ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE) Vol. IV, Issue VII, July 2018

Cancer Data Classification using Clustering Techniques

P. Deepalakshmi[1] , Dr.V.Umadevi [2]

1. Research Scholar, Department of Computer Science, Jairams Arts and Science College, Karur.

2. Research Director, Department of Computer Science, Jairams Arts and Science College, Karur.

Abstract: In a multilayered feedforward network, neurons are organized into layers. The input layer is not fully composed of neurons, but rather it consists of some values in a data record, that constitutes inputs to the next layer of neurons. The next layer is called a hidden layer; there may be many hidden nodes. The concluding layer is the output layer, there is only one node for each class. A single forward pass through the network results in the assignment of a value to each output node, and the record is assigned to whichever classifications node had the highest value. Multilayer feedforward networks are trained using the Backpropagation (BP) learning algorithm. Backpropagation training algorithm when applied to a feedforward multilayer neural network then it is known as Backpropagation neural network. Functional signals flows in the forward path and error signals transmit in backward path. That's why it is Error Backpropagation or shortly backpropagation network. The activation function that can be differentiated (such as sigmoid activation function) is chosen for hidden and output layer computational neurons. The algorithm is based on an error-correction rule. Learning is based upon mean squared error and generalized delta rule. The rule applied for weight updation is generalized delta rule.

1. Introduction

Much research is being done in the academics as well as the industries towards the application of bioinformatics that uses computational approaches to solve biological problems. The goal of this field is to retrieve, analyze and interpret the vast and complex genomic data sets that are uncovered in large volumes of genes in molecular biology. Biological data mining posses various challenges like gene discovery, drug discovery, gene finding, revealing unknown relationship with respect to structure and function of genes to understand biological systems. This field faces demands for immediate prediction and classification due to the availability of DNA cancer data, structure information of proteins and microarray technology to provide dynamic information about thousand of genes in data. The aims of Bioinformatics are:

1. To organize data in a way that allows researcher and practitioners to access existing information and to submit new entries as they are produced.
2. To develop tools, software's and resources that aid in analyzing and management of data.

3. Use of this data is to analyze and interpret the results in a biologically meaningful manner.
4. To help practitioners in the pharmaceutical industry by understanding the microarray cancer data structures which helps makes the disease prediction easy.

2. The Algorithm

1. Initialization of weights (w) and biases (b) to random small values and target (t) is fixed.
2. Forward computation: Output of each layer is y = (wx b). Where w = synaptic weight, x = input and b = bias value. Output of input layer is the input of hidden layer. In this way actual output is calculated.
3. Error is calculated by the difference of target and the actual output at output layer of neuron. Error e = t y.

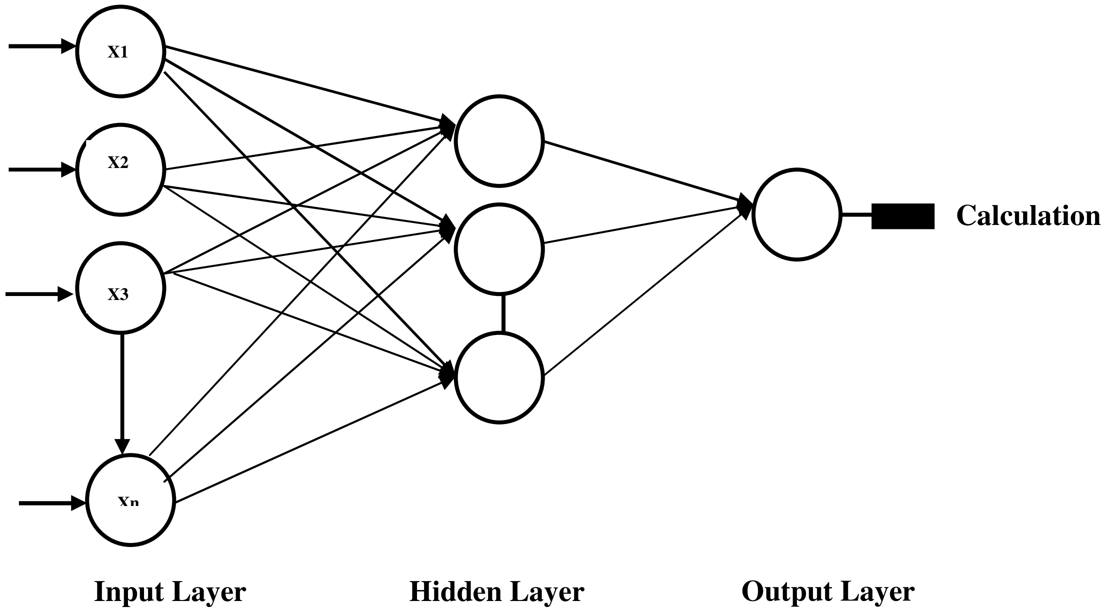


Figure 1: Multilayered Backpropagation Neural Network

4. Backward computation in NN: Each layer error is calculated by partial differentiation. For output layer error, e_0 = 0:5 (d (hidden) =dy (hidden)) e and For hidden layer error, e_h = (d (Y_{input}) =dY_{input}) $w_0$ut e_0.
5. Weights and biases in each layer are updated according to the computed errors. Updated weight, w_{new} = w_{old} lr e_{layer} x_{layer} layer. Updated bias, b_{new} = b_{old} lr e_{layer} layer where e_{layer} is the error of the particular layer and x_{layer} is the input that is fed to the layer and lr is the learning rate.
6. Step 2 to 5 is repeated until the acceptable minimized error.

3. BP Neural Network Classifier Hybrid with PCA Algorithm

Although back propagation is the most popular learning method in the neural network community, the drawbacks of it are often pointed out are:

1. Very slow computing speed
2. The possibility of getting trapped in local minima.
4. More hidden nodes lead to overwriting and greater capacity of assimilating data.
5. The convergence obtained from backpropagation learning is very slow.
6. The convergence in backpropagation learning is not guaranteed.

4. Why SVM for cancer classification

SVMs are used for cancer classification mainly due to following two reasons:

1. SVMs have demonstrated the ability not only to correctly separate the entities into appropriate classes, but also to identify instances whose established classification is not supports by the data.

2. SVM have many mathematical features that make them attractive for gene expression analysis, including their exibility in choosing a similarity function, sparseness of solution when dealing with the huge data sets, the capacity to hold that huge feature spaces, and the capacity to classify outliers.

5. The SVM Classifier and Kernel Selection

A support vector machine (SVM) [19] is a computer techniques used for the supervised learning process is to analyze and recognize patterns, are derived from statistical learning theory developed by Vladimir N.

Vapnik and Corinna Cortes in 1995. The goal of SVM is to produce a model (based on the training set) which predicts the target values of the test set making it as non-probabilistic linear classifiers. Viewing the input data as two sets of vectors in a d-dimensional space, an SVM constructs a separating hyperplane in that space, which maximizes the margin between the two classes of points. Instinctively, a superior partition is attained by the hyperplane that has the biggest distance to the neighboring data points of both classes. Larger margin or distance between these parallel hyperplanes indicates better generalization error of the classifier [19]. Implies that only support vectors machine matters and other training examples are ignorable.

The SVM is designed for binary-classification problems, assuming the data are linearly separable. Given the training data $(x_i; y_i)$; i = 1; 2::::m; x_i R^n; y_i f+1; 1g^t where, R^n : is the input space,

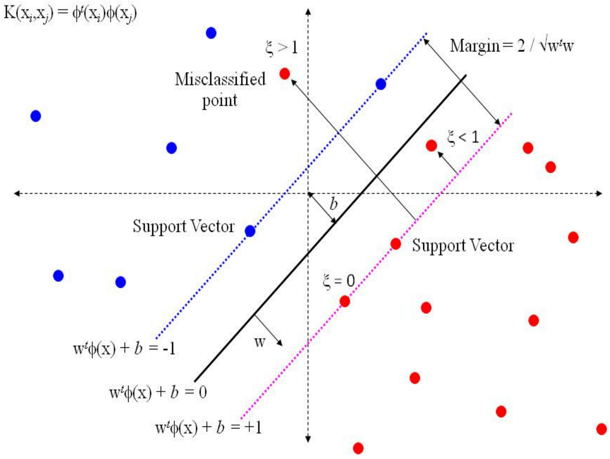


Figure 2: SVM Classifiers

6. Proposed Work:

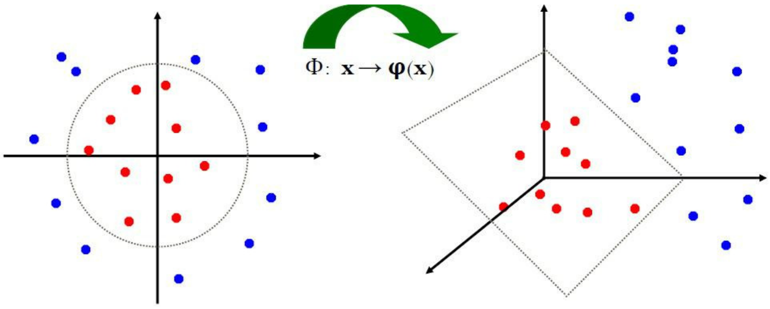
After the data set is normalized using the following equation, PCA is then implemented for reducing the high dimensional DNA microarray data. On the reduced data set feed forward neural network and SVM are implemented and their performance accuracies are compared.

6.1 Data Preprocessing and Cleaning

Filling in missing values, smoothing noisy data, identifying and removing outliers and resolving inconsistencies.

6.2 Data Normalization

Data normalization is followed after data preprocessing and cleaning. Data normalization is essential to the performance of classifiers. We use Z-min-max normalization method. It transforms the data into the desired range [0, 1].



$$X_{norm} = (X_{m\quad n}\quad min) = (max\quad min)$$

X_{norm} is the result of the normalization, $x_{m\quad n}$ is the feature (gene) to be normalized, max is upper bound of the gene expression value, and min is lower bound of the gene expression value.

SVM and BPNN often does not gives better accuracy for high dimension, to improve the efficiency, we proposed to apply Principal component analysis on the original data set, to obtain a reduced dataset containing possibly uncorrelated variables without any loss [5], [18]. Then the reduced data set will be applied to SVM and BPNN classifier to improve performance of the classifiers.

Our first contribution is to prove that PCA is able to reduce dimension of features and to provide classification competitive performance than traditional classifiers in terms of speed and predictive accuracy, and precision of convergence [20].

Hybrid approach is being proposed for reduction of features and structure modeling of classifiers using PCA

[16], [17]. After the implementation of PCA, two classifiers such as Feed Forward Neural Network (FFNN) trained using BP algorithm and SVM [19] are implemented. The general procedure of the algorithm explained in the Figure 2: the brief overview of our entire proposed process is shown below in Figure 3:

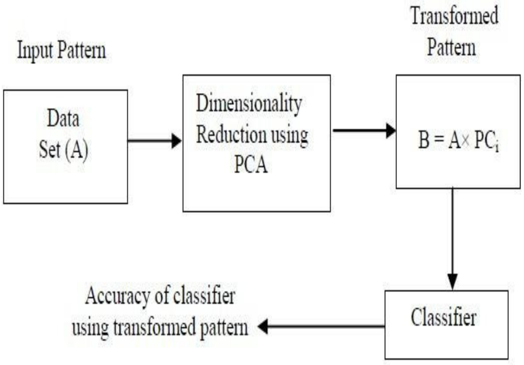

Input Pattern
Data Set (A)
Dimensionality Reduction using PCA
Transformed Pattern
B = A× PC$_i$
Classifier
Accuracy of classifier using transformed pattern

Figure 2: PCA-SVM or PCA-BPNN classifiers for cancer data

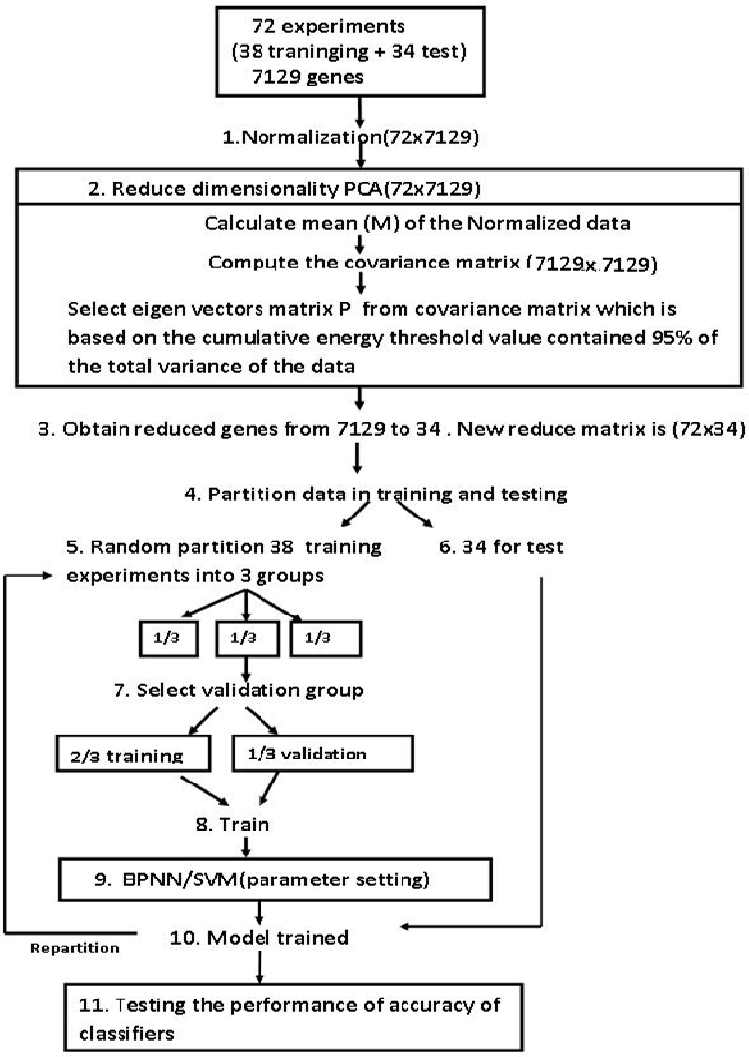

72 experiments
(38 traninging + 34 test)
7129 genes
1.Normalization(72x7129)
2. Reduce dimensionality PCA(72x7129)
Calculate mean (M) of the Normalized data
Compute the covariance matrix (7129x 7129)
Select eigen vectors matrix P from covariance matrix which is based on the cumulative energy threshold value contained 95% of the total variance of the data
3. Obtain reduced genes from 7129 to 34 . New reduce matrix is (72x34)
4. Partition data in training and testing
5. Random partition 38 training experiments into 3 groups
6. 34 for test
1/3
1/3
1/3
7. Select validation group
2/3 training
1/3 validation
8. Train
9. BPNN/SVM(parameter setting)
10. Model trained
Repartition
11. Testing the performance of accuracy of classifiers

Figure 3: Schematic illustration of the proposed method for Leukemia cancer data

The entire data set of all 72 experiments was first Normalized (step 1) and then the dimensionality was further reduced by principal component analysis (PCA) to 34 PCA projections, (2) from the original 7129 expression values. Next, the 34 test experiments were set aside (6) and the 38 training experiments were randomly partitioned into 3 groups from reduced matrix (5). One of these groups was reserved for validation and the remaining 2 groups for training (7). BPNN/SVM models were then trained using for each sample the 34 PCA values as input and the cancer category as output (9). The samples were again randomly partitioned and the entire training process repeated (10). The 34 test experiments were subsequently classified using all the trained models. The entire process (5-10) was repeated.

The goal of PCA is to derive another matrix P matrix which will describe a linear transformation of every column in X (every training gene) in the eigenfaces sub-space, in the form: W=PX, where W are the projections of the training genes on the subspace described by the eigenfaces. The rows of P matrix symbolize the principal components PC and they are orthogonal.

7 Implementation

The simulation process is carried on a machine having Intel(R) core (TM) 2 Duo processor 3.0 GHz and 3 GB of RAM. The MATLAB version used is R2012 (a). This was taken out with 3 microarray cancer data sets.

7.1 Data Sets

Data Set 1: Leukemia cancer

Number of Instances: 72 (consist of 2 classes for distinguishing: Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). The complete dataset contains 25 AML and 47 ALL samples. 38 samples for training set and 34 samples for test set are chosen for simulation).

Number of Attributes: 7129

Resultant data set (after PCA): 72x34.

The data sets taken from public Kent Ridge Biomedical Data Repository with URL: http://sdmc.lit.org.sg/GEDatasets/Datasets.html. or following

URL: http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html.

Data Set 2: Ovarian cancer

Number of Instances: 216 (consist of 2 classes for distinctive: Cancer and Normal. The complete dataset contains 121 ovarian cancer and 95 normal cancer samples. 119 samples for training set and 97 samples for test set are chosen for simulation).

Number of Attributes: 4000.

Resultant data set (after PCA): 216x28.

The data set taken from public Kent Ridge Biomedical Data Repository with the url followed as,

URL: http:// sdmc.lit.org.sg/GEDatasets/Datasets.html.

Data Set 3: Colon cancer

Number of Instances: 62 (consist of 2 classes for distinguishing: cancer biopsies and normal biopsies. The samples consist of 36 cancer biopsies collected from cancer data, and 27 normal biopsies collected from healthy part of the colons of the same patient.)

Number of Attributes: 2000.

Resultant data set (after PCA): 62x12.

The data sets taken from http://microarray.princeton.edu/oncolog.

7.2 Input Parameters

We have design BPNN architecture as 72x3x1 for Leukemia cancer, 216x3x1 for Ovarian and 62x3x1 for colon cancer data set.

BPNN Parameters: Number of nodes in hidden layer=3, learning rate=0.2, Number of iterations=1000.

SVM Parameters: C = 2, k= 8, d = 3.

The parameters that should be optimized include penalty parameter C and the kernel function parameters such as the (gamma) and d for the radial basis function (RBF) kernel. Generally d is set to be 2. Thus the kernel value is related to the Euclidean distance between the two samples is related to the kernel width. Correct parameters setting can develop the SVM classification accurateness.

7.3 Performance Measures

The measure used to evaluate the performance of classifiers:

Accuracy = (correctly classified instances) / (Total no. of instances) *100%

1. Accuracy =(TP+TN) / (TP+FP+TN+FN)

2. Sensitivity = (TP/TP+FN)*100%

3. Specificity = (TN/ TN+FP) * 100%

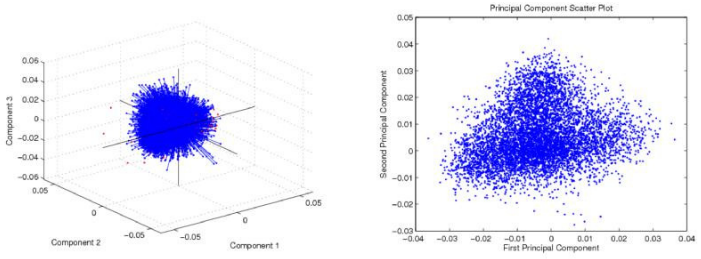
   Where, TP = true positive, TN = true negative FP = false positive, FN = false negative.

8. Numerical Simulation, Results and Discussion

Initially simulation was carried out considering the original features and BPNN and SVM classifiers. This classification approach is validated by considering three other data sets i.e. Leukemia cancer, ovarian cancer and colon cancer data.

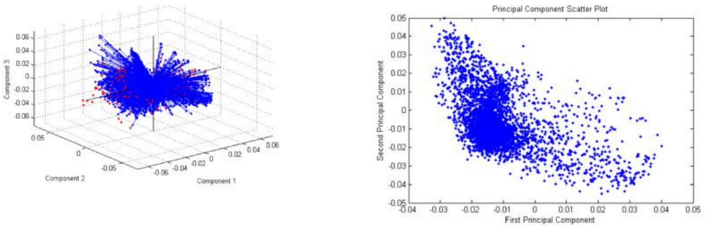
The correctness attained with usual BPNN and SVM were 91% and 93.1% taking Leukemia cancer and 87.1% and 96.2% taking ovarian cancer and 56.7% and 90.03% taking Colon cancer data respectively showing in Table 2 and Table 3.

After the implementation of PCA, the data distribution across the first three principal components (PC's) and first two principal components (PC's) are shown below in Figure 4for Leukemia cancer data set, Figure 5 for Ovarian cancer data set. The classification accuracy varying with number of principal components (PC's) are showing in Table 1. The Accuracy vs. graph is plotted for the principal component which has shown the maximum accuracy in Figure 6. The accuracy obtained with traditional BPNN and SVM were showing in Table 3. The data distribution across the first two features is shown in Figure 7.



(a) 3D representation (b) 2D representation

Figure 4: D and 2D Schematic representation of data across first three PC's and two PC's (Leukemia Cancer data set)



(a) 3D representation (b) 2D representation

Figure 5: D and 2D Schematic representation of data across first three PC's and two PC's (Ovarian Cancer data set)

Using PCA-based approach, the original number of features in Leukemia cancer got reduced from 7129 to 34 Latents (PC's) (i.e. reduced by 99.03%). It covers 95% of the total variance of the data. Therefore, there is hardly any loss of information along a dimension reduction. If the first 34 PC's are chosen, it gives best classification results. In Ovarian cancer Latents reduced from 4000 to 28(i.e. reduced by 82%) and Colon cancer from 2000 to 12 Latents (i.e. reduced by 86.05%) are reduced. Considering the reduced features, the accuracy obtained with PCA-BPNN and PCA-SVM were 97.3% and 98.08% for leukemia cancer and 96.2% and 98.09% for ovarian data set and 95.02% and 97.04% for Colon cancer data set respectively.

Table 1: Accuracy vs. No. of PC's using PCA-SVM (Leukemia Cancer data set)

No of PC's	Accuracy (%)
10	86.03
20	89.04
30	98.03
40	98.08
50	97.12
60	97.23
70	88.23
80	90.03
90	94.08
100	98.04

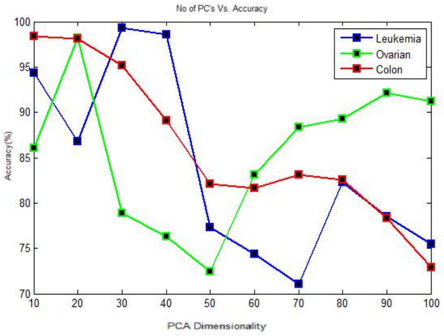


Figure 6: Plot showing Accuracy vs. No. of PC's using PCA

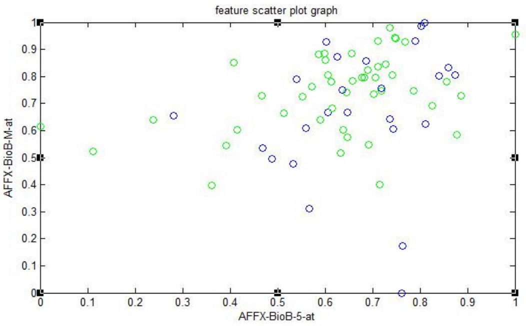


Figure 7: 2D Schematic representation of data across first two features (Leukemia data set)

Table 2: Classification Results: SVM Kernels

Data Set	Classifiers	Time (in sec)	Sensitivity (%)	Specificity (%)	Accuracy (%)
Leukemia (ALL vs. AML)	Linear	0.1548	100	93.33	96.08
	Polynomial	0.0696	100	83.33	87.08
	RBF	0.1548	100	93.3	98.08
	Sigmoid	0.0580	58.9	76.2	58.82
Ovarian (Cancer Vs. Normal)	Linear	0.1976	98.3	100	84.02
	Polynomial	0.1793	98.3	100	98.04
	RBF	0.0976	80	64.1	74.02
	Sigmoid	0.2818	34.4	76.9	59
Colon (Tumor biopsies Vs. Normal biopsies)	Linear	0.0956	98.3	100	84.02
	Polynomial	0.0451	97.03	98	99.02
	RBF	0.1146	85.2	94.4	84.8
	Sigmoid	0.2318	34.4	66.9	69

Table 3: Classification Results: Traditional BP, SVM, PCA-BP, and PCA-SVM

Data Set	Classifiers	Time (in sec)	Sensitivity (%)	Specificity (%)	Accuracy (%)
Leukemia (ALL vs. AML)	BP	6.17	97	86	91
	SVM	0.23	93	67.3	93.1
	PCA-BP	23.74	96	97	97.3
	PCA-SVM	0.1548	100	93.3	98.08
Ovarian (Cancer Vs. Normal)	BP	20.02	98	88.2	87.1
	SVM	9.45	68	81	96.2
	PCA-BP	20.02	98	98.2	96.2
	PCA-SVM	0.0976	98.3	100	98.09
Colon (Tumor biopsies Vs. Normal biopsies)	BP	20.02	48	58.2	56.7
	SVM	9.45	88	81	90.03
	PCA-BP	20.02	92.2	88.2	95.02
	PCA-SVM	0.0451	97.3	98	97.04

9 Conclusion

PCA-BP learning algorithm is designed to reduce network error between the actual output and the desired output of the network in a gradient descent manner.

Experimental outcome illustrate PCA-SVM method screening better results than PCA-BPNN, traditional BPNN and SVM, in terms of speed, accuracy and difficulty. The two stage approach of classification has shown promising results as they have outperformed traditional approaches. In this work the problem of cancer classification is solved successfully using PCA-SVM. If the data are intense over a particular linear subspace, PCA give a way to compress data and make simpler the representation without losing much information. But if the data are intense over a non-linear subspace, PCA fails to work well. Work can be further extended by implementing singular value decomposition or independent component analysis for dimension reduction. Accuracy can be checked by considering some more number of objectives (such as discarded features, weight value association with accuracy etc.) which can be efficiently solved using Genetic algorithm (GA) [21] and MultiObjective Genetic algorithm (MOGA) [15].

Reference

1. V. Vapnik. The nature of statistical learning theory. Springer, 1999
2. Guoqiang Peter Zhang. Neural networks for classification: a survey. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 30(4):451-462, 2000
3. D.E. Rumelhart, G.E. Hintont, and R.J. Williams. Learning representations by back-propagating errors Nature, 323(6088):533-536, 1986
4. J. Shlens. A tutorial on principal component analysis. Systems Neurobiology Laboratory, University of California at San Diego, 2005. [18] Imola K Fodor. A survey of dimension reduction techniques 2002.

5. Jolli e. Principal component analysis, Wiley Online Library, 2005
6. C.C. Chang and C.J. Lin. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3):27, 2011.
7. David E Goldberg. Genetic algorithms in search, optimization and machine learning 1989.
8. Kalyanmoy Deb. Multi-objective optimization, Multi-objective optimization using the evolutionary algorithms, pages 13-46, 2001.