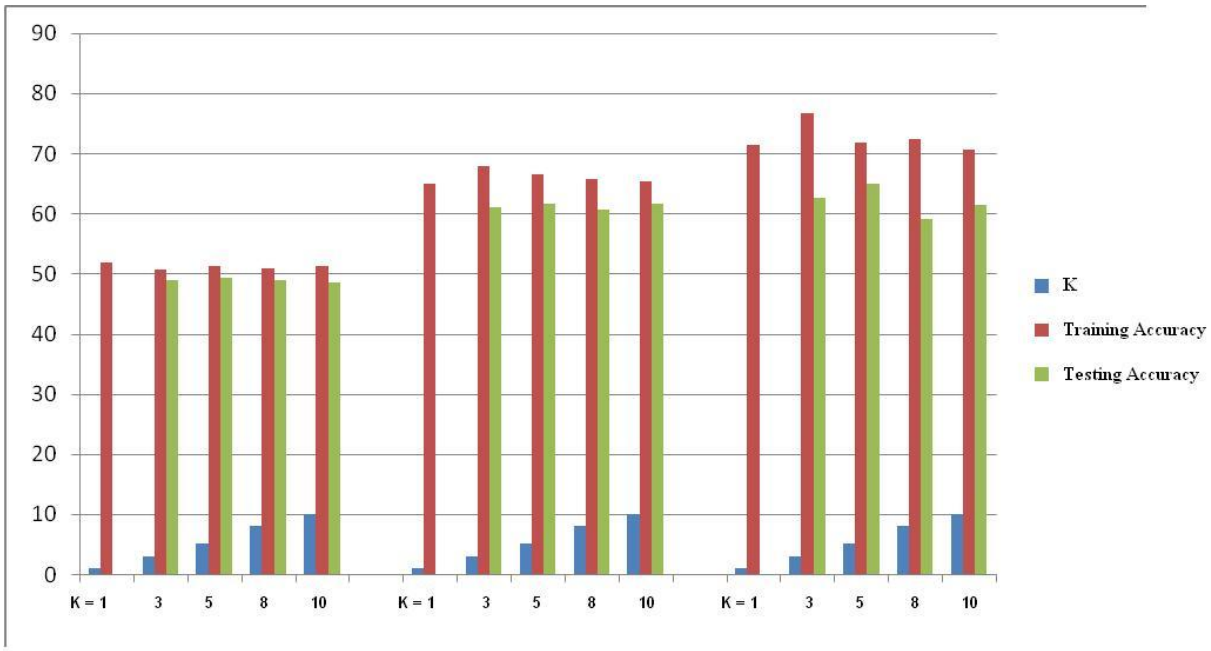


Figure 9. NSVM – Carcinoma

4. RESULTS AND DISCUSSION

The Linear SVM showcases an accuracy of 73.5% for adenoma dataset and 70% for carcinoma dataset. This accuracy is obtained by using the proposed two step Feature Extraction Method.

The Proximal SVM technique is applied on Adenoma and Carcinoma dataset after applying the proposed two step feature extraction method. The Newton SVM technique is

also applied on Adenoma and Carcinoma dataset after the application of the proposed two step feature extraction process. The results are clearly depicted by means of bar charts above. The result of applying Proximal SVM and Newton SVM on Adenoma dataset is tabulated as under:

Table 4. Results of PSVM on Adenoma Dataset – Accuracy – Training and Testing

	Entire Dataset					Absolute Scoring					Proposed Method				
K	1	3	5	8	10	1	3	5	8	10	1	3	5	8	10
Training Accuracy(%)	48.6	50.5	49.7	50.6	50.1	52.0	57.3	56.3	56.1	55.0	66.0	66.0	67.1	65.6	65.4
Testing Accuracy(%)	0.0	49.9	49.2	49.6	48.7	0.0	55.7	52.0	53.4	54.7	0.0	65.5	63.0	64.5	66.0

Table 5. Results of NSVM on Adenoma Dataset – Accuracy – Training and Testing

	Entire Dataset					Absolute Scoring					Proposed Method				
K	1	3	5	8	10	1	3	5	8	10	1	3	5	8	10
Training Accuracy(%)	48.7	51.9	50.0	50.3	51.0	52.0	56.7	57.5	55.3	55.0	71.5	69.7	70.6	71.3	71.1
Testing Accuracy(%)	0.0	51.7	49.1	49.2	49.8	0.0	51.0	49.3	51.9	51.7	0.0	71.0	68.0	70.5	70.0

It is clearly evident from the above Results Table 4 and 5 that the training and testing accuracy is higher when the proposed method is applied than applying the absolute scoring method or by applying the proximal SVM or Newton SVM on the entire dataset.

5. CONCLUSION AND CHALLENGES

The sample dataset from the Princeton University genome project is used for the above study. The results clearly depict that the proposed two step feature extraction method provides higher training and testing accuracy than the usage of the complete dataset and the absolute scoring of the genes when fed into the linear classifier, the proximal and Newton SVM classifier. The linear SVM consumes lot of time and performs more number of iterations if the dataset is very large. On the other hand, the proximal SVM and Newton SVM consume significantly lesser time compared to the linear SVM.

All the techniques have their own advantages and disadvantages. Partitive techniques that include supervised clustering and k-means clustering are not powerful for high dimensional and the nature of the data. Supervised techniques produce better results than non-supervised techniques because the knowledge of the training data set is available. The undeterministic character of the several clustering algorithms also makes them unreliable. The main challenge is to find the distance/proximity measure. Gene expression data contains a lot of clusters that are highly connected. The algorithms should be capable of handling these situations. The algorithms should also be able to operate under a noisy environment as most of the gene expression data that is captured would contain noise[16].

6. REFERENCES

[1] J.P.Florido, H.Pomares, I.Rojas, J.M.Urquiza, L.J.Herrera, M.G.Claros, "Effect of Pre-processing methods on Microarray-based SVM classifiers in Affymetrix Genechips", International Joint Conference on Neural Networks(IJCNN), pp 1-6, 2010

[2] Wei Du, Yan Wang,De-Ping Wang, Zhong-Bo Cao, Ying Sun and Yan-Chun Liang, "An Improved Normalized Signal to Noise Ratio Method for Irrelevant Genes Removing", 3rd International Conference on Biomedical Engineering and Informatics (BMEI 2010), pp 2275-2279, 2010

[3] Azadeh Mohammadi, Mohammad Hossein Saraee, "Dealing with Missing Values in Microarray Data, International Conference on Emerging Technologies" ,IEEE-ICET 2008,Rawalpindi, Pakistan, pp 258-263, 18-19 October, 2008

[4] Nicholas A. Furlotte, Lijing Xu,Robert W.Williams, Ramin Homayouni, "Literature-based Evaluation of Microarray Normalization", IEEE International Conference on Bioinformatics and Biomedicine, pp 608-612, 2011

[5] http://genomics-pubs.princeton.edu/oncology/

[6] Jinn-Yi Yeh, Tai-Shi Wu, Min-Che Wu, Der-Ming Chang, "Applying Data Mining Techniques for Cancer Classification from Gene Expression Data", International Conference on Convergence Information Technology, IEEE Computer Society,pp 703-708, 2007

[7] Huang, D.; Chow, T.W.S.; Ma, E.W.M.; Jinyan Li, Efficient selection of discriminative genes from microarray gene expression data for cancer diagnosis, IEEE Transactions on Circuits and Systems, Volume 52, Issue 9, pp 1909-1918, 2005

[8] Seeja.K.R, Shweta, "Microarray Data Classification Using Support Vector Machine", International Journal of

Biometrics and Bioinformatics (IJBB), Volume (5): Issue (1): 2011

[9] Tang, Yuchun; Zhang, Yan-Qing; Huang, Zhen, Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 4, Issue 3, pp 365-381, 2007

[10] http://bioinfo.mbb.yale.edu/mbb452a/intro

[11] Ghorai, S.; Mukherjee, A.; Sengupta, S.; Dutta, P.K., Cancer Classification from Gene Expression Data by NPPC Ensemble, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 8, Issue 3, pp 659-671, 2011

[12] Guang-bin Huang; Hongming Zhou; Xiaojian Ding; Rui Zhang, Extreme Learning Machine for Regression and Multiclass Classification, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol 42, Issue 2, pp 513-529, 2012

[13] Glenn Fung & O. L. Mangasarian, "Finite Newton Method for Lagrangian Support Vector Machine Classification"

[14] Xiong Fu-song, NLSSVM: "Least Square Support Vector Machine based on Newton optimization", IEEE International Conference on Computer Science and Automation Engineering (CSAE), pp 140-144, 2011

[15] Ruopeng Wang; Hongmin Xu; Hong Shi, "Newton's Method for L∞ Support Vector Machine Via Smoothing technique", Sixth International Conference on Natural Computation, Vol 1, pp 436-440, 2010

[16] Mitra, S.; Das, R.; Hayashi, Y., "Genetic Networks and Soft Computing", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 8, Issue 1, pp 94-107, 2011

[17] Alireza Osareh,Bita Shadgar, "Microarray Data Analysis for Cancer Classification", 5th International Symposium on Health Informatics and Bioinformatics, Turkey, pp 125-132, April 20-22, 2010

[18] Jinn-Yi Yeh,Tai-Shi Wu,Min-Che Wu,Der-Ming Chang, "Applying Data Mining Techniques for Cancer Classification from Gene Expression Data", International Conference on Convergence Information Technology, IEEE Computer Society, pp 703-708, 2007

[19] Chen Liao,Shutao Li,Zhiyuan Luo, "Gene Selection for Cancer Classification using Wilcoxon Rank Sum Test and Support Vector Machine", International Conference on Computational Intelligence, pp 368-373, 2006

[20] Yuchun Tang, Yan-Qing Zhang, and Zhen Huang, "Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 4, NO. 3, Published by the IEEE CS, CI, and EMB Societies & the ACM, pp 365-381, July-September 2007

[21] Patharawut Saengsiri,Sageemas Na Wichian,Phayung Meesad,Unger Herwig, "Comparison of Hybrid Feature Selection Models on Gene Expression Data", Eighth International Conference on ICT and Knowledge Engineering, pp 13-18, 2010

[22] Wai-Ho Au, Keith C.C. Chan, Andrew K.C. Wong, and Yang Wang, "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 2, NO. 2, pp 83-101, April-June 2005

[23] John Phan, Richard Moffitt, Jennifer Dale, John Petros, Andrew Young, and May Wang, "Improvement of SVM Algorithm for Microarray Analysis Using Intelligent Parameter Selection", Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China, pp 4838-4841, September 1-4, 2005

[24] Osman Abul, Reda Alhajj, and Faruk Polat, "A Powerful Approach for Effective Finding of Significantly Differentially Expressed Genes", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 3, NO. 3, pp 220-231, July-September 2006

[25] Shutao Li, Chen Liao, James T. Kwok, "Wavelet-Based Feature Extraction for Microarray Data Classification", International Joint Conference on Neural Networks,Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada , pp 5028-5033, July 16-21, 2006

[26] Taysir Hassan A. Soliman,Adel A. Sewissy,Hisham AbdelLatif, "A Gene Selection Approach for Classifying Diseases Based on Microarray Datasets", 2nd International Conference on Computer Technology and Development (lCCTD 2010), pp 626-631, 2010

[27] http://bmbolstad.com/talks/Bolstad%20GenentechTalk.pdf

[28] http://www.bioinformatics.wsu.edu/bioinfo_course/notes/Lecture16.pdf

[29] T.R. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science, vol. 286, pp. 531-537, 1999.

[30] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," Artif. Intell., vol. 97, pp. 245–271, 1997.