

Research Article

A Weighted-SNR Feature Selection from Independent Component Subspace for NB Classification of Microarray Data

***Rabia Aziz, C.K. Verma and Namita Srivastava**

Department of Mathematics and Computer Application,
Maulana Azad National Institute of Technology Bhopal-462051 (M.P.) INDIA
*Corresponding Author: Email- rabia.aziz2010@gmail.com, Tel: +91-9479967401

[Received-28/04/2015, Published-27/05/2015]

ABSTRACT

One of the major challenges is a small sample size as compared to large features (genes) number of microarray data. Feature selection is an important process for accurate classification of microarray data. This paper proposed a new combination of feature selection/extraction approach for the Naive Bayes (NB) classification of high dimensional cancer microarray data, which uses extraction technique such as Independent Component Analysis (ICA) and filtering technique Signal to noise ratio (SNR). The proposed method is divided into two stages. In the first stage the original DNA microarray gene expression data are modeled by ICA, SNR score is used to rank the ICA feature vector. To validate the efficiency, we apply the proposed method to the five different DNA microarray data sets for the experiment. The experimental results show that our method of gene selection obtains better classification result of NB classifier than other method. Which demonstrates that the proposed feature selection method can obtain more correct and informative gene subset in comparison with the simple ICA for NB classifier and accuracy increases up to 97.42 % for Acute leukemia data using the Gaussian distribution.

Keywords: Feature selection, Independent component analysis (ICA), Signal to noise ratio (SNR), Naïve Bayes (NB), Classification.

1. INTRODUCTION

The genetic information of human beings is very helpful in cancer diagnosis. Cancer is a chronic disability for which a group of cells undergoes uncontrolled growth. It demolishes the adjacent tissues and sometimes spreads to other locations in the body via lymph or blood. Various research efforts, including ones based on surgery, chemotherapy, radiotherapy, are being established to fight against cancer. Recent research [1, 2] has shown a profound impact on cancer disease via a gene expression profiles, obtained by microarray technology. Microarrays provide information about the expression level of genes represented on the array. In some cases correlations between the

expression levels of a gene or a set of genes, and clinically relevant sub classifications of specific tumor subtypes have been examined [3, 4]. These inferences from observation shows that the true molecular classification and sub staging of multiple tumor types may be possible, leading to prognosis and patient management. Using microarray technologies while analyzing cancer cells, researchers have identified subgroups of tumors that differ according to types of tumor and histological subclasses of tumor, exist among carcinogenic patients [5]. Microarrays can thus be used to help classify, and predict different types of cancers. Traditional methods for identifying cancer

cells are mainly based on its morphological behavior. However, sometimes it is extremely difficult to find clear distinctions between some types of cancers according to their appearances. Hence the microarray technology stands to provide a more quantitative means for cancer diagnosis [6]. Several machine learning methods are presently used for microarray data analysis. Two major challenges of Microarray data analysis are small numbers of microarray data samples are available from a small number of patients and curse of dimensionality, thousands or tens of thousands of genes. Many genes from the dataset contain irrelevant information for the accurate classification of disease. There is highly redundant and irrelevant information in dataset needed to be removed. To select a small and discriminative subset of genes from tens of thousands of genes is extremely difficult. Therefore, the feature (gene) extraction/selection becomes the most needed requirement for accurate classification. For this cause, many researchers have used different techniques to take a little subset of informative genes that can classify different subgroups of cancers accurately. So, the microarray classification process is a two stage process: feature selection and classification. Feature selection techniques reduce large dimensional data set into smaller gene set capable of distinguishing between infected and normal samples [7-9]. There are a large number of methods which have been built up and applied to do geneselection. A typical gene selection method has two parts - an evaluation criterion and a searching strategy. As many evaluation criteria and searching schemes already exist, it is possible to develop many gene selection methods by just combining different evaluation criteria and searching schemes. Since, many of these combinations of evaluation criteria and searching schemes actually perform similarly, it is sufficient to compare amongst the most commonly used combinations instead of all possible combinations [10]. There are three widely known feature selection approaches: filter, wrapper and embedded methods. Filter methods rank genes according to certain intrinsic characteristics of gene

expressions with the class label. Filters are of two types, univariate or multivariate. The univariate filter considers intrinsic properties of each feature individually ignoring feature dependencies. T-statistics, Chi-square, Signal-to-Noise ratio (SNR), Information Gain (IG), Gain Ratio (GR) are univariate filters. The multivariate filters take feature dependencies into consideration. Correlation based feature selection (CFS) is a multivariate filter. The wrapper methods interact with the classifier while gene ranking. In embedded methods the gene selection process is embedded in constructing classifier [11, 12]. The basic idea of a feature extraction is simply to transform a high-dimensional feature vector into a low-dimensional space such that the transformed variables give information on the data which is otherwise hidden in the large data set. Principal Component Analysis (PCA), Linear Discriminant Analysis, independent component analysis (ICA) is some widely used feature extraction method. The success of feature extraction methods depends on the appropriate choice of best gene subset, two broad categories of feature subset selection are filter and wrapper [13]. In this paper, SNR ranking is being used to rank the most discriminant genes extracted by the ICA for classification. There have been many methods for performing the classification tasks such as decision tree induction, Bayesian classifier, k-nearest neighbor (k-NN), case-based reasoning, Naïve Bayes, support vector machine (SVM), genetic algorithm, neural network, logistic discrimination, and quadratic discriminant analysis etc. In this study, Naïve Bayes classifiers are used for classification. The Naïve Bayes (NB) classifier is a simple Bayesian network classifier which is established upon the firm assumption that different attributes are independent with each other given the class of education. The two major restrictions that may seriously affect the successful application of NB classifier for microarray data analysis. The first is the conditional independence assumption rooted in the classifier itself, which is hardly satisfied by the microarray data [14]. This limitation could be successfully resolved as the components extracted

by the ICA are statistically independent; therefore gene extraction by ICA could effectively improve the performance of a NB classifier for microarray data. Second limitation is that, all the attributes have an influence on the classification, hence the feature subset selection from the ICA feature vector improve the performance of a NB classifier during cross validation. It is therefore necessary to select genes to reduce the dimensionality of microarray data before applying a NB classifier [15].

In this paper, the features extracted by the ICA are ranked by the SNR-test of the DNA microarray data for NB classification. The proposed approach consists of two main steps, feature extraction by FastICA and extract feature ranked by SNR-test technique, which will be introduced in section 2. The next section explains the classification procedure of NB, followed by the details of used datasets and preprocessing step of datasets. Section 5, represent the experimental results on five microarray datasets, which shows that the proposed approach can not only improve the average classification accuracy rates but also reduce the variation in classification performance of NB. Discussions and conclusions are presented in section 6.

2. PROPOSED APPROACH:

2.1. FEATURE EXTRACTION BY ICA:

Independent Component Analysis (ICA), a computationally efficient blind source separation algorithm, ICA models, observations as a linear combination of latent feature variables, or components, which are chosen to be as statistically independent as possible, which was proposed by Hyvarinen and has been proven successful in many applications [16]. The microarray data, observations consist of microarray gene expression measurements, and independent components are interpreted to be transcriptional modules that often correspond to specific biological processes. ICA has been widely used in a variety of biological inquiries, including identifying oscillating regulatory modules in yeast cell cycle data, investigating tumor-related pathways, classifying

disease datasets, characterizing transcriptional regulators, identifying disease-specific biomarkers and examining response to bacterial infection successfully. Furthermore, ICA outperforms PCA and other unsupervised methods, in contrast to PCA, the goal of ICA is to find a linear representation of non-Gaussian data so that the components are statistically independent [17]. ICA provides a more biologically plausible model for gene expression data by assuming a non-Gaussian data distribution. ICA provides a data-driven method for exploring functional relationships and grouping genes into transcriptional modules.

In the simplest form of ICA we observe that the expression levels of all genes are n scalar random variables x_1, x_2, \dots, x_n , which are assumed to be linear combinations of m unknown independent components S_1, S_2, \dots, S_m are mutually statistically independent and zero-mean. Let us arrange the expression levels x_j into a vector $X = (x_1, x_2, \dots, x_n)^T$ which are modeled as linear combination of m random variable $S = (s_1, s_2, \dots, s_m)^T$ [18]:

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \dots + a_{jm}s_m, \text{ for all } j = 1, \dots, n \tag{1}$$

$$X = AS, \quad \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & & a_{nm} \end{bmatrix} \begin{bmatrix} s_1 \\ \vdots \\ s_m \end{bmatrix} \tag{2}$$

Where X , is $(n \times m)$ matrix which denote microarray gene expression data, with n genes and m samples, and a_{ij} , $(i=1, \dots, n, j=1, \dots, m)$ in X are some real ratio of intensities, represent the expression level of i^{th} genes in the j^{th} sample, and number of genes are much greater than that of the sample m where, $n \gg m$. This is a basic ICA model of microarray gene expression data. Since we assume that the observed variables are independent components, these are latent variable, which cannot be directly observed. Also the mixing matrix A is assumed to be unknown matrix. We only observe the random

variable x_j and we estimate both the matrix S and A using X , since we can invert the mixing matrix as

$$U = S = A^{-1}X = WX \quad (3)$$

Then ICA can be applied to find a matrix W that provides the transformation $U = u_1, u_2, \dots, u_m = WX$ of the observed matrix X under which, the transformed random variables u_1, u_2, \dots, u_m called the independent components are as independent as possible.

A fixed point algorithm is a computationally highly efficient method for performing the estimation of ICA for microarray data [19]. It is based on a fixed-point iteration scheme that has been found in independent experiments to be 10-100 times faster than conventional gradient descent methods for ICA. In the fixed point algorithm of ICA (FastICA), maximizing negentropy is used as the contrast function since negentropy is an excellent measure of non-gaussianity and is approximated by

$$J(u) = H(u_G) - H(u) \quad (4)$$

where u_G is a Gaussian random vector of the same covariance matrix as vector u . Mutual information I , is known as natural measure independence of random variables; it is widely used as the criterion in ICA algorithm and can be measured by

$$I = \sum_i J(u_i) - J(u) \quad (5)$$

where $J(u_i) = -\int p(s_i) \log p(s_i) ds_i$ is the marginal entropy of the variable u_i , $p(\cdot)$ is a probabilistic density function. The independent components are determined, when mutual information I is minimized. From equation (5) it is clearly shown that minimizing the mutual information I is equivalent to maximizing the negentropy $J(u)$. To estimate the negentropy of $u_i = w^T x$, an approximation to identify independent components one by one is designed as follows:

$$J_G(w) = [E\{G(w^T x)\} - E\{G(v)\}]^2 \quad (6)$$

Where, G can be practically any non-quadratic function, $E(\cdot)$ denotes the expectation, and v is a

Gaussian variable of zero mean and unit variance [20].

2.2 FEATURE SELECTION BY SNR TECHNIQUE:

The SNR test is used to select the best feature subset from the ICA feature vector for good separability of the classification task. A central issue associated with ICA is, generally extracted number of component which are equal to the observational variables m for which again 2^m feature subsets exist [21]. The evaluation of all possible feature subsets leads to computational problem for large values of m . To solve this problem of identifying the most relevant gene subsets we applied SNR technique. In this paper, we use Signal-to-noise Ratio Score for ranking the genes which are extracted by ICA.

The SNR score identifies the expression patterns with a maximal difference in mean expression between two groups and minimal variation of expression within each group. In this method genes are first ranked according to their expression levels using SNR test Statistic. The SNR is defined as follows:

$$\text{Signal to noise ratio} = (\mu_1 - \mu_2) / (\sigma_1 + \sigma_2) \quad (7)$$

Here μ_1 and μ_2 denote the mean expression values for the sample class 1 and class 2 respectively. σ_1 and σ_2 are the standard deviations for the samples in each class [22].

3. CLASSIFIERS:

3.1 NAIVE BAYES CLASSIFIER:

Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumption. Bayesian network classifier computes the posterior probability that the sample belongs to class H by using the Bayes theorem for multiple evidences as follows [23]:

$$P(H | E_1, E_2, E_3, \dots, E_n) = \frac{P(E_1, E_2, E_3, \dots, E_n | H) \times P(H)}{P(E_1, E_2, E_3, \dots, E_n)} \quad (8)$$

Since it is difficult to compute the joint conditional probability in Eq. (8), two assumptions are often imposed to simplify the computation. One is that all

the features $E_1, E_2, E_3, \dots, E_n$ are independent with each other given the class variable H . The other is that all the features are directly dependent on the class variable H . The two assumptions make the computation of the joint conditional probability equivalent to the product of all the marginal conditional probabilities. With these assumptions the Bayes theorem can be written as:

$$P(H|E_1, E_2, E_3, \dots, E_n) = \frac{P(E_1|H) \times P(E_2|H) \times \dots \times P(E_n|H) \times P(H)}{P(E_1, E_2, E_3, \dots, E_n)} \quad (9)$$

Since $P(E_1, E_2, E_3, \dots, E_n)$ is a common factor for a certain sample, it can be ignored in the classification process. In addition, since the attribute variables are continuous in microarray data analysis, we can use the probability density value $f(E_i|H)$ to replace the probability value $P(E_i|H)$. The class-conditional probability density $f(.|H)$ for each attribute and the prior $P(H)$ can be obtained from the learning process. For the estimation of $f(.|H)$ we use the nonparametric kernel density estimation method [15, 24]. As a result, the general Bayesian classifier given by Eq. (8) can be simplified as the Naïve Bayes classifier given by Eq. (10). Figure 3 shows the simplified form of Bayesian classifier as the Naïve Bayes classifier.

$$H' = \arg \max_{H \in \omega} P(H) \prod_{i=1}^n f(E_i | H) \quad (10)$$

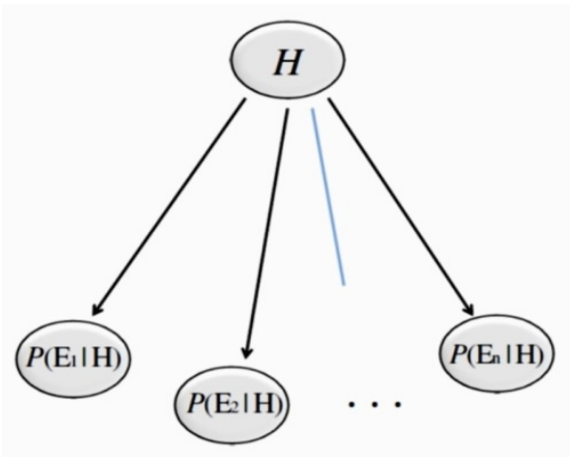


Fig. 1: Naïve Bayes Classifier.

4. DATASET:

We evaluate the performance of the proposed feature selection approach on five publicly available microarray data sets of Colon cancer [25], Acute leukemia [26], and Prostate cancer [27], high-grade Glioma data [28], and Lung cancer-II [29] dataset, taken from Kent ridge an online repository of high-dimensional biomedical data sets, (<http://datam.i2r.astar.edu.sg/datasets/krbd/index.html>) to study the cancer classification problem. Table 1 shows the five datasets with their properties. These datasets are preprocessed by setting thresholds and log-transformation on the original data. Threshold technique is mostly achieved by restricting gene expression levels to be larger than 20. In other words, the expression levels that are smaller than 20 will be set to 20. Regarding the log-transformation, the natural logarithm of the expression levels is usually taken. In addition, no further preprocessing is applied to the rest of the dataset. After preprocessing the data, independent component analysis is performed on the gene expression data set. For ICA, the FastICA algorithm software package for Matlab (R2010a) is applied. Then SNR-test is used to rank the independent component feature vectors. For evaluating the performance of the proposed method using NB classifier, the dataset is classified with these reduced numbers of genes by using above mentioned classifier. The classifier and feature selection method was implemented with MATLAB software.

Table 1: Summary of five microarrays the datasets.

Data set	No. of classes	No. of features	No. of samples
Colon cancer (<i>Alon et al., 1999</i>)	2	2000	62
Acute leukemia (<i>Golub et al., 1999</i>)	2	7129	72
Prostate tumor (<i>Dinesh Singh et al., 2002</i>)	2	12600	102
High-grade Glioma (<i>C.L. Nutt et al., 2003</i>)	2	12625	64
Lung cancer II (<i>Gorden et al., 2002</i>)	2	12533	181

5. EXPERIMENTAL RESULT:

To check the performance of the proposed approach with NB classifier, the above mentioned combination has been applied on the five DNA microarray gene expression datasets. Since all data samples in the five datasets have already been assigned to a training set or test set. The training dataset is used to do gene selection and then built the model for classification of the test dataset to evaluate the performances of classifiers. To demonstrate the efficiency and feasibility of our proposed method, the results of the other three gene selection methods for the same classifier are also computed for comparison. In method 1, all the features are extracted by principle component analysis for NB classification and the same is applied for method 2 except using ICA for feature extraction. Method 3 is similar to our proposed method where PCA is used with SNR for NB classification and in method 4 ICA with SNR. The classification for pure Naive Bayes classifier was not included due to its extremely time-consuming computations.

Due to the small sample size of microarray data Leave-One-Out Cross-Validation (LOOCV) accuracy rates are used to give a relatively comprehensive comparison of the performances of alternative methods. In LOOCV method of cross validation the number of partitions of data set is equal to the number of sample size (m). Each test set consists of a different singleton set and each training set consists of all ($m-1$) cases not in the corresponding test set. Given a dataset containing m samples, ($m-1$) samples are used to construct a classifier and then apply the remaining one data sample to test this classifier. By repeating this process of successively using each data samples (x_i) as the testing data sample, totally m prediction $e_i = c(x_i)$ ($i = 1-m$) are obtained. The performance of the classifier is then measured by the average misclassification rate:

$$E_r = \frac{1}{m} \sum_{i=1}^m \delta(e_i, y_i),$$

Where y_i , is the true class label for instance x_i , and

$$\delta(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$$

A comparison between different methods on the different datasets has been shown in Table 2, Table 3, Table 4, Table 5 and Table 6 respectively. It can be seen from Table 2-6 the classification accuracy of classifier with our proposed method compared to other three gene selection methods is more accurate, feasible and also reduce the variation of classification performance. So, the proposed approach improves the classification performance of the NB classifiers for microarray data.

Table 2: Classification accuracy of NB classifier by using Gaussian distribution and Kernel density estimation for Colon cancer data.

S. No.	Classifier	Method	Mean accuracy	Variance
1.	NB with kernel density estimation	PCA+ NB	76.71	0.065
2.		ICA+ NB	80.19	0.062
3.		PCA+SNR+ NB	85.33	0.032
4.		ICA+SNR+ NB	91.88	0.026
1.	NB with Gaussian distribution estimation	PCA+NB	76.88	0.069
2.		ICA+ NB	79.81	0.059
3.		PCA+SNR+NB	81.65	0.036
4.		ICA+SNR+ NB	89.46	0.014

Table 3: Classification accuracy of NB classifier by using Gaussian distribution and Kernel density estimation for Acute leukemia data.

S. No.	Classifier	Method	Mean accuracy	Variance
1.	NB with kernel density estimation	PCA+ NB	77.66	0.067
2.		ICA+ NB	89.13	0.038
3.		PCA+SNR+ NB	92.53	0.029
4.		ICA+SNR+ NB	95.30	0.011
1.	NB with Gaussian distribution estimation	PCA+NB	69.53	0.063
2.		ICA+ NB	85.31	0.061
3.		PCA+SNR+NB	95.62	0.029
4.		ICA+SNR+ NB	97.42	0.021

Table 4: Classification accuracy of NB classifier by using Gaussian distribution and Kernel density estimation for Prostate tumor data.

S. No.	Classifier	Method	Mean accuracy	Variance
1.	NB with kernel density estimation	PCA+NB	77.73	0.110
2.		ICA+NB	82.65	0.089
3.		PCA+SNR+NB	84.63	0.066
4.		ICA+SNR+NB	90.11	0.033
1.	NB with Gaussian distribution estimation	PCA+NB	74.73	0.097
2.		ICA+NB	79.45	0.087
3.		PCA+SNR+NB	87.62	0.055
4.		ICA+SNR+NB	89.19	0.041

Table 5: Classification accuracy of NB classifier by using Gaussian distribution and Kernel density estimation for High-grade Glioma data.

S. N	Classifier	Method	Mean accuracy	Variance
1.	NB with kernel density estimation	PCA+NB	70.72	0.051
2.		ICA+NB	74.21	0.047
3.		PCA+SNR+NB	79.32	0.039
4.		ICA+SNR+NB	81.21	0.037
1.	NB with Gaussian distribution estimation	PCA+NB	70.88	0.039
2.		ICA+NB	74.30	0.044
3.		PCA+SNR+NB	77.42	0.028
4.		ICA+SNR+NB	79.33	0.025

Table 6: Classification accuracy of NB classifier by using Gaussian distribution and Kernel density estimation for Lung cancer II data.

S. No.	Classifier	Method	Mean accuracy	Variance
1.	NB with kernel density estimation	PCA+NB	75.33	0.085
2.		ICA+NB	81.12	0.093
3.		PCA+SNR+NB	86.21	0.066
4.		ICA+SNR+NB	92.23	0.026
1.	NB with Gaussian distribution estimation	PCA+NB	82.54	0.060
2.		ICA+NB	87.52	0.079
3.		PCA+SNR+NB	93.32	0.033
4.		ICA+SNR+NB	96.52	0.010

In order to study the behavior of a proposed feature selection approach, we applied it to the Colon, Leukemia, Prostate, High-grade Glioma and Lung cancer II data set for NB classification kernel density and Gaussian distribution estimation, a graph is plotted between the number of selected genes and classification accuracy rates.

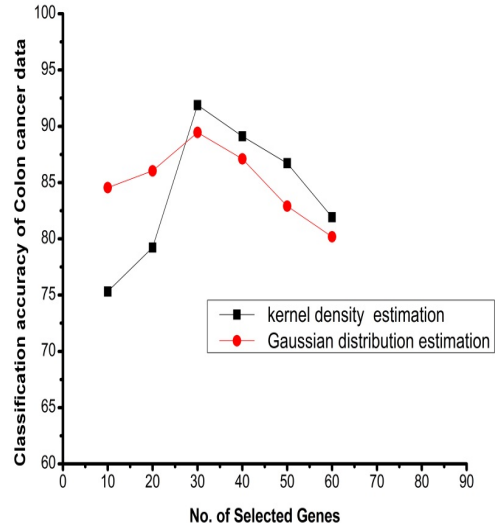


Fig.2: Number of selected genes V/s Classification accuracy, using NB classifier with kernel density and Gaussian distribution estimation of Colon cancer data, based on proposed method.

Figure 2 shows the graph between the number of selected genes and the classification accuracy, using kernel density and Gaussian distribution estimation with NB classifier for colon cancer data based on the proposed gene selection method. Colon Cancer dataset consist of 62 samples with 2000 (genes) features of two classes. Here by ranking the gene with SNR-technique, we managed to enhance the mean classification accuracy significantly. The mean improvement in classification accuracy was verified by adding 10 genes, each time in training sets. The peak of the graphs shows the best means classification accuracy for Colon data sets. The best mean accuracy of a NB classifier with the proposed method was found 91.88 % and 89.46 % for 30 selected genes with kernel density and Gaussian distribution estimation respectively.

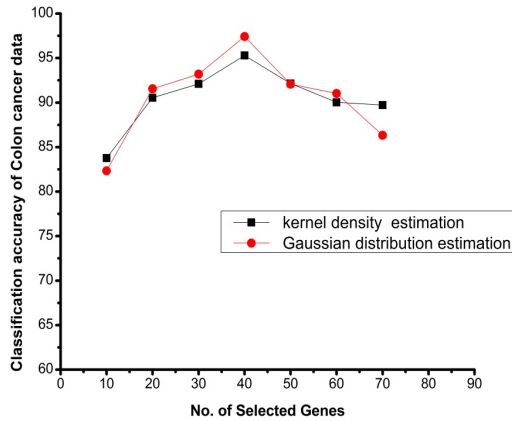


Fig.3: Number of selected genes V/s Classification accuracy, using NB classifier with kernel density and Gaussian distribution estimation of Acute leukemia cancer data, based on proposed method.

Acute leukemia dataset consists of 72 samples with 7129 genes of two classes. From figure 3, we can see the results of classification accuracy with the number of selected genes for leukemia dataset. As shown in Table 3, with this data set using kernel and Gaussian estimation for NB classifier with ICA feature vector, the highest mean accuracy obtained was 89.73% and 86.33%. When SNR-technique is employed to rank the independent components feature vector, to get 95.3% and 97.42% mean classification accuracies of NB classifier with kernel density and Gaussian distribution respectively.

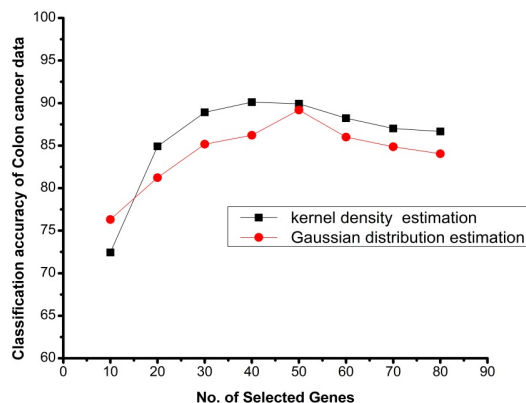


Fig.4: Number of selected genes V/s Classification accuracy, using NB classifier with kernel density and

Gaussian distribution estimation of Prostate tumor cancer data, based on proposed method.

Figure 4, shows the graph for the classification accuracy of the prostate cancer dataset with a number of selected genes using the SNR+ICA approach with NB classifier. The peak of the graphs shows those here, 40 genes for kernel density and 50 genes for Gaussian distribution were used for best NB classification. It can be seen from the graph that the highest mean accuracies was found 90.11 and 89.19 (with the difference of 10 genes) for kernel density and Gaussian distribution estimation of NB classification respectively. These results clearly show that the SNR-technique with ICA performs better than the other existing methods.

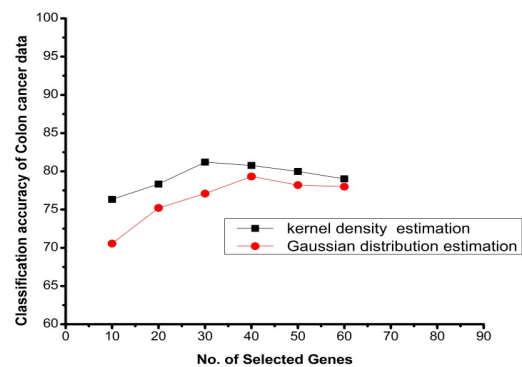


Fig.5: Number of selected genes V/s Classification accuracy, using NB classifier with kernel density and Gaussian distribution estimation of High-grade Glioma cancer data, based on proposed method.

High-grade Glioma dataset consist of 64 samples with 12625 genes of two classes. From this data set 63 genes are extracted by FastICA from the training set. Figure 5, shows the classification accuracy graph of High-grade Glioma data by ranking of the genes with SNR-technique, using kernel density and Gaussian distribution estimation for NB classifier. The values of mean classification accuracy with the proposed method for kernel density and Gaussian distribution for NB classifiers are 81.21 % and 79.33 %, respectively, which is very low as compared to the accuracies of other datasets.

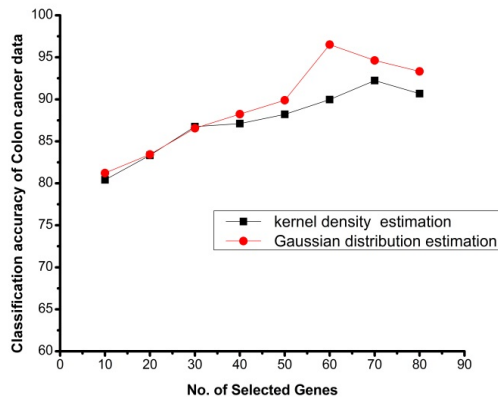


Fig.6: Number of selected genes V/s Classification accuracy, using NB classifier with kernel density and Gaussian distribution estimation of Lung cancer II cancer data, based on proposed method.

In lung cancer-II data set there were 181 samples with 12533 genes. Figure 6 clearly shows the difference between the classification accuracies of this dataset using kernel density and Gaussian distributions for NB classifier. Interestingly, for both kernel density and Gaussian distribution gives highest mean classification accuracy of a NB classifier with the same number of selected genes. Classification accuracy of a NB classifier of this dataset with Gaussian distribution is more as compared with kernel density with the same number of 50 selected genes.

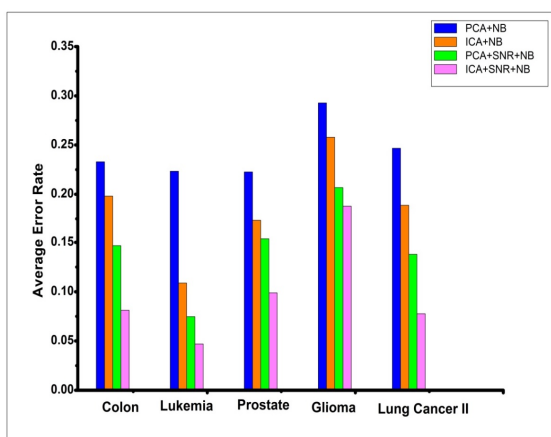


Fig.7: Average error rate of NB classifier with kernel density estimation for the five datasets with different gene selection method.

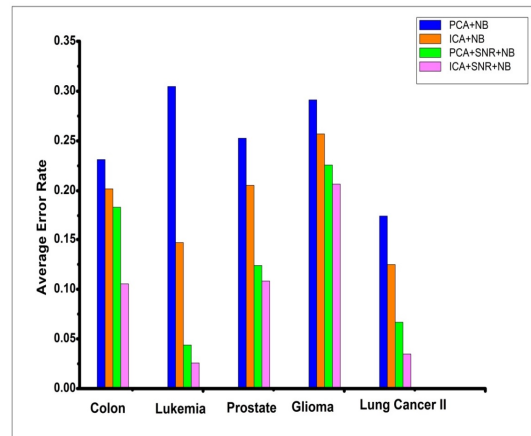


Fig.8: Average error rate of NB classifier with Gaussian distribution estimation for the five datasets with different gene selection method.

Figure 7 and 8 shows the graph of the average error rate of a NB classifier with two Kernel function for the five datasets with different gene selection methods. It clearly shows from the figures that NB classifier with kernel density performs better than a Gaussian distribution function because of the reduced error rate. It is evident from the graph that when we use top ranked genes based on SNR-technique from PCA then the percentage error rate is minimized, so the PCA+SNR - test method performs better than PCA method with NB classifier. Also from figure 7,8 see that the proposed method ICA + SNR-test with NB gives the minimized error rate, which shows the significance of the proposed method with the other existing methods.

6. CONCLUSION:

This paper presents a SNR-test based feature selection approach in ICA feature vector for NB classification of microarray data where the methodologies involve dimension reduction of microarray data using ICA, followed by the feature ranking using SNR-test. The approach was tested by classifying five data sets. The experimental results show that our combination of gene selection methods of an existing algorithm together with NB classifier is giving better results as compared to other existing approaches. Our experimental results

on five microarray datasets demonstrate the effectiveness of the proposed approach in improving the classification performance of the NB classifier in microarray data analysis. It is also found that the proposed method can obtain better classification accuracy with a smaller number of selected genes than the other existing methods, so our proposed method is effective and efficient for NB classifier.

ACKNOWLEDGMENT:

The author would like to acknowledge the support of the Director (Dr. AppuKuttan K.K.), Maulana Azad National Institute of Technology Bhopal-462051 (M.P.) India for providing basic facilities in the institute. The support of the Dr. Sanjay Sharma (Prof & Head) Department of Mathematics and Computer Application, Maulana Azad National Institute of Technology Bhopal-462051 (M.P.) India is kindly acknowledged.

7. REFERENCE:

- Liu, Y., *Prominent feature selection of microarray data*. Progress in Natural Science, 2009. **19**(10): p. 1365-1371.
- George, G. and V.C. Raj, *Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile*. arXiv preprint arXiv:1109.1062, 2011.
- Bhattacharjee, A., et al., *Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses*. Proceedings of the National Academy of Sciences, 2001. **98**(24): p. 13790-13795.
- Garber, M.E., et al., *Diversity of gene expression in adenocarcinoma of the lung*. Proceedings of the National Academy of Sciences, 2001. **98**(24): p. 13784-13789.
- De, R.K. and A. Ghosh, *Interval based fuzzy systems for identification of important genes from microarray gene expression data: Application to carcinogenic development*. Journal of biomedical informatics, 2009. **42**(6): p. 1022-1028.
- Chu, F. and L. Wang, *Applications of support vector machines to cancer classification with microarray data*. International journal of neural systems, 2005. **15**(06): p. 475-484.
- Saeys, Y., I. Inza, and P. Larrañaga, *A review of feature selection techniques in bioinformatics*. bioinformatics, 2007. **23**(19): p. 2507-2517.
- Debnath, R. and T. Kurita, *An evolutionary approach for gene selection and classification of microarray data based on SVM error-bound theories*. Biosystems, 2010. **100**(1): p. 39-46.
- Díaz-Uriarte, R. and S.A. De Andres, *Gene selection and classification of microarray data using random forest*. BMC bioinformatics, 2006. **7**(1): p. 3.
- Tang, E.K., P.N. Suganthan, and X. Yao. *Feature selection for microarray data using least squares svm and particle swarm optimization*. in *Computational Intelligence in Bioinformatics and Computational Biology, 2005. CIBCB'05. Proceedings of the 2005 IEEE Symposium on*. 2005. IEEE.
- Ammu, P. and V. Preeja, *Review on Feature Selection Techniques of DNA Microarray Data*. International Journal of Computer Applications (0975-8887) vol, 2013: p. 39-44.
- Du, D., et al., *A novel forward gene selection algorithm for microarray data*. Neurocomputing, 2014. **133**: p. 446-458.
- Bartenhagen, C., et al., *Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data*. BMC bioinformatics, 2010. **11**(1): p. 567.
- Patil, T.R. and M. Sherekar, *Performance analysis of Naive Bayes and J48 classification algorithm for data classification*. Int J Comput Sci Appl, 2013. **6**: p. 256-261.
- Fan, L., K.-L. Poh, and P. Zhou, *A sequential feature extraction approach for naive bayes classification of microarray data*. Expert Systems with Applications, 2009. **36**(6): p. 9919-9923.
- Hyvarinen, A., *Fast and robust fixed-point algorithms for independent component analysis*. Neural Networks, IEEE Transactions on, 1999. **10**(3): p. 626-634.
- Naik, G.R. and D.K. Kumar, *An overview of independent component analysis and its applications*. Informatica: An International Journal of Computing and Informatics, 2011. **35**(1): p. 63-81.
- Engreitz, J.M., et al., *Independent component analysis: Mining microarray data for fundamental human gene expression modules*. Journal of biomedical informatics, 2010. **43**(6): p. 932-944.
- Hyvärinen, A., J. Karhunen, and E. Oja, *Independent component analysis*. Vol. 46. 2004: John Wiley & Sons.
- Capobianco, E., *Exploration and reduction of high dimensional spaces with independent component analysis*. 2004.
- Zheng, C.-H., et al., *Gene expression data classification using consensus independent component analysis*. Genomics, proteomics & bioinformatics, 2008. **6**(2): p. 74-82.

22. Hengprapromh, S., *GA-Based Classifier with SNR Weighted Features for Cancer Microarray Data Classification*. 2013.
23. Zhang, B.-T. and K.-B. Hwang, *Bayesian network classifiers for gene expression analysis*, in *A Practical Approach to Microarray Data Analysis*. 2003, Springer. p. 150-165.
24. Baldi, P. and A.D. Long, *A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes*. *Bioinformatics*, 2001. **17**(6): p. 509-519.
25. Alon, U., et al., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*. *Proceedings of the National Academy of Sciences*, 1999. **96**(12): p. 6745-6750.
26. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*. *science*, 1999. **286**(5439): p. 531-537.
27. Singh, D., et al., *Gene expression correlates of clinical prostate cancer behavior*. *Cancer cell*, 2002. **1**(2): p. 203-209.
28. Nutt, C.L., et al., *Gene expression-based classification of malignant gliomas correlates better with survival than histological classification*. *Cancer research*, 2003. **63**(7): p. 1602-1607.
29. Gordon, G.J., et al., *Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma*. *Cancer research*, 2002. **62**(17): p. 4963-4967.

First Author Rabia Aziz (2nd July, 1982) has done her Bachelor of Science in 2001 (Allahabad University, Allahabad), M.Sc. (Mathematics) in 2009 (Barkatullah University, Bhopal) and M.Phil. in Mathematics from Institute for excellence in Higher Education Bhopal-462061, India in 2012. Her research interest includes Micro-array data



analysis, financial mathematics, Data mining and Machine learning. She is having 6 years of experience and is working as Research Scholar in the Department of Mathematics and Computer Application at Maulana Azad National Institute of Technology, Bhopal-462051 (M.P.) India.

Dr. Namita Srivastava (9th October 1965) has done her Bachelor of Science in 1985, M.Sc. (Mathematics) in 1987 and Ph. D. in Mathematics from Barkatullah University in 1992. Her research interest includes fracture mechanics, financial mathematics, parallel computing and parallel mining. She is having 25 years of experience and is working as Professor in the Department of Mathematics at Maulana Azad National Institute of Technology, Bhopal India. She has published 35 papers in international journal, 25 papers in national journal and 20 papers in proceedings of international and national conference. She guided more than 8 Ph.D.



Dr. C. K. Verma (8th June 1975) has done Master of Science in Mathematics from Govt. Science College Jabalpur India in 1998 and has Qualified National Eligibility test (NET) in 2000. He has done his Ph.D from National Institute of Technology, Bhopal in 2012. His area of research includes Computational Biology. He is having 15 years of teaching experience and is working as Assistant Professor in the Department of Mathematics at Maulana Azad National Institute Bhopal, India. He has published 08 papers in international journal, 02 papers in national journal and 04 papers in proceedings of international and national conference.

