International Journal of Bioinformatics Research
ISSN: 0975-3087 & E-ISSN: 0975-9115, Volume 6, Issue 1, 2015, pp.-305-312.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BIJ0000001

t-INDEPENDENT COMPONENT ANALYSIS FOR SVM CLASSIFICATION OF DNA- MICROARRAY DATA

AZIZ R.*, SRIVASTAVA N. AND VERMA C.K.

Department of Mathematics & Computer Application, Maulana Azad National Institute of Technology Bhopal - 462 051, MP, India.
*Corresponding Author: Email- rabia.aziz2010@gmail.com

Received: January 23, 2015; Revised: March 12, 2015; Accepted: March 16, 2015

Abstract- Classification analysis of microarray data is known to be hard because it involves thousands of features (genes) values, so it is necessary to reduce the number of features to obtain a manageable size of data for classification. In the present work two existing feature extraction/selection algorithms, namely Independent component analysis (ICA) and t-test are used which is a new combination of extraction/selection. The main objective of this paper is to rank the independent components of the DNA microarray data using t-test to improve the performance of Support Vector Machine (SVM)) classifier. To show the validity of the proposed method, it is applied to reduce the number of genes of five DNA microarray datasets then classify these datasets by using the SVM classifier. Experimental results on five datasets demonstrate that genes selected by proposed approach effectively improve the performance of SVM classifiers in terms of classification accuracy. We compare our proposed method with several existing methods and find that the proposed method can obtain better classification accuracy, using SVM classifier and accuracy increased up to 94.42 % of Acute leukemia data using the RBF kernel.

Keywords- Independent component analysis (ICA), t-test, Support vector machine (SVM), feature selection, classification

Citation: Aziz R., Srivastava N. and Verma C.K. (2015) t-Independent Component Analysis for SVM Classification of DNA- Microarray Data. International Journal of Bioinformatics Research, ISSN: 0975-3087 & E-ISSN: 0975-9115, Volume 6, Issue 1, pp.-305-312.

Introduction

A wide range of changes in genes occurs during the transformation of a normal cell to a cell capable of forming a cancerous growth. Usually cancer is caused by multiple changes in several different genes, although the genes that contribute to the development of cancer fall into broad categories. Every case of cancer is unique, with its own set of genetic changes and growth properties. Some cancers grow quickly while others can take years to become dangerous to the patient. When a normal tissue becomes cancerous, the expression levels of genes also change since transcriptional changes accurately reflect the status of disease, including cancers, by identifying these changes in gene expression, the tissues can be classified as cancerous and normal. The differences between cases of cancer, even of the same organ, are one of the main reasons that treatment is so difficult. High-density DNA microarray measures the activities of several thousand genes simultaneously and the gene expression profiles have been used for the cancer classification recently [1]. Currently, cancer diagnosis highly depends on a variety of histological observations, including immune histo chemical assays, which detect cancer biomarker molecules. However, these assays have limitations due to morphological similarity and lack of available biomarkers of cancers. Microarray technology is a hybridization technique which allows monitoring the quantity of messenger RNA present in a cell for several thousand genes simultaneously in a single experiment on a small chip. By submitting cells to various experimental conditions and comparing the expression profiles of different genes, a better understanding of the regulation mechanisms and functions of each gene is expected [2]. The output of these microarray experiments are the expression levels of different genes and these data are publicly available. This revolutionized the approach has provided a large amount of data from which a lot of knowledge can be explored. These datasets include a large number of gene expression values and need to have a good data mining method to extract knowledge from these microarray gene expression datasets [3]. A reliable and accurate classification is essential for successful diagnosis and treatment of cancer [4]. Microarrays have thousands to tens-of-thousands of gene features, but only a few hundred patient samples are available. However, among the large amount of genes, only a small fraction is effective for performing a classification task, so the dimension reduction is one of the important procedures for DNA-microarray data. In conjunction with this invention, identifying gene markers that present the maximum discrimination power between cancerous and normal cells has become one of the vital research areas in microarray data analysis. This trouble can be alleviated by using two types of methods: Feature Extraction and Feature selection. The goal of both the methods is to determine a small subset of informative features that reduces processing time and provides higher classification accuracy [5]. 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. These methods include clus-

tering, basic linear transforms of the input variables (Principal Component Analysis/Singular Value Decomposition, Linear Discriminant Analysis), spectral transforms, wavelet transforms or convolution of kernels. Feature selection aims at selecting a subset of features relevant in terms of discrimination capability. It avoids the drawback of the output interpretability, because the selected features represent a subset of the given ones. The feature selection methods are classified as filters, wrappers and embedded, depending upon the criteria used to evaluate the feature subsets [6]. The filter approach is widely used based on gene ranking, yet the drawback of this selection procedure is that, it is independent of the specific required prediction/classification task. The wrapper method, such as sequential forward selection and particle swarm optimization usually consists of the search process and the evaluation criterion. However, an exhaustive search of all subsets is too expensive to implement for a high dimensional feature space. Unlike the filter and wrapper methods that separate the variable selection and training process, the embedded methods incorporate feature selection into the construction process of the classifier or regression model [7]. A large number of gene selection & extraction approaches exist, such as t-test, relief-F, information gain, and Principal Component Analysis (PCA), Linear Discriminant Analysis, independent component analysis (ICA). These methods are capable of selecting a smaller subset of genes for sample classification [8]. Recently Independent component analysis (ICA) methods have received growing attention as effective data-mining tools for microarray gene expression data. As a technique of higher-order statistical analysis, ICA is capable of extracting biologically relevant gene expression features of microarray data [9]. The success of ICA methods depends on the appropriate choice of best gene subset from given ICA feature vector and choice of an appropriate classifier [10]. Several machine learning techniques, such as Artificial neural networks (ANN), k-nearest neighbor (KNN), support vector machine (SVM), Naïve Bayes, Decision Tree, Random Forest and kernel-based classifiers, have been successfully applied to microarray data and also for other biological data analyses in recent years [4,11,12]. Statistical and machine learning approaches are popularly used to construct a predictive model for classifying cancer patients from normal ones based on gene expression data. A few of such approaches include the SVM-based classifier is superior, as it is less sensitive to the curse of dimensionality and more robust than other non-SVM classifiers [13]. The biggest drawback of an SVM is that it cannot directly obtain the genes of importance. Thus, during the fitting of an SVM model, a careful gene selection has to be done first and then the selected genes should be used to obtain improved classification results. If genes are not appropriately chosen, there may be a large number of redundant variables in the model, severely affecting its performance [14]. From the study of L Chun-Hou Zheng (2006), we see that SVM is the best classifiers with ICA for microarray data, and feature subset selection from the ICA feature vector can significantly improve the performance of classifiers [3]. In this study, the most discriminant features extracted by the ICA are ranked by the t-test. The t-test compares the actual difference between two means in relation to the variation in the data.

In this paper, the features extracted by the ICA are ranked by the t-test of the DNA microarray data for support vector machine (SVM) classification. The proposed approach consists of two main steps, feature extraction by FastICA and extract feature ranked by t-test technique, which will be introduced in section 2. The next section explains the classification procedure of SVM, 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 of classification performance. Finally, the concluding section discusses the applicability of our proposed methods used [Fig-1].

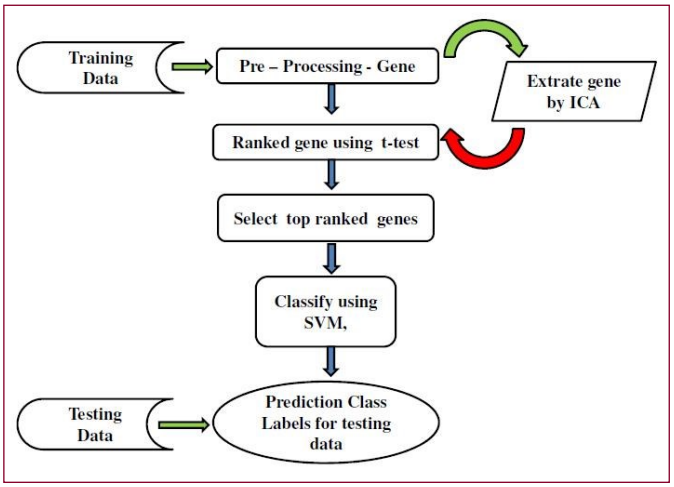

Fig. 1- The procedure of the method used in this paper.

Proposed Approach

Feature Extraction by ICA

ICA is a statistical method for transforming an observed multidimensional random vector into components that are mutually as independent as possible, which was proposed by Hyvarinen and has been proven successful in many applications [15]. ICA is a useful extension of PCA that has been acquired in context with blind separation of independent sources from their linear mixtures. PCA projects the data into a new space spanned by the principal components. 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 [16]. 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. Independent component analysis (ICA) is a signal processing technique whose goal is to express a set of random variables as linear combinations of statistically inde-pendent component variables. Two interesting applications of ICA are blind source separation and feature extraction.

In the simplest form of ICA we observe the expression levels of all genes are n scalar random variables $x_1, x_2, .. x_n$, which are assumed to be linear combinations of m unknown independent components $S_1, S_2, \ldots S_m$ that is mutually statistically independent, and zero-mean. Let us arrange the expression levels xj into a vector $X = (x_1, x_2, \ldots, x_n)^T$ which are modeled as linear combination of m random variable $S = (s_1, s_2, \ldots s_m)^T$ [17]:

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \ldots + a_{jm}s_m, \text{ for all } j = 1, \ldots\ldots\ldots, n \quad (1)$$

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

Where X, is (n × m) matrix which denote microarray gene expres-

sion data, with n genes and m samples, and 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 >> 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 of the observed matrix X under which, the transformed random variables called the independent components are as independent as possible. Theoretical framework of ICA algorithms of microarray gene expression data shown in [Fig-2], as previously demonstrated by Wei Kong et al [18].

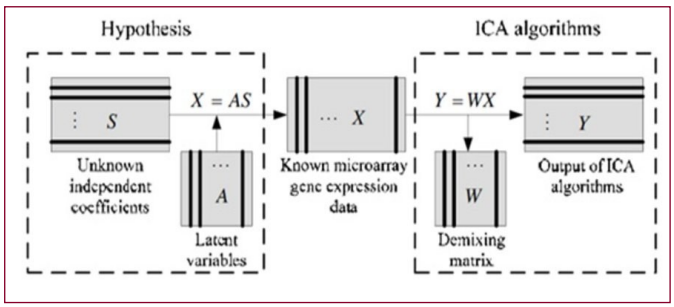

Fig. 2- Theoretical framework of ICA algorithms of microarray gene expression data

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 nonguassianity 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 *ui*, p (.) is a probabilistic density function. The independent components are determined, when mutual information *I* is minimized. From [Eq-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 (.) denotes the expectation, and v is a Gaussian variable of zero mean and unit variance [20].

Feature Selection by t-Test Technique

The t-statistic measure is used to select the best gene from the given ICA feature vector for good separability of the classification task. A central issue associated with ICA is, it generally extracts the number of components, which are equal to the observational variables m for which again 2m gene subsets exist [5]. The evaluation of all possible gene subsets leads to computational problem for large values of m. To solve this problem of identifying the most relevant gene we applied t-test method.

In this paper, we use t-test for ranking the genes which is extracted by ICA, t-test (t-score or TS) is a statistical method and is used to measure how large the difference is in between the distributions of two groups of samples. We shall focus on the classification problems with two classes, labeled by 1 and 2, respectively, and let m_k denote the sample size for class k; i.e., $m_1 + m_2 = m$. The response variable yj, j = 1. . . m, takes on the values of +1 or −1 for the two classes, respectively. The t-statistic measures the separability between classes using a standardized distance for a single gene, which gives a relevance score for each gene [21]. The ranking criterion is given as:

$$TS_i = \frac{\overline{x_{1i}} - \overline{x_{2i}}}{\sqrt{\frac{s^2_{1i}}{m_1} + \frac{s^2_{2i}}{m_2}}}$$

Where, *TSi* is a t-statistic measures for the gene *i*, and x_{ki} to denote the vector of values on the i^{th} row of X for gene *i*, that belong the class k $\in$ {1, 2}. The mean of the values in x_{ki} is denoted by $\overline{x_{ki}}$, and the sample standard deviation by s_{ki} [22].

The genes with largest TS, put in the first place in the ranking list, followed by the gene with the second largest TS, and so on. To measure the relevance of a gene, the t-test is widely used, assuming that there are two classes of samples in a gene expression data set. The t-test is computed for one class versus the other classes. It compares the actual difference between two means in relation to the variation in the data. The test values are calculated as the test statistic t is used to detect the difference between the means of two populations and it has two versions depending on whether or not the two variances of the two populations are equal. The statistics is not only used for two class prediction problems, but they also apply to the class discovery [23].

Classifiers

SVM Classifier

The support vector machine (SVM) is a widely used tool for 2-class classification and it is inspired by the idea of maximizing the geometric margin. SVM performs classification by constructing an optimal hyperplane which separates the data into different classes [24]. The SVM is a linear classifier that maximizes the margin between the separating hyperplane and the training data points. It has no local minima, i.e. it works out a convex optimization problem. The algorithm can automatically define a network architecture. For these causes, it is a lot more attractive in application areas than the other neural networks. Basically SVM is designed for binary classification problems, and many different forms of SVM algorithms have been introduced for different purposes. In case of linearly separable data, the goal of training phase of SVM is to find the linear function:

$$f(x) = W^T X + b \quad (7)$$

which is the border for two different data classes and divides the

space into two classes according to the condition:

$$W^T X + b > 0 \qquad W^T X + b < 0$$

The separating plane is defined by $W^T X + b = 0$, and the distance between the two parallel hyperplane is equal to $2/\|W\|^2$

This quantity is termed as the classification margin as shown in [Fig -3]. For maximizing the classification margin the SVM requires the solution of the following optimization problem [25]:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|W\|^2 \\ &\text{subject to} && Y_i(W^T X_i + b) \geq 1 \end{aligned} \tag{8}$$

In case of nonlinearly separable data, SVM has to map the data from the input space into a higher-dimensional feature space, where the classes can then be separated by a hyperplane. The function that performs this mapping is called a kernel function. In SVM the following four basic Kernel functions are used [26]:

1. Linear : $K(X_i, X_j) = X_i^T X_j$
2. Polynomial: $K(X_i, X_j) = (\gamma X^T{}_i, X_j + r)^d, \ \gamma > 0$
3. *Radial basis function (RBF):*
$$K(X_i, X_j) = \exp(-\gamma \|X_i - X_j\|)^2, \quad \gamma > 0$$
4. *Sigmoid :* $K(X_i, X_j) = \tanh(\gamma X_i^T, X_j + r)$

where r, d and γ is a kernel parameter.

For nonlinearly separable data, SVM requires the solution of the following optimization problem:

$$\begin{aligned} &\text{min imize} && \frac{1}{2}\|W^T\|^2 + C\sum_{i=1}^{n} \xi_i \\ &\text{subject to:} && Y_i(W^T X_i + b) \geq 1 - \xi_i \\ & && \xi_i \geq 0 \end{aligned} \tag{9}$$

where $\xi_i \geq 0$ are slack variables that allow the elements of the training data set to be at the margin or to be misclassified. So, these points which are on the margin are called support vectors [27]. In this study, SVM with RBF kernel as the classifier is used.

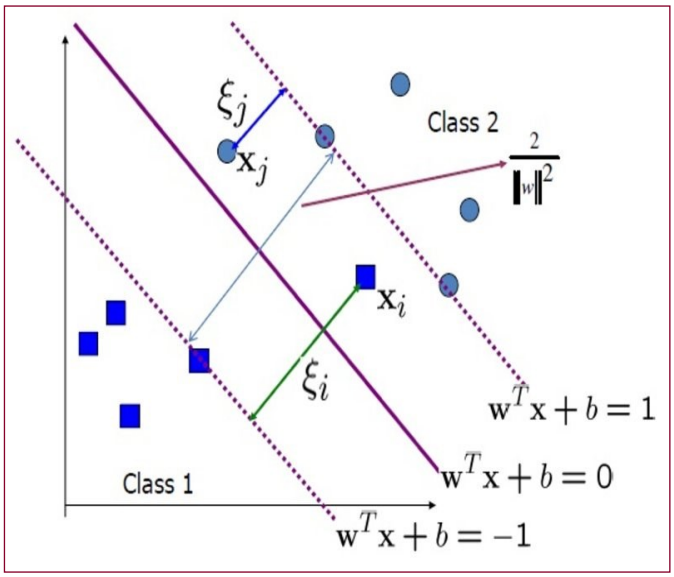

Fig. 3- Maximum margin hyperplanes for SVM divides the plane into two classes

Dataset

We evaluate the performance of the proposed feature selection approach on five publicly available microarray data sets of Colon cancer [28], Acute leukemia [29], Prostate cancer [30] Lung cancer-II [31], and high-grade Glioma data [32] 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 an overview of the characteristics of five datasets for the five binary cancer classification problems.

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, it is divided into training and test set, further independent component analysis is performed to reduce the dimensionality of train data. For ICA, the FastICA algorithm software package for Matlab (R2010a) is applied. Then t-test is used to rank the genes of independent component feature vectors. For validation, the data are classified with these reduced numbers of features by using the SVM classifier. The classifier and feature selection method was implemented with MATLAB™ software.

Table 1- Summary of five high dimensional biomedical microarray Datasets (Kent ridge online repository)

Data set	No. of Classes	No. of Features	No. of Samples	(+\-)
Colon cancer [28]	2	2000	62	(40\22)
Acute leukemia [29]	2	7129	72	(47\25)
Prostate tumor [30]	2	12600	102	(52\50)
High-grade Glioma [32]	2	12625	64	(35\29)
Lung cancer II [31]	2	12533	181	(31\150)

Experimental Result

To check the performance of the proposed approach with SVM 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 show the efficiency and feasibility of our proposed method, the results of the other five methods with the same classifier are also listed in [Table-2], [Table-3], [Table-4], [Table-5] for comparison. We use 4 kernels of the SVM classifier to check the performance of SVM with our proposed methods using five DNA microarray datasets. [Table-2], [Table-3], [Table-4], [Table-5] shows the classification accuracy of SVM using Linear, Polynomial, Radial basis function and Sigmoid Kernels with each datasets. In method 1, the microarray data are classified by SVM directly with all features. In Method 2, the features are selected by t-test for classification. In the Method 3, all the features are extracted by principle component analysis and the same is applied for method 4 except using ICA for feature extraction. In Method 5 and 6 t-test is used to rank the ICA and PCA features vector for SVM classification.

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 (xi) 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 yi, is the true class label, for instance xi, and

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

It can be seen from [Table-2], [Table-3], [Table-4], [Table-5] that t-test+PCA and t-test+ICA perform better than PCA and ICA in microarray data analysis, which demonstrates the effectiveness of the proposed approach.

Table 2- Classification Accuracy (CA) rates and variance (V) in (%) on five data set with different genes selection method using a linear kernel function with SVM classifier.

SNo	Datasets	Feature selection method	Mean Accuracy	Error	Variance
1	Colon cancer	1.SVM	88.29	11.71	0.071
		2.t-test +SVM	75.23	24.77	0.069
		3. PCA+SVM	75.15	24.85	0.061
		4. ICA+SVM	78.09	21.91	0.059
		5. PCA+t-test+SVM	82.44	17.56	0.042
		6. ICA+t-test+SVM	89.09	10.91	0.036
2	Acute leukemia	1.SVM	89.21	10.79	0.072
		2.t-test +SVM	84.32	15.68	0.066
		3. PCA+SVM	76.02	23.98	0.059
		4. ICA+SVM	86.13	13.87	0.042
		5. PCA+t-test+SVM	89.23	10.77	0.036
		6. ICA+t-test+SVM	92.28	7.72	0.023
3	Prostate tumor	1.SVM	78.26	21.74	0.106
		2.t-test +SVM	79.53	20.47	0.096
		3. PCA+SVM	73.23	26.77	0.098
		4. ICA+SVM	79.88	20.12	0.089
		5. PCA+t-test+SVM	82.13	17.87	0.078
		6. ICA+t-test+SVM	86.12	13.88	0.039
4	High-grade Glioma	1.SVM	69.93	30.07	0.068
		2.t-test +SVM	70.12	29.88	0.069
		3. PCA+SVM	69.62	30.38	0.052
		4. ICA+SVM	70.23	29.77	0.046
		5. PCA+t-test+SVM	72.32	27.68	0.045
		6. ICA+t-test+SVM	76.21	23.79	0.038
5	Lung cancer II	1.SVM	75.21	24.79	0.084
		2.t-test +SVM	74.33	25.67	0.079
		3. PCA+SVM	74.2	25.8	0.071
		4. ICA+SVM	78.12	21.88	0.081
		5. PCA+t-test+SVM	82.21	17.79	0.052
		6. ICA+t-test+SVM	89.21	10.79	0.034

Table 3- Classification Accuracy(CA) rates and variance(V) in (%) on five data set with different genes selection method using polynomial kernel function with SVM classifier.

SNo	Datasets	Feature selection method	Mean Accuracy	Error	Variance
1	Colon cancer	1.SVM	87.71	12.29	0.056
		2.t-test +SVM	77.51	22.49	0.049
		3. PCA+SVM	75.23	24.77	0.054
		4. ICA+SVM	77.99	22.01	0.063
		5. PCA+t-test+SVM	83.34	16.66	0.038
		6. ICA+t-test+SVM	90.01	9.99	0.022
2	Acute leukemia	1.SVM	91.21	8.79	0.071
		2.t-test +SVM	86.91	13.09	0.076
		3. PCA+SVM	78.97	21.03	0.054
		4. ICA+SVM	87.33	12.67	0.049
		5. PCA+t-test+SVM	89.33	10.67	0.037
		6. ICA+t-test+SVM	91.2	8.8	0.019
3	Prostate tumor	1.SVM	78.13	21.87	0.104
		2.t-test +SVM	80.82	19.18	0.981
		3. PCA+SVM	74.53	25.47	0.106
		4. ICA+SVM	79.99	20.01	0.099
		5. PCA+t-test+SVM	81.39	18.61	0.079
		6. ICA+t-test+SVM	86.22	13.78	0.049
4	High-grade Glioma	1.SVM	68.33	31.67	0.065
		2.t-test +SVM	69.68	30.32	0.056
		3. PCA+SVM	69.92	30.08	0.049
		4. ICA+SVM	70.19	29.81	0.047
		5. PCA+t-test+SVM	72.36	27.64	0.046
		6. ICA+t-test+SVM	76.92	23.08	0.039
5	Lung cancer II	1.SVM	76.01	23.99	0.076
		2.t-test +SVM	81.82	18.18	0.069
		3. PCA+SVM	75.43	24.57	0.079
		4. ICA+SVM	78.88	21.12	0.089
		5. PCA+t-test+SVM	83.44	16.56	0.066
		6. ICA+t-test+SVM	89.44	10.56	0.027

As for the comparison between the former two classification rules, t-test+ICA perform obviously better than t-test+PCA in terms of classification accuracy. It is clear that the classification accuracy of SVM with our proposed method compared to other five gene selection methods with same classifiers is more accurate, feasible and reduces the variation of classification performance. From the accuracy table of different Kernel function 2-5 with five datasets, the performance of the proposed method is better as compared with the all other 5 methods simultaneously the results are much better with RBF Kernel function compared with other three Kernels. So, the proposed approach improves the classification performance of the SVM classifier for microarray data.

[Fig-4], [Fig-5], [Fig-6], [Fig-7] shows the graph of the average error rate of the SVM classifier with four Kernel function for the five datasets with different gene selection methods. It clearly shows from the figures that SVM classifier with RBF kernel performs better than other kernel function because of the reduced error rate. It is evident from the graph that when we use top ranked genes based on t-test from PCA then the percentage error rate is minimized, so the PCA+t - test method performs better than PCA method with SVM classifier. In [Fig-8] the proposed method ICA + t-test with SVM gives the minimized error rate, which shows the significance of the proposed method with the other existing methods.

Table 4- Classification Accuracy (CA) rates and variance (V) in (%) on five data set with different genes selection method using RBF kernel function with SVM classifier.

SNo	Datasets	Feature selection method	Mean Accuracy	Error	Variance
1	Colon cancer	1. SVM	89.17	10.83	0.051
		2.t-test +SVM	82.11	17.89	0.05
		3. PCA+SVM	74.13	25.87	0.052
		4. ICA+SVM	78.19	21.81	0.062
		5. PCA+t-test+SVM	84.44	15.56	0.036
		6. ICA+t-test+SVM	91.09	8.91	0.021
2	Acute leukemia	1.SVM	91.21	8.79	0.071
		2.t-test +SVM	87.21	12.79	0.067
		3. PCA+SVM	77.17	22.83	0.044
		4. ICA+SVM	88.33	11.67	0.039
		5. PCA+t-test+SVM	90.13	9.87	0.031
		6. ICA+t-test+SVM	93.2	6.8	0.017
3	Prostate tumor	1.SVM	78.43	21.57	0.102
		2.t-test +SVM	81.22	18.78	0.991
		3. PCA+SVM	75.43	24.57	0.101
		4. ICA+SVM	80.45	19.55	0.092
		5. PCA+t-test+SVM	82.23	17.77	0.076
		6. ICA+t-test+SVM	87.12	12.88	0.043
4	High-grade Glioma	1.SVM	69.23	30.77	0.067
		2.t-test +SVM	74.65	25.35	0.054
		3. PCA+SVM	69.72	30.28	0.042
		4. ICA+SVM	70.21	29.79	0.043
		5. PCA+t-test+SVM	72.32	27.68	0.047
		6. ICA+t-test+SVM	77.21	22.79	0.041
5	Lung cancer II	1.SVM	76.21	23.79	0.074
		2.t-test +SVM	83.22	16.78	0.065
		3. PCA+SVM	75.23	24.77	0.081
		4. ICA+SVM	79.12	20.88	0.091
		5. PCA+t-test+SVM	84.21	15.79	0.062
		6. ICA+t-test+SVM	90.23	9.77	0.024

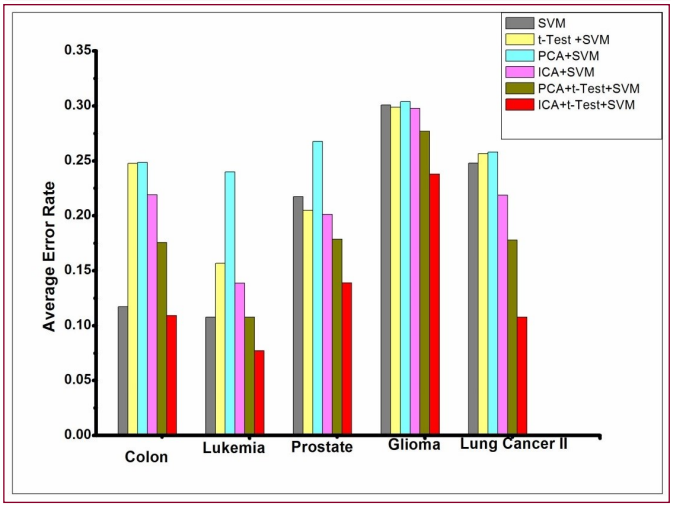

Fig. 4- Average error rate of classifier for five datasets with different gene selection methods Using Linear Kernel.

Therefore, with this proposed approach, discarding redundant, noise-corrupted or unimportant genes, we can reduce the dimensionality of any type of microarray data to speed up the classification process of SVM, increase the accuracy rate of the classification and making the computational expenses affordable.

Since, a small number of features are not enough for classification, while a large number of features may add noise and cause over fitting, we used t-test to rank the ICA feature vectors and the termination criterion in our method is based on the classification rate of the classifier.

Table 5- Classification Accuracy (CA) rates and variance (V) in (%) on five data set with different genes selection method using sigmoid kernel function with SVM classifier.

SNo	Datasets	Feature selection method	Mean Accuracy	Error	Variance
1	Colon cancer	1.SVM	88.19	11.81	0.061
		2.t-test +SVM	81.33	18.67	0.051
		3. PCA+SVM	75.15	24.85	0.053
		4. ICA+SVM	79.19	20.81	0.052
		5. PCA+t-test+SVM	82.34	17.66	0.032
		6. ICA+t-test+SVM	90.09	9.91	0.026
2	Acute leukemia	1.SVM	92.21	7.79	0.071
		2.t-test +SVM	89.23	10.77	0.065
		3. PCA+SVM	76.67	23.33	0.054
		4. ICA+SVM	88.23	11.77	0.039
		5. PCA+t-test+SVM	91.23	8.77	0.03
		6. ICA+t-test+SVM	92.99	7.01	0.013
3	Prostate tumor	1.SVM	78.43	21.57	0.102
		2.t-test +SVM	81.23	18.77	0.093
		3. PCA+SVM	75.43	24.57	0.101
		4. ICA+SVM	80.45	19.55	0.092
		5. PCA+t-test+SVM	83.23	16.77	0.076
		6. ICA+t-test+SVM	88.12	11.88	0.043
4	High-grade Glioma	1.SVM	69.21	30.79	0.067
		2.t-test +SVM	71.34	28.66	0.051
		3. PCA+SVM	69.72	30.28	0.039
		4. ICA+SVM	70.21	29.79	0.041
		5. PCA+t-test+SVM	73.32	26.68	0.043
		6. ICA+t-test+SVM	75.66	24.34	0.042
5	Lung cancer II	1.SVM	76.34	23.66	0.084
		2.t-test +SVM	79.99	20.01	0.054
		3. PCA+SVM	75.69	24.31	0.079
		4. ICA+SVM	80.12	19.88	0.091
		5. PCA+t-test+SVM	83.21	16.79	0.064
		6. ICA+t-test+SVM	89.46	10.54	0.029

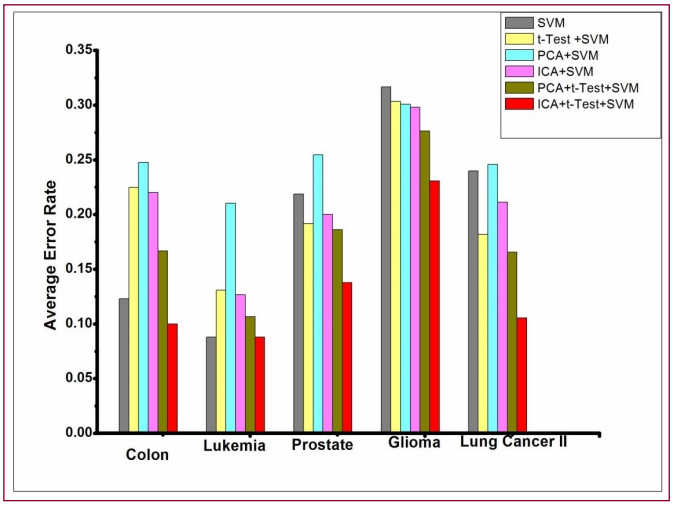

Fig. 5- Average error rate of classifier for five datasets with different gene selection methods using Polynomial Kernel.

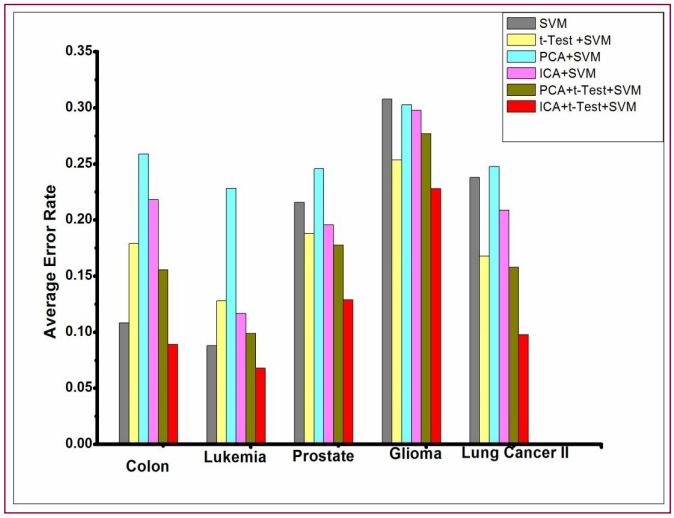

Fig. 6- Average error rate of classifier for five datasets with different gene selection methods using RBF Kernel.

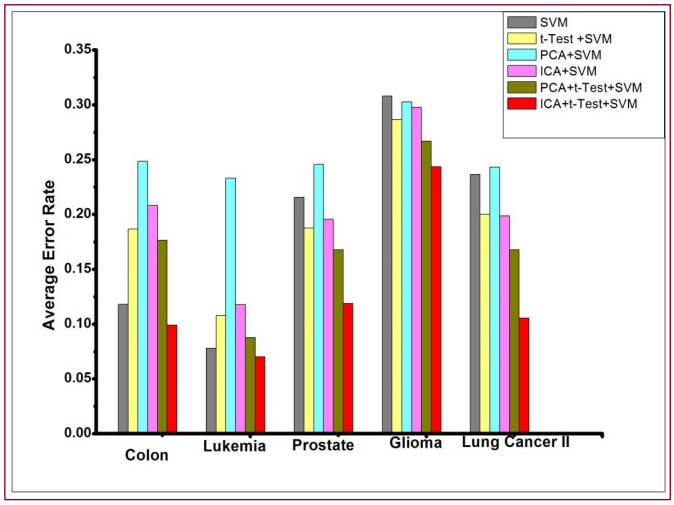

Fig. 7- Average error rate of classifier for five datasets with different gene selection methods using Sigmoid Kernel.

Table 6- Classification accuracy rates using RBF Kernel function with different numbers of genes for 5 datasets

SNo	Data sets	20	30	40	50	60	70	80
1	Colon cancer	86.04	91.09	87.11	82.91	79.19	_	_
2	Acute leukemia	91.55	93.2	92.11	92.07	91.03	88.77	_
3	Prostate tumor	81.23	85.17	86.22	88.12	86.01	84.87	84.05
4	High-grade Glioma	77.21	75.09	75.22	72.05	70.99	_	_
5	Lung cancer II	84.44	86.56	88.23	89.89	90.23	88.63	87.33

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 SVM classification, a graph is plotted between the number of genes and classification accuracy rates.

[Fig-8] shows the graph between the number of selected genes and the classification accuracy, using SVM classifier with RBF Kernel for five data sets based on the proposed gene selection method. The t-test technique is used to rank the independent components feature vector. With the help of top ranked gene, 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 five data sets. As shown in [Table-6], with five data sets using SVM classifier with RBF Kernel with ICA feature vector, the highest mean accuracy obtained was 78.19%, 88.33%, 80.45 %, 70.21% and 79.12 % respectively. When the t - test is used to rank independent component feature vector, one managed to get 91.09%, 93.20 %, 87.12%, 77.21% and 90.23 % mean classification accuracies with 30, 30, 50, 20 and 60 genes respectively for SVM classifier with RBF Kernel. These results clearly show that the t-test approach with ICA performs better than the other existing methods.

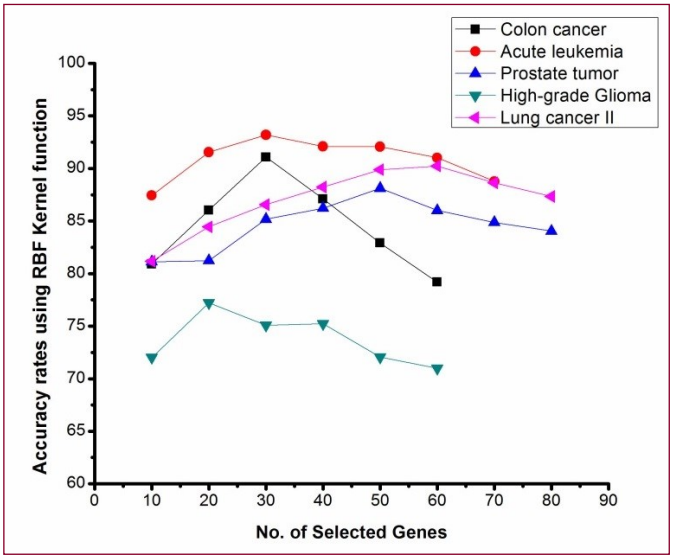

Fig. 8- Number of selected genes V/s Classification accuracy using SVM classifier with RBF Kernel on five datasets, based on proposed methods.

Conclusion

This paper presents a t-test based feature selection approach in ICA feature vector for SVM classification of microarray data where the methodologies involve dimension reduction of microarray data using ICA, followed by the feature ranking using t-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 SVM 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 SVM 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 SVM classifier.

Acknowledgment: The author would like to acknowledge the support of the Director (Dr. Appu Kuttan 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.

Conflicts of Interest: None declared.

References

[1] Peng Y. (2006) *Computers in Biology and Medicine*, 36(6), 553 -573.

[2] Vilda P.G., Díaz F., Martínez R., Malutan R., Rodellar V. & Puntonet C.G. (2006) *Robust preprocessing of gene expression microarrays for independent component analysis*, Independent Component Analysis and Blind Signal Separation, Springer, 714 -721.

[3] Debnath R. & Kurita T. (2010) *Biosystems*, 100(1), 39-46.

[4] Kar S., Sharma K.D. & Maitra M. (2015) *Expert Systems with Applications*, 42(1), 612-627.

[5] Gutkin M. (2008) *Feature selection methods for classification of gene expression profiles*, Tel-Aviv University.

[6] Ammu P. & Preeja V. (2013) *International Journal of Computer Applications*, 61(12), 39-44.

[7] Du D., Li K., Li X. & Fei M. (2014) *Neurocomputing*, 133, 446-458.

[8] Bartenhagen C., Klein H.U., Ruckert C., Jiang X. & Dugas M. (2010) *BMC Bioinformatics*, 11(1), 567.

[9] Frigyesi A., Veerla S., Lindgren D. & Höglund M. (2006) *BMC Bioinformatics*, 7(1), 290.

[10] Zheng C.H., Huang D.S., Kong X.Z. & Zhao X.M. (2008) *Genomics, Proteomics & Bioinformatics*, 6(2), 74-82.

[11] Statnikov A., Henaff M., Narendra V., Konganti K., Li Z., Yang L., Pei Z., Blaser M.J., Aliferis C.F. & Alekseyenko A.V. (2013). *Microbiome*, 1(1), 11.

[12] Mohan A., Rao M.D., Sunderrajan S. & Pennathur G. (2014) *Interdisciplinary Sciences: Computational Life Sciences*, 6(3), 176-186.

[13] Anand A. & Suganthan P. (2009) *Journal of Theoretical Biology*, 259(3), 533-540.

[14] Chakraborty S. & Guo R. (2011) *Computational Statistics & Data Analysis*, 55(3), 1342-1356.

[15] Hyvärinen A. & Oja E. (1997) *Neural Computation*, 9(7), 1483-1492.

[16] Naik G.R. & Kumar D.K. (2011) *Informatica: An International Journal of Computing and Informatics*, 35(1), 63-81.

[17] Engreitz J.M., Daigle B.J., Marshall J.J. & Altman R.B. (2010) *Journal of Biomedical Informatics*, 43(6), 932-944.

[18] Kong W., Vanderburg C.R., Gunshin H., Rogers J.T. & Huang X. (2008) *Biotechniques*, 45(5), 501.

[19] Hyvärinen A., Karhunen J. & Oja E. (2004) *Independent component analysis*, John Wiley & Sons.

[20] Capobianco E. (2004) *Exploration and reduction of high dimensional spaces with independent component analysis.*

[21] Chu F. & Wang L. (2005) *International Journal of Neural Systems*, 15(06), 475-484.

[22] Yang T., Kecman V., Cao L. & Zhang C. (2010) *Combining support vector machines and the t-statistic for gene selection in DNA microarray data analysis*, Advances in Knowledge Discovery and Data Mining, Springer, 55-62.

[23] Sreekumar J. & Jose K. (2008) *Indian Journal of Biotechnology*, 7(4), 423-436.

[24] Huang H.L. & Chang F.L. (2007) *Biosystems*, 90(2), 516-528.

[25] Kostka P.S. & Tkacz E.J. (2008) *Feature extraction based on time-frequency and Independent Component Analysis for improvement of separation ability in Atrial Fibrillation detector*, 30th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society.

[26] Hsu C.C., Chen M.C. & Chen L.S. (2010) *Computers & Industrial Engineering*, 59(1), 145-156

[27] Durgesh K.S. & Lekha B. (2010) *Journal of Theoretical and Applied Information Technology*, 12(1), 1-7.

[28] Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D. & Levine A.J. (1999) *Proceedings of the National Academy of Sciences*, 96(12), 6745-6750.

[29] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D. & Caligiuri M.A. (1999) *Science*, 286 (5439), 531-537.

[30] Singh D., Febbo P.G., Ross K., Jackson D.G., Manola J., Ladd C., Tamayo P., Renshaw A.A., D'Amico A.V., Richie J.P., Lander E.S., Loda M., Kantoff P.W., Golub T.R. & Richie J.P. (2002) *Cancer Cell*, 1(2), 203-209.

[31] Gordon G.J., Jensen R.V., Hsiao L.L., Gullans S.R., Blumenstock J.E., Ramaswamy S., Richards W.G., Sugarbaker D.J. & Bueno R. (2002) *Cancer Research*, 62(17), 4963-4967.

[32] Nutt C.L., Mani D., Betensky R.A., Tamayo P., Cairncross J.G., Ladd C., Pohl U., Hartmann C., McLaughlin M.E., Batchelor T.T., Black P.M., von Deimling A., Pomeroy S.L., Golub T.R., & Batchelor T.T. (2003) *Cancer Research*, 63(7), 1602-1607.