International Journal of ChemTech Research
CODEN (USA): IJCRGG, ISSN: 0974-4290, ISSN(Online):2455-9555
Vol.10 No.14, pp 190-197, 2017

A Novel Feature Selection Algorithm for Dimensionality Reduction in Microarray Datasets

A.K. Shafreen Banu[1]*, S. Hari Ganesh[2]

[1]Dept. of Information Technology,Bishop Heber College (Autonomous),Tiruchirappalli.620017, India
[2] Dept. of Computer Science, H.H The Rajah's College, Pudukottai-622 001, India

Abstract : Dimensionality reduction is one of the vital and challenging tasks of feature selection techniques in data mining, as it requires an intrinsic analysis of data distribution with respect to class label. Despite the *non-linear* distribution of data attributes, the *linear attributes* have gained more attention by the researchers as it could build an effective knowledge prediction model with maximized accuracy. The objective of this paper is to propose another *feature selection* algorithm that is designed to process linear data attributes for reducing the *dimensions* of *microarray* datasets. The algorithm is also to be well compared with the existing algorithms to prove its efficacy in terms of usefulness.
Keywords : microarray, feature selection, high dimensional, non-linear and linear attributes.

Introduction

The enormous growth of communication and information technologies has led the world be drowned with lots of data that contains hidden potential facts.The size of the data may vary from lower to higher dimensions. Effective manipulation of high dimensional data has always been a challenging task since the manual processing of voluminous data is highly impractical. The art of dimension reduction is the one and only concern of feature selection techniques of data mining that encompasses a collection of statistical and machine learning algorithms that is intended to discover meaningful attributes from the large sets of data. In general a dataset may either be linear or non-linear, where the linear dataset consists of high the quotient of correlated attributes through which the class attributes can be easily separable, whereas non-linear datasets, in contrast does not have linear relationship between the attributes. The extraction of linear attributes is under the scope of regression, correlation and machine learning algorithms. On the other hand, the extraction of meaningful attributes can be accomplished using MDS and ISOMAP algorithms.

The objective of feature selection techniques is to extract the most meaningful attributes from both linear and non-linear datasets. The output produced by the feature selection techniques is analyzed by the classification algorithms so as to predict the classes of unknown data that aims at producing highest prediction accuracy. Though there has been numerous research works carried out in feature selection, there is no single algorithm works well for selecting best features that produce effective results. Thus the aim of this paper is to propose a novel feature selection algorithm that processes high dimensional data to select linear data attributes and to transform non-linear data into linear form through PCA. The features that are selected with PCA are executed upon Support Vector Machine, yet another linear classifier to identify the hidden classes of the dataset. The remaining section of the paper is organized as follows: Section 2 consists of review of literature, section 3 describes the methodology, section 4 explains the experimentation and result discussions and finally section 5 concludes the findings of the research.

Review of Literature

Malhi and Gao[1] have presented a feature selection scheme based on principal component analysis method for machine defect classification. The authors have considered two scenarios for identifying the severity level of bearing defects, when no prior knowledge on the defect conditions is available. The first scenario of the proposed work is the supervised training, where the applicability of PCA to select suitable features as input to feed forward networks and radial basis functions are investigated for defect classification. The second scenario is the unsupervised training that computes the most sensitive features from the vibration signals of defective bearings which are identified based on the defect conditions. The features are then used as inputs for unsupervised competitive learning scheme to classify the defective bearing based on the size of the defect. The authors have claimed that the results obtained by the PCA are convincing and applicable to wide range of problems.

Lu et al.[2] have introduced a dimensionality reduction method called, PFA (Principal Feature Analysis) that chooses the feature subset that contains the essential information from original dataset. The authors have trained the advanced statistical modeling and optimization techniques Space tracking and content based image retrieval problems to obtain the principal features. The authors have claimed that, unlike PCA, PFA does require only fewer sensors or fewer features to compute and the selected features that have original physical meaning and stated that PFA features are ranked as the top 5% of all possible combinations of the experiment. The authors have also claimed that the proposed PFA obtains highest accuracy.

Dang et al.[3] have demonstrated a framework for selecting good feature subsets from all the principle components that enhances the prediction accuracy of gene expression in microarray data. The authors have employed PCA for dimension reduction, decision tree for feature selection and multilayer perceptron for classification. The authors have stated that the proposed method improves the performance on the gene expression of micro array data in terms of accuracy and denoted that not all the top eigenvectors of PCA are meaningful for classification. Therefore, PCA should be collaborated with feature selection or feature extraction for dimension reduction for analyzing the problems with high dimensionality.

Inan et al.[4] have presented a hybrid feature selection method combining the association rules and PCA with artificial neural network classifier for diagnosing the breast cancer disease. The authors have executed the Apriori algorithm as feature selection to analyze all the inputs and the attributes that are significantly least are eliminated. The PCA algorithm is then employed over the filtered attributes of Apriori algorithm to obtain the most meaningful data attributes. The output of PCA is applied to multi-layered feed-forward back-propagation neural network. The authors have stated that the hybrid feature selection method with feed-forward neural network classifier has achieved highest accuracy than the state of art methods.

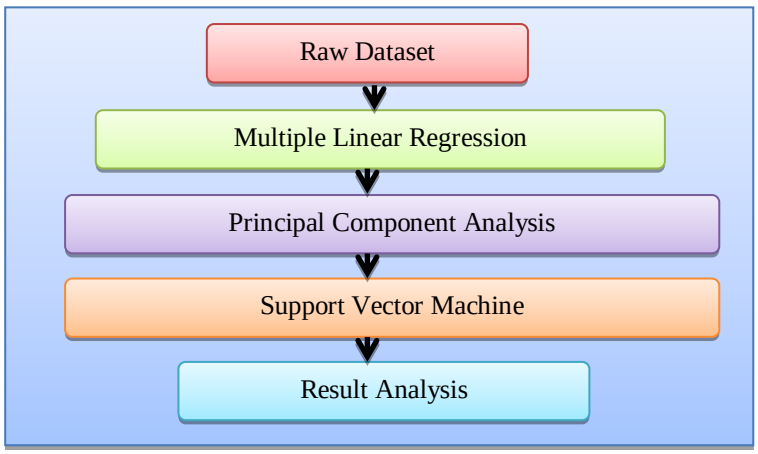
Yuce et al.[5] have integrated PCA and ANN, to improve the quality control for the identification of wood veneer defects. The method consists of the procedure that identifies the principal components from the overall attributes with the objective of detecting the defects in quality control with minimal error and maximized accuracy by reducing the number of inputs to be given to the ANN. The authors have claimed that the best performance with one hidden layer is found to be with more than 10 neurons and the reduction of features reduces the number of iterations as PCA reduces training time and increases the testing performance. In addition, a reduction of 61 epochs increased the quality of outputs in testing stage by 18%.

Morchid et al[.6] have analyzed the specific tweet features through principal component analysis to understand the behavior of highly forwarded tweets in contrast to those retweeted only few times. The authors have also proposed a method that automatically detects the massively retweeted messages with the objective to select the best features allowing the best classification performance. The authors have insisted upon the selection of best correlated features for increasing the prediction accuracy. The authors have claimed that the choice of most relevant features has a real impact on the massive retweet detection with the gain of 2.4% with SVM and 0.7% with Naïve Bayes algorithms on the F-measure using the correlated features.

Methodology

The proposed novel feature selection algorithm MLR-PCA combines the essentials concepts of Multiple Linear Regression (MLR) and Principal Component Analysis (PCA) for selecting the meaningful features. The raw dataset is first inputted into MLR for understanding the hidden linear relationship between the

attributes and the class variable. The dataset is then transformed into a linear form as directed by the coefficients of MLR. The linearly transformed raw dataset is processed by PCA algorithm for selecting the best subset of features. The attributes that are suggested by PCA is then classified using Support Vector Machine (SVM) classifier to evaluate the prediction accuracy. The architecture of the proposed methodology is shown in Fig.(1).



Multiple Linear Regressions

MLR is one of the most common and frequently used linear regression analysis techniques which explain the relationship between a dependent and one or more independent continuous and categorical variables [7]. MLR centers around the task of fitting a single line through a scatter plot. In a single linear regression model, a dependent variable Y is related to a single independent variable which can be denoted using the equation 1.

$$E(Y|X) = \alpha + \beta X \quad \ldots (1)$$

But, most of the real time problems consist of more than one independent variable. This leads to the derivation of multiple linear regression model which is denoted using the equation 2.

$$E(Y|X) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad \ldots (2)$$

Where,

α denotes the line intercept , β is called the slopes or coefficients.

Principal Component Analysis

PCA examines the interrelations among a set of variables in order to understand the itsunderlying structure. The objective of PCA is to explain the maximum amount of variance with the fewest number of principal components [8]. Consider a data matrix shown in equation 3,

$$A = \{a_{ij}\} \in R^{n \times p} \quad \ldots (3)$$

Where n is the number of rows and p is the number of columns. PCA is mathematically denoted as orthogonal linear transformation which transforms the data to a new form in such a way that the greatest variance by any projection of data come to lie on the first principle component and so on. The raw data may change with number of intervals to standardize the values of A as shown in equations 4, 5 and 6.

$$\hat{A} = \{\hat{a}_{ij}\} \quad \ldots (4)$$

With

$$\hat{a}_{ij} = \frac{a_{ij} - f_i}{\sqrt{n}} \quad \ldots (5)$$

$$f_i = \frac{1}{n}\sum_{i=1}^{n} a_{ij} \quad \ldots (6)$$

Where, f_i is the average value of j^{th} column of A denoted by

Then, $\hat{A}$ is used to compute the covariance matrix of MLR dataset which can be denoted using the equation 7.

$$S = \hat{A}^T . \hat{A} \quad \ldots (7)$$

Where $S \in R^{p\times p}$ with a matrix size p×p. PCA then calculates the eigenvalues and eigenvectors and sorts them in descending order. Supposing the p eigenvalues of S are $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots . \geq \lambda_p$ and p eigenvectors are denoted as t1,t2,t3…tp. Then the axis of the new space eigenvector is ti. If the first k-dimensions are used then the new matrix created from the eigenvectors is as shown in equation 8.

$$T = [t_1|t_2|t_3|\ldots|t_k] \in R^{p\times k} \quad \ldots (8)$$

And the co-ordinates are denoted as shown in equation 9.

$$C = \hat{A}.T \quad \ldots (9)$$

Support Vector Machine

SVM is a popular and widely used algorithm for classification and regression tasks as it process the high dimensional data and show good generalization behavior. SVMs are computed by solving quadratic programming problems. The method has its foundation in classification and has later been extended to regression. SVM trains the dataset with n samples, denoted as $\{x_i, y_i\}_{i=1}^{n}$, where xi is an input and yi is an output, and yi∈-1, +1}to find the optimal classes separation hyperplane , which is given by f(xi)=wφ(xi)+b. where w is the optimal set of weights and b is optimal bias, and the φ is the non-linear mapping applied to input vectors [9]. SVM optimizes the hyperplane by maximizing the distance between the hyperplane and its closest data points.

Illustration

This section illustrates the understanding of the proposed method using iris dataset [10]. The dataset has four attributes and 150 instances which contains the three classes of 50 instances each, where each class refers to a type of iris plant. Among the three, one class is linearly separable from the other two. The illustration is performed using Weka 3.6 data mining tool. The transformation of raw dataset into linear data is done in MS-Excel and converted as .CSV extension file format. Figure (2)explains the execution of iris original dataset with PCA and SVM on Weka tool. Figure (2.a) denotes the loading of original iris dataset on to Weka, fig(2.b) depicts the values of original iris dataset, fig(2.c) shows the execution of PCA on to original iris dataset and finally fig (2.d) depicts the execution SVM on the principle attributes suggested by PCA. As PCA chooses the first two attributes as principal components for the classifying the iris plants, the first two attributes Sepal length and Sepal width alone are used for the classification with SVM classifier. Out of 150 instances, SVM correctly classified 121 instances and 29 are incorrectly classified. Thus, the accuracy obtained by PCA with original dataset is 81%. In addition, the taken to build the model is 0.9 seconds.

Figure.2.a). Loading of Iris Original

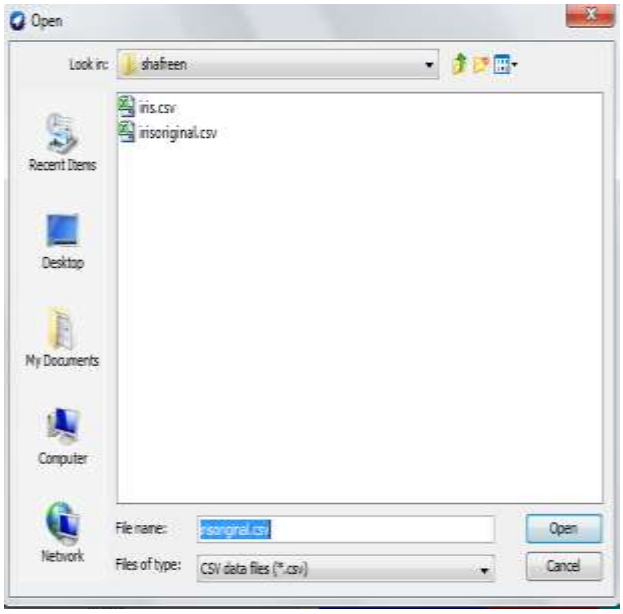


Figure 2.b). Data Viewer – Iris Original

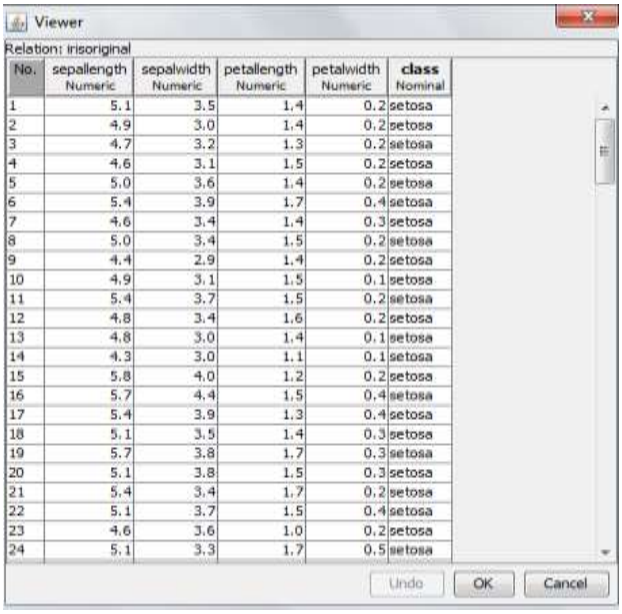


Figure 2.c). PCA Execution – Iris Original

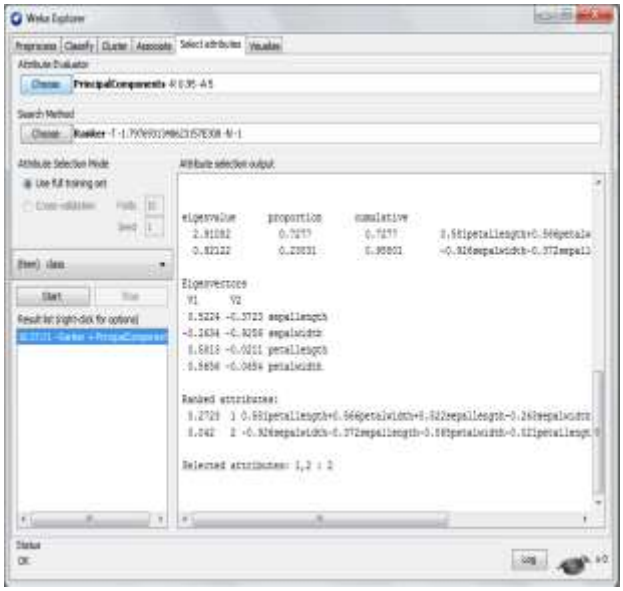
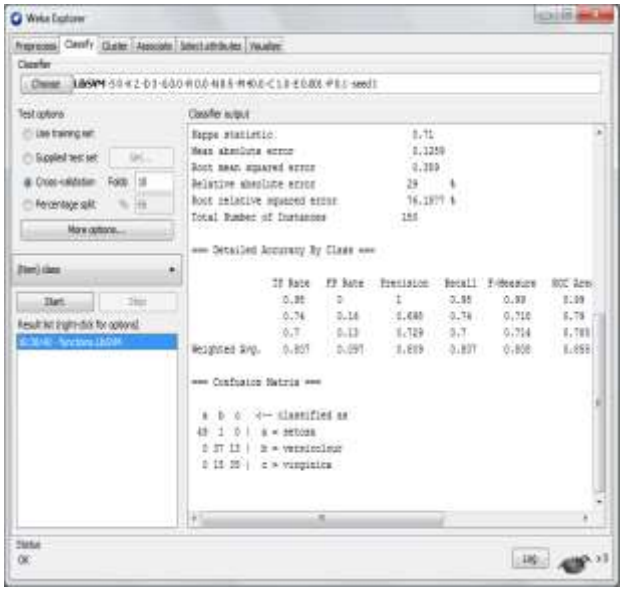


Figure.2.d). SVM Execution – Iris Original



Fig(3).describes the execution of linearly transformed Iris dataset using MLR technique with PCA and SVM classifier. Figure (3.a) denotes the loading of linear iris dataset on to Weka, fig(2.b) depicts the values of linear iris dataset, fig (2.c) shows the execution of PCA on to linear iris dataset and finally fig (2.d) depicts the execution SVM on the principle attributes suggested by PCA. As PCA chooses the first three attributes as principal components for the classifying the iris plants, the first three attributes Sepal length, Sepal width and petal length are used for the classification with SVM classifier. Out of 150 instances, SVM correctly classified 145 instances and only 5 instances are incorrectly classified. Thus, the accuracy obtained by PCA with original dataset is 97%. In addition, the taken to build the model is 0.06 seconds. Thus, the results obtained by the MLR-PCA outperform the traditional PCA in terms of time and accuracy. Table 1 denotes the comparative results of PCA and MLR-PCA.

Figure.3.a). Loading of Linear Iris

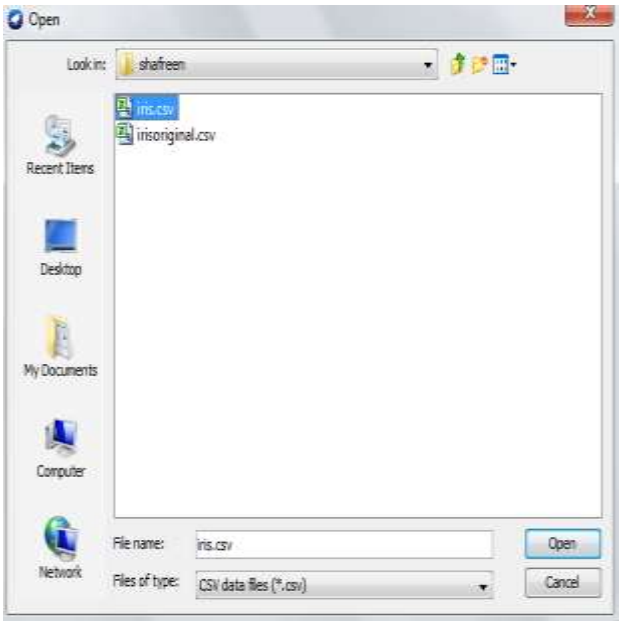


Figure. 3.b). Data Viewer – Linear Iris

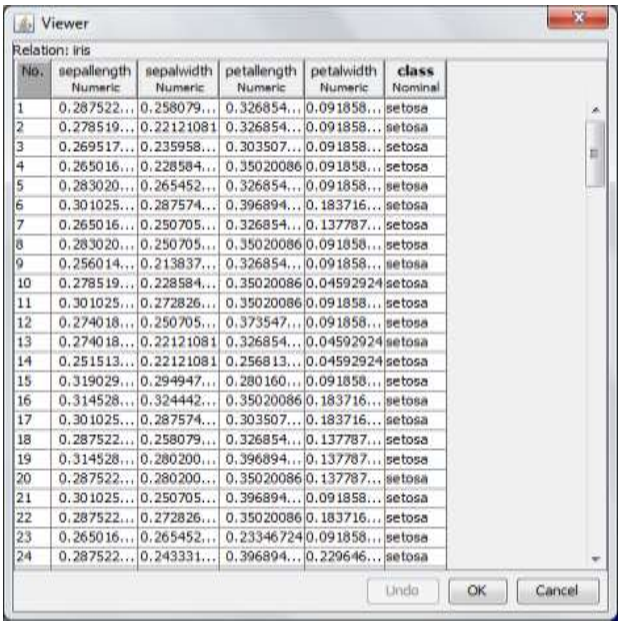


Figure.3.c). PCA Execution – Linear Iris

Figure.3.d). SVM Execution – Iris Original

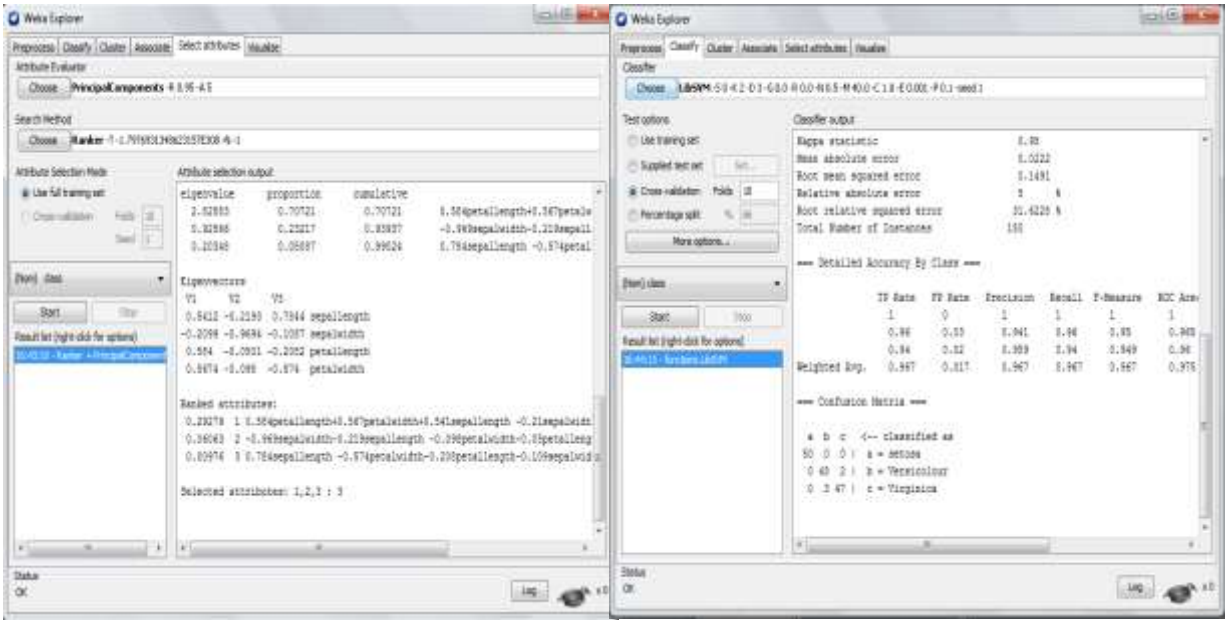


Table 1. Comparative results of PCA Vs MLR - PCA

Algorithm	Time	Accuracy	Precision	Recall
PCA	0.9 seconds	81%	1	0.98
MLR-PCA	0.06 seconds	97%	1	1

Experimentation and Result Discussions

The experimentation is performed over the three gene expression of microarray datasets on leukemia, Lymphoma and Colon, as objective of the proposed work is to reducing the dimensionality of the high dimensional microarray data. The datasets are downloaded from- http://www.ntu.edu.sg/home/elhchen/data.htm[10]. The description of the experimental datasets is explained in table (2). The graphical representation of the comparative results is shown in Fig.(4).

Table 2. Experimental Dataset Description

Name of the datasets	Number of Samples	Number of Genes	Number of classes	Description
Leukemia	72	7129	2	47 ALL and 25 AML
Lymphoma	62	4026	3	11 DLBCL, 42 FL, and 9 CLL
Colon	62	2000	2	40 Tumor, 22 Normal

Table 3. Results Summary of Experimental Datasets

Name of the datasets	Number of genes		Time (in sec)		Accuracy (in %)	
	PCA	**MLR-PCA**	**PCA**	**MLR-PCA**	**PCA**	**MLR-PCA**
Leukemia	153	112	413	258	78	83
Lymphoma	326	259	571	486	69	75
Colon	124	91	327	254	82	89

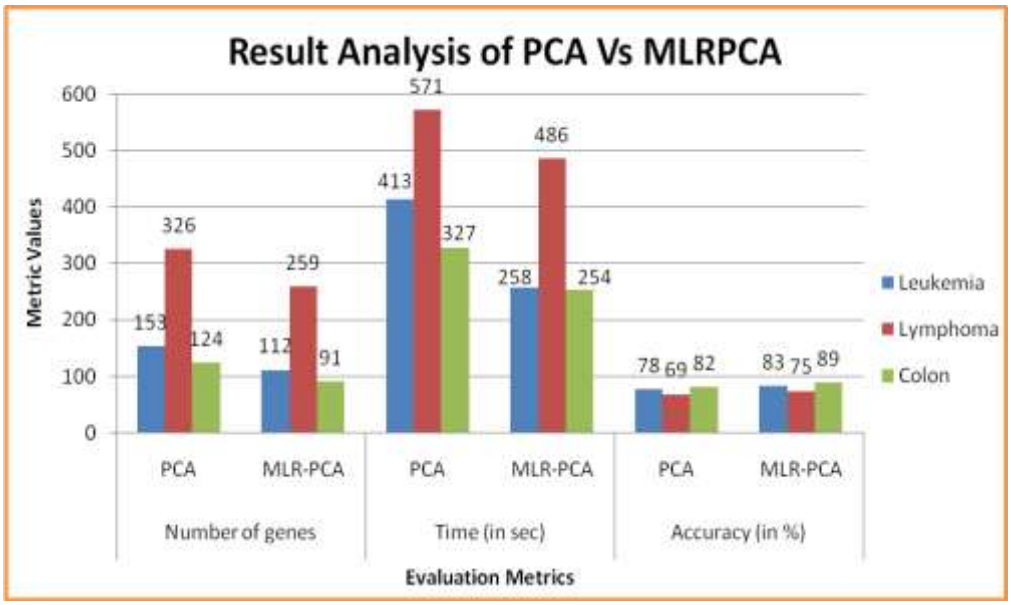


Figure. 4.Comparative Analysis of MLR-PCA with PCA

Conclusion

This paper presents MLR-PCA, a feature reduction method using linear regression concepts with the intension of reducing meaningful gene expressions in microarray datasets. The proposed MLR-PCA is a generic method which can be applied to any real time high dimensional datasets. The proposed algorithm is well demonstrated with four sample datasets and the results shows that the proposed method achieves better accuracy and minimized computational time with minimized attributesfor all four datasets. Thus, the method outperforms the state -of-the art PCA method. In future, the suitable linear function for SVM classifier has to be developed for processing linear data retrieved by the MLR-PCA.

References

1. Malhi, Arnaz, and Robert X. Gao. "PCA-based feature selection scheme for machine defect classification." IEEE Transactions on Instrumentation and Measurement 53, no. 6 (2004): 1517-1525.
2. Lu, Yijuan, Ira Cohen, Xiang Sean Zhou, and Qi Tian. "Feature selection using principal feature analysis." In Proceedings of the 15th ACM international conference on Multimedia, pp. 301-304.ACM, 2007.
3. Dang, Thuy Hang, Trung Dung Pham, HoaiLinh Tran, and Quang Le Van. "Using dimension reduction with feature selection to enhance accuracy of tumor classification."In Biomedical Engineering (BME-HUST), International Conference on, pp. 14-17.IEEE, 2016.
4. Inan, Onur, Mustafa SerterUzer, and NihatYilmaz. "A new hybrid feature selection method based on association rules and PCA for detection of breast cancer." International Journal of Innovative Computing, Information and Control 9, no. 2 (2013): 727-729.
5. Yuce, Baris, Ernesto Mastrocinque, Michael Sylvester Packianather, Duc Pham, Alfredo Lambiase, and Fabio Fruggiero. "Neural network design and feature selection using principal component analysis and Taguchi method for identifying wood veneer defects." Production & Manufacturing Research 2, no. 1 (2014): 291-308.
6. Morchid, Mohamed, Richard Dufour, Pierre-Michel Bousquet, Georges Linarès, and Juan-Manuel Torres-Moreno. "Feature selection using Principal Component Analysis for massive retweet detection." Pattern Recognition Letters 49 (2014): 33-39.
7. Christy, A. Joy, and S. Hari Ganesh. "Linear Regressive Percentage Split Distribution Clustering."I J C T A, 9(27), 2016, pp. 495-502.
8. Shlens, Jonathon. "A tutorial on principal component analysis." arXiv preprint arXiv:1404.1100 (2014).
9. Suthaharan, Shan. "Support Vector Machine." In Machine Learning Models and Algorithms for Big Data Classification, pp. 207-235. Springer US, 2016.
10. https://archive.ics.uci.edu/ml/machine-learning-databases/iris/
11. http://www.ntu.edu.sg/home/elhchen/data.htm
12. A.K.Shafreen Banu" A Study of Feature Selection Approaches/or Classification"978-1-4799-6818-3/15/$31.00 © 2015 IEEE
13. A.K.Shafreen Banu" Tumor Cells Classification –A Comparative Study" International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.85 (2015) © Research India Publications; http/www.ripublication.com/ijaer.htm pp797-800

*****