Performance of Feed Forward Neural Network for a Novel Feature Selection Approach

Barnali Sahu#, Debahuti Mishra*

Department of computer science and Engineering, Trident Academy and Technology, Bijupattnaik University
Bhubaneswar, Odisha, India

* *Department of Computer Science and Engineering, Institute of Technical Education and Research, Siksha O Anusandhan University, Bhubaneswar, Odisha, India*

Abstract— **Feature selection for classification of cancer data is to discover gene expression profiles of diseased and healthy tissues and use the knowledge to predict the health state of new sample. It is usually impractical to go through all the details of the features before picking up the right features. The differentially expressed genes or biomarker gene selection is the pre-processing task for cancer classification. In this paper, we have compared the results of two approaches for selecting biomarkers from Leukaemia data set for feed forward neural networks. The first approach for feature selection is by implementing k-means clustering and signal-to-noise ratio (SNR) method for gene ranking, the top scored genes from each cluster is selected and given to the classifiers. The second approach uses signal to noise ratio ranking only for feature selection. For validation of both the approaches we have used Holdout validation and compared the results.**

Keywords— ***Differentially Expressed Genes, Feature Selection, K-means, Signal to Noise ratio, Feed forward neural network***

I. INTRODUCTION

Cancer diagnosis is one of the most important emerging clinical applications of gene expression microarray technology. With the development of genomic techniques, research on molecular biology has shifted from individual genes to the entire genomes. Microarray technology can measure the expression levels of thousands of genes in a single experiment. With a certain number of samples, investigations can be made into whether there are patterns or dissimilarities across samples of different types, such as cancerous versus normal or even within subtypes of disease. The problem is referred to as sample classification. In a microarray chip, the number of genes available is far greater than that of samples, however most genes in microarray give little benefits to the sample classification problem. Therefore, prior to sample classification, it is important to perform gene selection whereby more interpretable genes are identified as biomarkers, so that a more efficient, accurate and reliable performance in classification can be expected

This many high level data analysis techniques such as clustering and classification algorithms work well with small number of genes. This approach usually covers one or more components of microarray data analysis that include dimensionality reduction through a gene subset selection, the construction of new predictive features and model inference [1]. The gene expression microarray technology allows us to measure expressions of thousands of genes simultaneously in a single experiment. This technique presents gene expression data of an organism in different environment or different expression of a gene in different organism. Microarray data are generally high dimensional data having large number of genes in comparison to the number of samples or conditions. Hence, it suffers from a very well known problem of “curse of dimensionality”. Due to this problem it is very complex to analyse microarray data. There are many efficient methods for the analysis of microarray data such as clustering, classification and feature selection. Feature selection is the pre-processing task for classification. As classification does not work well with large numbers of features hence prior to sample classification feature (gene) selection is essential, where by more relevant and interpretable genes can be filtered. These relevant genes are known as discriminative genes or Biomarkers. By training the classifiers with the biomarkers we can achieve better classification accuracy with a low risk of misclassification. The benefits obtained from gene selection are not only to get better classification accuracy but also to decrease the cost in a clinical setting. It also enhances the interpretability of genetic nature of the disease for biologists [2], [3], [4], [5].

As microarray data are high dimensional data, there may be noise present in the data. With noisy data the performance and efficiency of the model may decrease. There are several feature selection methods available to resolve the problem and to increase the efficiency of the model [6]. The well known feature selection methods are: filter and wrapper method. Filter method rank the features according to their discriminative power with regard to the class labels of samples where as wrapper approach selects a subset of features from the original feature set with respect to a classifier. Filter methods such as signal-to-noise ratio [7], t-Statistics [8], F-test [9] have been shown to be effective scores for measuring discriminative power of features in microarray data. In all cases genes are ranked according to their statistical scores and a certain number of highest ranking genes are selected for the purpose of classification.

A. Goal of the Paper

The goal of this paper is to find differentially expressed genes by applying clustering technique to group similar genes before implementing filtering techniques to filter relevant gene subset and to enhance the accuracy of the filtering technique. We have adopted two different approaches for relevant gene selection. In first approach we have used k-means clustering technique for grouping the features in the data set, as genes in a cluster are more correlated with each other with respect to genes present in different clusters. After that we have implemented different filtering technique to rank the genes in each cluster. The best scored features in each cluster are then selected. After that the data with these features are tested using feed forward Neural Network classifiers, and the performance is compared in two approaches.

B. **Paper Lay out**

The rest of the paper is organized as follows: in the section II shows the related work on discovering differentially expressed genes, section III describes our proposed work, section IV gives a brief introduction to gene expression data, k-means clustering, signal-to-noise ration and feed forward neural network classification model Holdout validation technique. Section V gives experimental validation and comparison among both the approaches and also the result analysis. Finally, section VI gives the conclusion and future work.

II. RELATED WORK

Supoj Hengpraprohm et.al. [10] Proposed a method for selecting informative features (genes) were using k-means clustering and SNR ranking. Genetic programming is used as a classifier. They have used eight datasets such as leukaemia, breast cancer, CNS, colon cancer, ovarian cancer, prostate cancer, lung cancer and lymphoma. They have compared the experimental results with many feature selections and classifiers among them only kNN using Pearson's coefficients correlation and information gain as feature selection showed the better result than the proposed approach. Wai-Ho Au et.al.[11]proposed a clustering algorithm known as k-modes Attribute Clustering Algorithm (ACA). It follows the idea of k-means clustering algorithm. The authors have taken colon cancer and leukaemia data set for their experiment. The proposed method groups interdependent attributes into clusters by optimizing a criterion function. The experimental result of ACA is compared with those of t-test, k-means algorithm, Kohonen's SOM, biclustering algorithm, MRMR algorithm, and RBF algorithm. Hualong Yu et.al. [12] Proposed a marker gene selection approach. The authors have selected top ranked informative genes by applying SNR score. After that PSO was applied to select a few marker genes and SVM for evaluation. Yukee Leung et.al. [13] Proposed a multi filter-multi wrapper approach for selecting informative genes or biomarkers. Multiple filters are SNR method, Pearson correlation and t-test. For wrapper they have used SVM, weighted Voting, 3NN as classifiers. The proposed mode was evaluated by six DNA microarray data sets such as LEU, COL62, BR-ER49, LYM77, PROS102 and UNG181. Shamsul Huda et.al. [14] Proposed a hybrid wrapper and filter feature selection algorithm by introducing filters feature ranking score in wrapper stage to get a more compact feature set. They have hybridized mutual information based maximum relevance filter ranking method with artificial neural network based wrapper approach to get the accuracy. Chenn-Jung Huang et.al.[15] the authors have under gone a comprehensive study on the capability of probabilistic neural network associated with SNR scoring method for cancer classification. The experimental results shows that the combination of the probabilistic neural network with the signal-to-noise method can achieve better classification results for two types of acute leukaemia and five categories of embryonic tumours of central nervous system with satisfactory computation speed. Piyushkumar A. Mundra et. al.[16] have decomposed the t-statistics in to two parts, corresponding to relevant and irrelevant data points. The relevant data points were selected using SVM and then t-statistic was used for feature selection. Jooyong shim et.al. [17] Proposed an algorithm for selecting marker genes. The algorithm was based on support vector machine and supervised weighted kernel clustering (SWKC/SVM). They have used a simulated data set and 6 real data sets and compared the method with three existing methods (PAM, SVM-REF, and SPM) and the result of (SWKC/SVM) method was having a lower mean error than the existing methods.

III. PROPOSED MODEL

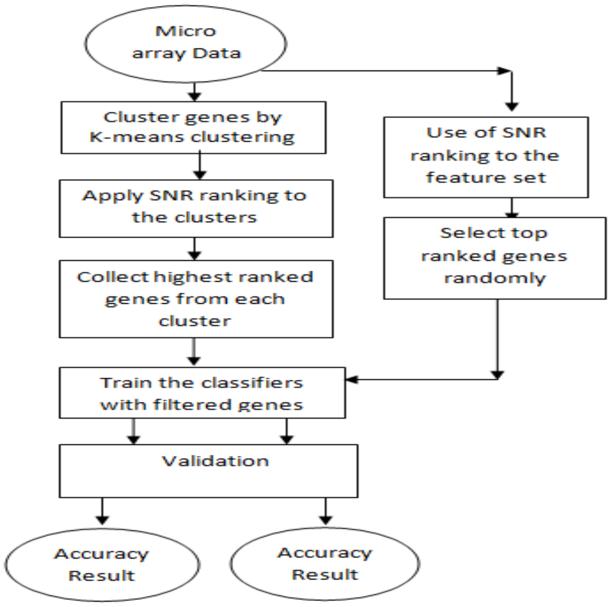

Micro array Data
Cluster genes by K-means clustering
Apply SNR ranking to the clusters
Collect highest ranked genes from each cluster
Use of SNR ranking to the feature set
Select top ranked genes randomly
Train the classifiers with filtered genes
Validation
Accuracy Result
Accuracy Result

Fig. 1 Model for feature (gene) selection Approach

IV. PRILIMINARIES

All the methods and technologies used for feature selection, classification and validations are described below.

A. Gene expression data

A Microarray data set can be represented as an expression table. Where, each row corresponds to a particular gene and each column to a sample. $E = \{X_{ij} \,|i=1\ldots\ldots k, j=1\ldots n\}$ where $X_{ij} \in R$ is the expression level of gene g_i sample S_i [10].

B. Signal-to-noise ratio feature selection

The signal-to-noise ratio (SNR) test identifies the expression patterns with a maximal difference in mean expression between two groups and minimal variation of expression within each group [18]. In this method genes are first ranked according to their expression levels using SNR test Statistic. The SNR is defined as follows

$$\text{signal-to-noise ratio} = \left|\frac{\mu_I - \mu_U}{\sigma_I + \sigma_U}\right| \qquad (1)$$

Where, μ_I and $\mu_U$denote the mean expression values for the sample class I and class U respectively.σ_I, σ_U are the standard deviation for the samples in each class.

C. K-means clustering

lustering algorithm partitions the tuples of large data sets into groups based on their similarity. Hence tuples in a cluster are more similar to each other than those belongs to different cluster. In our approach we have used k-means clustering algorithm due to its simplicity and faster execution capability.

K-means clustering Algorithm:

```
Input:  k   is the Number of clusters
        P is the data set containing n features (n number of
            genes)

1.   Select number of cluster k.
2.   Randomly choose k features from the data set as the
     initial cluster centre.
3.   Repeat until the termination criteria fulfilled
     3.1   Assign each feature to one of the clusters
           according to the similarity measure
     3.2   Update the cluster means.
4.   until no change in the value of cluster's mean
```

In this approach we have used Euclidean distance as distance measure. The reasons for the popularity of *k*-means are ease and simplicity of implementation, scalability, speed of convergence and adaptability to sparse data.

D. Feed forward neural network

A Neural network implements a non-linear function f(x, w) where f is the output of the function for input x and network parameters w. Given a training set, i.e., set of pairs of the form$\langle x_i, y_i \rangle$, i =1...N the neural network can be trained to model the data as closely as possible. The behaviour of the neural network depends on the interaction between the neurons. The architecture of neural network consists of 3 types of neuron layers 1, input layer 2 and hidden 3 outputs layers. In feed forward neural network the signal flow is from input to output units in a feed forward direction [19]. In particular, for modelling gene expression data it is necessary to model the correlations between weight vector components because genes act in concert with a collection of other genes forming gene networks.

E. *Holdout validation method*

To avoid over-fitting, an independent test set is preferred. A natural approach is to split the available data into two non-overlapped parts: one for training and the other for testing. The test data is held out and not looked at during training. Hold-out validation avoids the overlap between training data and test data, yielding a more accurate estimate for the generalization performance of the algorithm. The downside is that, this procedure does not use all the available data and the results are highly dependent on the choice for the training/test split.

V. EXPERIMENTAL EVALUATION

We have used leukaemia data set of cancer microarray data from Biological data analysis web site [20]. The leukaemia dataset consists of 72 Microarray experiments (samples) with 7129 gene expression levels. The problem is to distinguish between two types of Leukaemia, Acute Myeloid Leukaemia (AML) and Acute Lymphoblastic Leukaemia (ALL). The complete data set contains 25 AML samples and 47 ALL samples. As in other experiments 38 out of 72 samples used as training data (27 ALL samples and 11 AML samples) and the training samples (20 ALL samples and 14 AML samples) are used as test data. We have taken 50 genes and 72 samples (47 class1, 25 class2) of original data set. The experiment is done in MATLAB version 7.6.0.324 (R2008a), windows XP, PC of Intel Pentium dual CPU. We have implemented two different approaches of feature selection used for classification model to discover differentially expressed genes.

A. *Algorithm for Approach I for classification model to discover differentially expressed genes using k-means and SNR*

1. Microarray data set with
   Dom (C) = *{I, U}*
   C is the random variable for class label.

   1.1. I= $[x_{ij}]$; *i* represents genes and j∈ *(1, M)* samples.
   U= $[x_{ij}]$; *i* represents genes and j∈ *(1, N)* samples.
2. Each gene *i* of the data set is clustered using k-means algorithm, where each *i* is associated with a cluster number.
3. For each i∈ S_1^n ,
   SNR (i) = $\left|\frac{\mu_I-\mu_U}{\sigma_I+\sigma_U}\right|$ is calculated.
   Where S represents clusters from 1 to n, n is total number of clusters
4. Top scored *SNR (i)* is collected from each clusters and the significant features will act as an input to the classifier. Training set T of n tuples is formed.
5. The feed forward neural network classifier is trained with T and tested with test set with holdout validation method and accuracy is measured.

For Approach I we have taken 5, 10 and 20 clusters and for Approach II 5, 10 and 20 top scored genes are selected randomly. These significant features (genes) will act as input to the feed forward neural network and the performance of the classifier is measured and compared. Training stops when any of these conditions occurs:

* The maximum number of epochs (repetitions) is reached.
* The maximum amount of time is exceeded.
* Performance is minimized to the goal.

TABLE I
CLASSIFICATION ACCURACY OF FFNN USING APPROACH I

No of clusters	No of genes selected	Performance of Feed forward Neural Network
5	5	2.52e-28
10	10	4.28e-28
20	20	3.89e-23

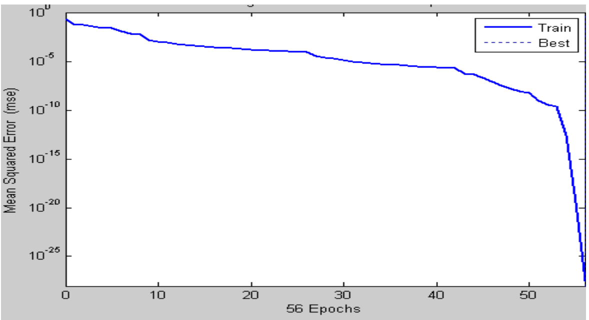


Fig. 2 performance plot of FFNN for 5 numbers of genes

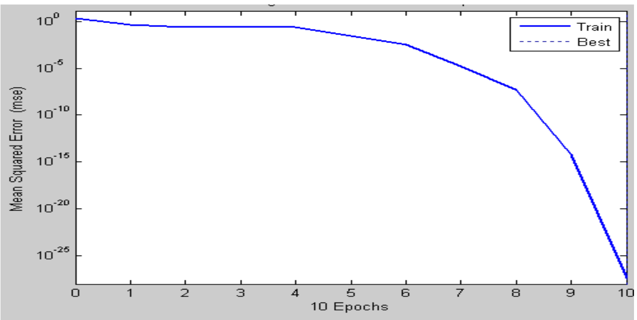


Fig. 3 Performance plot of FFNN for 10 number of genes

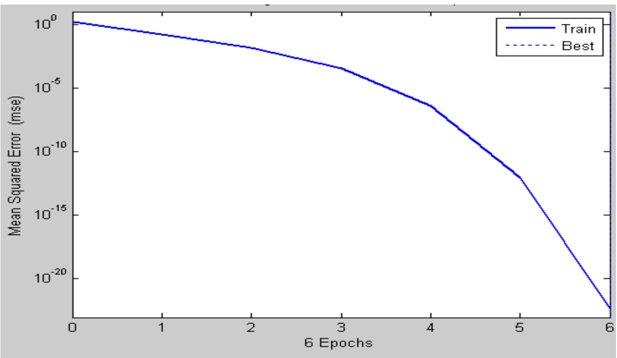


Fig. 4 Performance Plot of FFNN for 20 numbers of genes

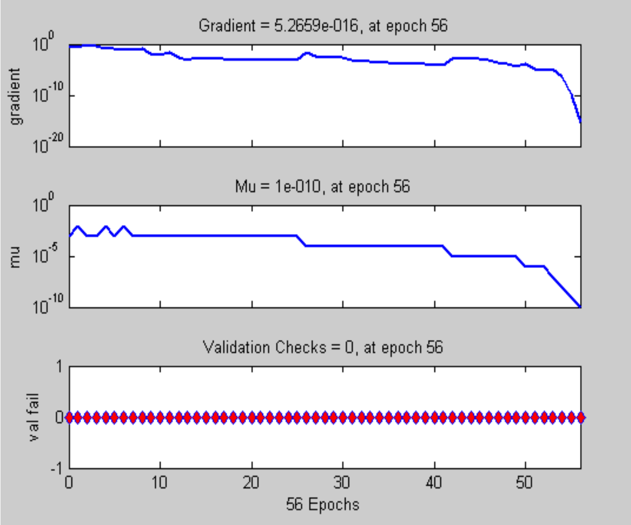


Fig. 5 Plot for training state of FFNN for 5 number of genes

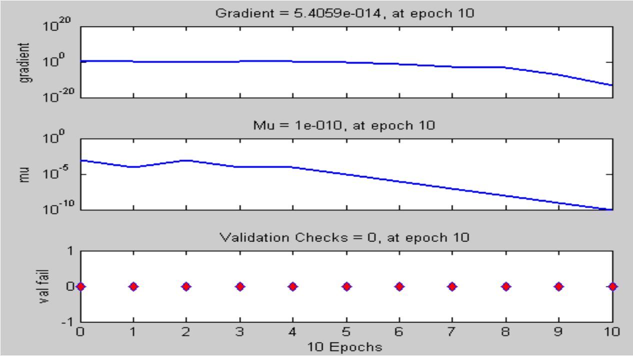


Fig. 6 Plot for training state of FFNN for 10 number of genes

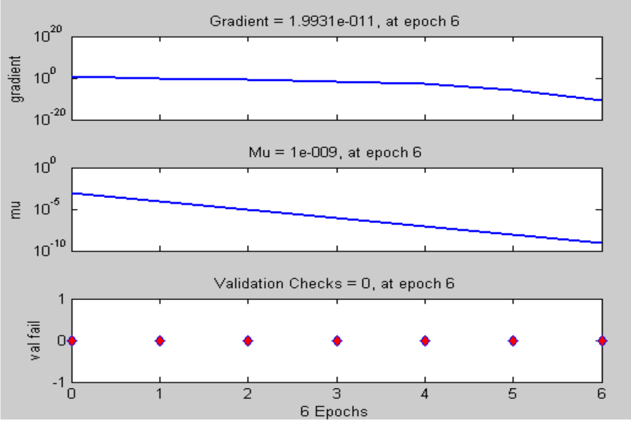


Fig. 7 Plot for training state of FFNN for 20 number of genes

B. *Algorithm for Approach II for classification model to discover differentially expressed genes without using k-means*

1. Microarray data set D with
   Dom (C) = *{I, U}*
   C is the random variable for class label.

   1.1 *I= [x_{ij}]; i* represents genes and *j*∈ *(1, M)* samples.
   U= [x_{ij}]; i represents genes and *j*∈ *(1, N)* samples.

2. For each *i*∈ *D* ,
   SNR (i) = $\left|\frac{\mu_I-\mu_U}{\sigma_I+\sigma_U}\right|$ is calculated.
3. Random number of Top scored *SNR (i)* is collected and training set *T* of *n* tuples are formed.
4. The feed forward neural network classifier is trained with T and tested with test set with holdout validation method and accuracy is measured.

TABLE 2
CLASSIFICATION ACCURACY OF FFNN IN APPROACH II

No. of genes selected	Performance of feed forward neural network (FNN)
5	8.88e-29
10	3.41e-34
20	1.30e-31

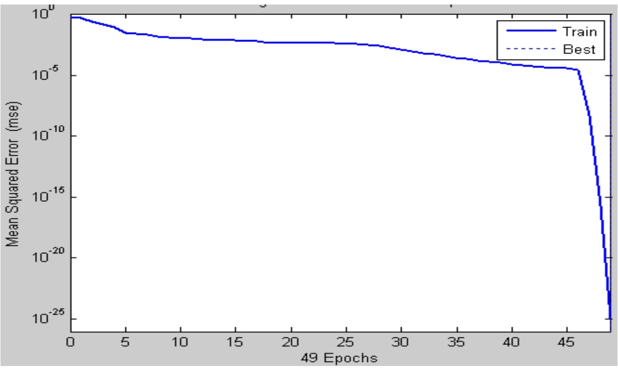


Fig. 8 Performance plot of FFNN for 5 number of genes

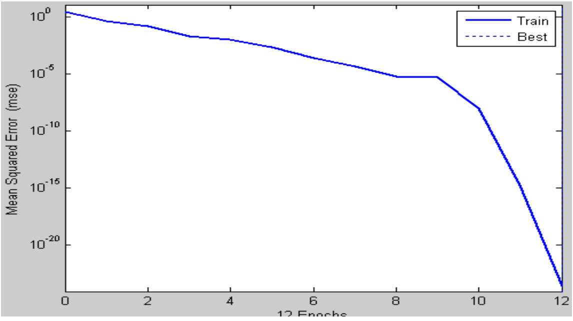


Fig. 9 Performance plot of FFNN for 10 number of genes

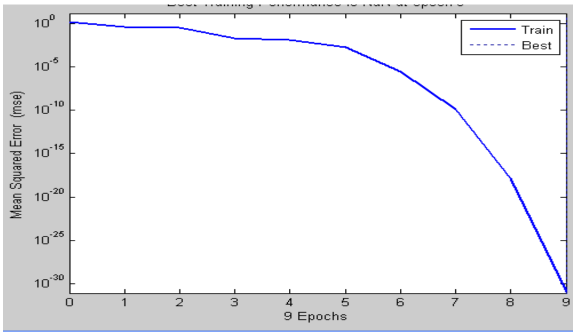


Fig. 10 Performance plot of FFNN for 20 number of genes

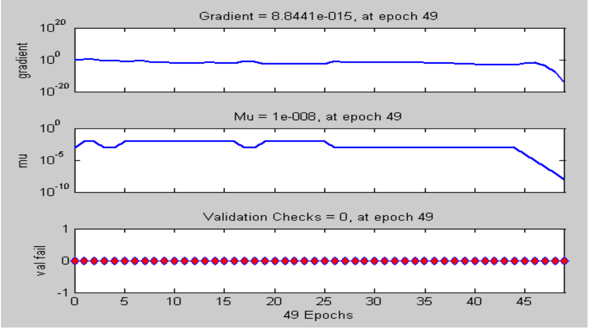


Fig. 11 Plot for training state of FFNN for 5genes

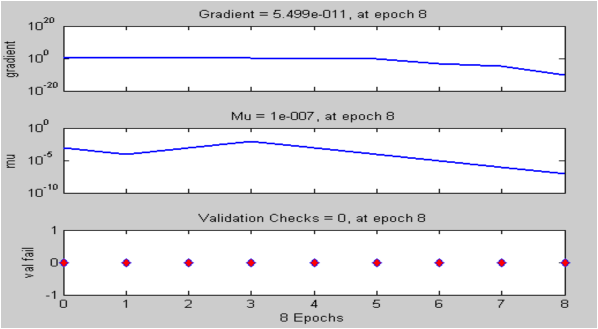
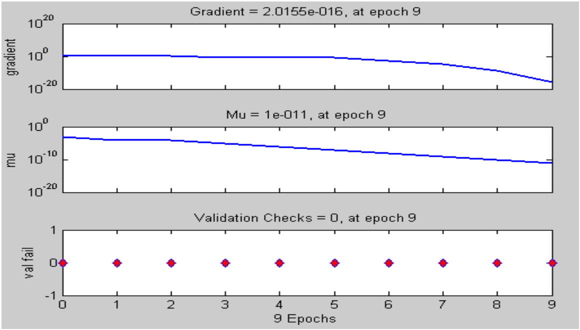


Fig. 12 Plot for training state of FFNN for 10 genes



. Fig. 12 Plot for training state of FFNN for 10 genes

Table I and table II listed the performance of feed forward neural network for first approach and second approach respectively. By observing the results of both the tables the accuracy or performance result of FFNN for first approach is better than the second one. Again FFNN perform better with 10 genes in comparison to 5 or 20 genes.

Figure 2, 3 and 4 show s the performance plot for 5, 10 and 20 numbers of genes in first approach. Figure 2 for 5 numbers of genes it takes 56 epochs to train the FFNN classifier, where as in figure 3 and 4 for 10 and 20 numbers of genes FFNN takes 10 and 6 epochs respectively. Here an epoch is the presentation of the entire training set to the neural network or an epoch is one sweep through all the records in the training set.

Back propagation is used to calculate derivatives of performance with respect to the weight and bias variables. Each variable is adjusted according to gradient descent with momentum, validation vector are used to stop training early if the network performance on the validation vectors fails to improve or remains the same for maximum epochs in a row. Test vectors are used as a further check that the network is generalizing well, but do not have any effect on training. Figure 5, 6 and 7 shows the training state plot for 5 and 20 numbers of genes in first approach respectively. The training state plots gives the gradient value, mu and validation checks for 56 epochs, Like wise figure 6 and 7 shows the gradient value, mu and validation check for 10 and 6 epochs.

Figure 8, 9 and 10 shows the performance plot for 5, 10 and 20 genes in second approach. Figure 11, 12 and 13 shows the training state of FFNN for 49, 8 and 9 epochs.

VI. Comparison and Conclusion

It is significant to note that after applying the first approach of feature selection the feed forward neural network perform well in comparison to the second approach. With 20 numbers of relevant genes selected by first approach is taken as input to the feed forward neural network to train the classifier and tested with a test set applying hold out validation method and the performance achieved is 3.89e-23. Whereas after applying the second approach of feature selection with 5, 10 and 20 numbers of relevant features the feed forward neural network is trained and tested and by applying holdout validation method the performance is measured and they are 8.88e-29, which is lower than the result in first approach.

Feature selection means choosing the feature (gene) subset in gene expression data analysis which enhances the prediction and classification accuracy of a model. There are several objectives for feature selection first to get relevant features; second relevant features may be redundant so removal of some redundant features may enhance the accuracy of the model. The best feature subset always contains minimum number of features (genes) which contribute towards the accuracy of the model.

In our paper [21] we have got 100% and 99.3% accuracy for SVM with holdout validation and 10 fold cross validation method in first approach respectively.

Our approach is to satisfy the second objective of feature selection approach i.e. to remove redundant features to avoid unnecessary complexity and enhance the performance of the FFNN classifier. Experimental result shows the better performance of FFNN with this approach than in general filtering technique.

REFERENCES

[1] Minca Mramor Gregor Leban, Janez Demšar, Blaž Zupan. *Visualization-based cancer microarray data classification analysis*. Bioinformatics, 2007,vol .23:2147-2154

[2] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "*Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", Science*,1999, Vol. 286, no. 5439, pp. 531-537.

[3] C.-A. Tsai,Y.-J. Chen, and J.J. Chen "*Testing for Differentially Expressed Genes with Microarray Data*," Nuclear Acids Research, Vol. 31, No.9, pp. 52, 2003.

[4] M.S. Pepe, G. Longton, G.L. Anderson, and M. Schummer, "*Selecting Differentially Expressed Genes from Microarray Experiments*," *Biometrics*, Vol. 59, No. 1, pp. 133-142, March 2003.

[5] P. Broberg, "Statistical Methods for Ranking Differentially Expressed Genes", *Genome Biology*, Vol. 4, No. 6, p. R41, 2003.

[6] I. Guyon and A. Elisseeff, " *An Introduction to Variable and Feature selection*," Journal of Machine Learning Rearch, vol. 3, p.1157-1182, 2003.

[7] Jirapech-Umpai, T. and S. Aitken, "*Feature selection and classification for microarray data analysis:Evolutionary methods for identifying predictive genes*," *BMC Bioinformatics*, Vol. 6, No. 148, 2005.

[8] T.P.Speed, "*Statistical analysis of Gene Expression Microarray Data"*. Chapman & Hall/CRC,2003

[9] C.ding and H.Peng, "*Minimum Redundancy Feature selection from microarray Gene Expression data",* proc.IEEE computer Soc. Bioinformatics Conf.(CSB '03), pp.523-528, 2003.

[10] Supoj Hengpraprohm , Prabhas Chongstitvatana, "Selecting Informative Genes from Microarray Data for Cancer Classification with Genetic Programming Classifier using K-Means Clustering and SNR Ranking", *Frontiers in the Convergence of Bioscience and Information Technologies* , pp211-216, 2007.

[11] Wai-Ho Au,Keith C.C.Chan,Andrew K.C. Wong, Yang Wang,2005. Attribute clustering for Grouping ,Selection and Classification of Gene Expression Data,IEEE/ACM Transactions on computational biology and Bioinformatics, vol 2.,No 2,pp83-101

[12] Hualong Yu,Guochang Gu,Haibo Liu,Jing Shen, Changming Zhu,. A Novel Discrete Particle Swarm Optimization Algorithm for Microarray Data-based Tumor Marker Gene Selection, *International Conference on Computer science and software Engineering,*pp1057-1060,2008

[13] Yukyee Leung, Yeungsam Hung, A Multi-Filter-Multi-Wrapper Approach to Gene Selection and Microarray Date Classification", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vo1. 7, No .1 , pp 108-117, 2010.

[14] Shamsul Huda, John Yearwood, Andrew Stranieri, "Hybrid wrapper-filter approach for input feature selection using Maximum Revalance and Artificial Neural Network Input Gain Measurement Approximation", *Fourth International conference on Network and system security*, pp442-449, 2010.

[15] Chenn-Jung Huang ,Wei-Chen Liao, "A Comparative Study of Feature Selection Methods for Probabilistic Neural Networks in Cancer Classification", *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence* (ICTAI'03),Vol 3, pp1082-3409, 2003.

[16] Piyushkumar A. Mundra and Jagath C. Rajapakse," Gene and sample selection for cancer classification with support vectors based t-statistic", *Neurocomputing* Vol 73 pp2353-2362, 2010

[17] Jooyong Shim,Insuk Sohn, Sujong Kim, jae Won Lee, Paul E. Green ,Changha Hwang, "Selecting marker genes for cancer classification using supervised weighted kernel clustering and the support vector machine", *Computational Statistics and* Data *Analysis,* Vol 53 pp1736-1742, 2009.

[18] Miroslava Cuperlovic-Culf, Nabil Belacel, Rodney J. Ouellette, "Determination of tumour marker genes from gene expression data", DDT, Vol-10, Number 6 pp429-437, 2005

[19] Aboul Ella Hassanien , Mariofanna G. Milanova, Tomasz G. Smolinski, Ajith Abraham,2008.Computational Intelligence in solving Bioinformatics problems:Reviews, perspectives, and challenges, Springer-Verlag Berlin Heidelberg SCI 151, pp. 3–47

[20] http://sdmc.lit.org.sg/GEDatasets/

[21] Mishra Deahuti, Sahu Barnali," Feature selection for cancer classification: A signal to noise Approach",International Journal of scientific and Engineering research ,2011,vol.2 issue 4 pp 1-7