A Classification Technique for Microarray Gene Expression Data using PSO-FLANN

Jayashree Dev[1], Sanjit Kumar Dash[2], Sweta Dash [3], Madhusmita Swain [4]

[1, 2, 4]College of Engineering and Technology
Biju Patanaik University of Technology
Bhubaneswar, Odisha, India

[3] Synergy Institute of Engineering and Technology
Biju Patanaik University of Technology
Dhenkanal, Odisha, India

***Abstract*— Despite of an increased global effort to end breast cancer, it continues to be most common cancer deaths in women. This problem reminds that new therapeutic approaches are desperately needed to improve patient survival rate. This requires proper diagnosis of disease and classification of tumor type based on genomic information according to which proper treatment can be provided to the patient. There exists a no. of classification techniques to classify the tumor types. In this paper we have focused on three different classification techniques: BPN, FLANN and PSO-FLANN and found that the integrated approach of Functional Link Artificial Neural Network (FLANN) and Particle Swarm Optimization (PSO) can better predict the disease as compared to other method.**

Keywords- Backpropagation (BPN), FLANN, PSO, PSO-FLANN

I. INTRODUCTION

Breast cancer is a heterogeneous disease with respect to molecular alteration, cellular composition, and clinical outcome. This diversity creates a challenge in developing tumor classifications that are clinically useful with respect to prognosis or prediction. There are two types of breast cancer: ductal carcinomas and lobular carcinomas. Cancers originating from ducts are known as ductal carcinomas and cancers originating from lobules are known as lobular carcinomas. The high incidence of breast cancer is emerging as a public health problem in the country in past decades. While there is no single reason for the escalating incidence of breast cancer, obesity, dietary habits, physical inactivity and the over-use of hormonal pills are the main causes. Prognosis and survival rates for breast cancer vary greatly depending on the cancer type, stage, treatment and geographic location of the patient. Breast cancer is the most commonly diagnosed form of cancer in women accounting for about 30% of all cases. Some studies, which were based on low capability and poorly calibrated equipment, infrared imaging has been shown to be well suited for task of detecting breast cancer, in particular when the tumor is in its early stages or in dense tissue [1].

Microarrays are a powerful tool for biologists as they enable the simultaneous measurement of the expression levels of thousands of genes per tissue sample [10]. Microarray analysis is a widely used technology for studying gene expression on a global scale. Gene expression profiling by DNA microarray has become an important tool for studying the transcriptome of cancer cells and has been successfully used in many studies of tumor classification and identification of marker genes associated with cancer. With an increasing number of microarray data becoming available, the comparative of study on normal tissue versus tumor tissue has gained high importance.

Microarray breast cancer event prediction, however, has proven to be difficult, as few classification rules are able to obtain a balanced accuracy rate of over 70%, when properly validated. These performance indicators are also often associated with wide confidence intervals. Signature composition strongly depends on the subset of patient samples used for feature selection. In recent years many different signatures have been proposed, mostly derived using different patient populations and/or array technologies. Although the overall performance of these signatures is comparable, there is often a high level of inconsistency between class assignments obtained using different signatures. One of the challenging aspects of microarray data is that they are subject to various sources of technical variation, arising from the many experimental laboratory steps needed to get from a tissue sample to an array scan. The noises can be removed from microarray data using some preprocessing methods.

The goal of this study is to investigate the benefit of performing supervised classification analyses on microarray data [6]. Methods of supervised classification analysis render it possible to automatically build classifiers that distinguish among specimens on the basis of predefined class label information (phenotypes) and

in many cancer research studies; the application of these methods has shown promising results of improved tumor diagnosis and prognosis. In this paper, an integrated approach of functional link artificial intelligence (FLANN) [2] and particle swarm optimization (PSO) [3] is used to build a more reliable Classifier. This approach incorporates gene features and FLANN parameters into one common solution code. The problem is solved by selecting gene features and optimizing the parameters of the FLANN classifier.

II. CLASSIFICATION WITH BPN MODEL

Neural networks with back propagation technique can be used for classification task such as character recognition, voice recognition, medical application etc [7, 8, 9]. Neurons in neural network have weighted inputs, threshold values, activation function and an output where activation function=f (∑ (inputs * weights). Threshold values play an important role in deciding the output. Backpropagation technique can be applied to multi-layered neural network. Neural networks learn by example. That means, during training we have to apply input-output pair from the training set to the neural network. Training continues till the network is well-trained. After that the performance of the network can be checked by applying the patterns belonging to testing set. The increasing no. of hidden layers results in the computational complexity of the network. The time taken for convergence and to minimize the error may be very high due to this reason.

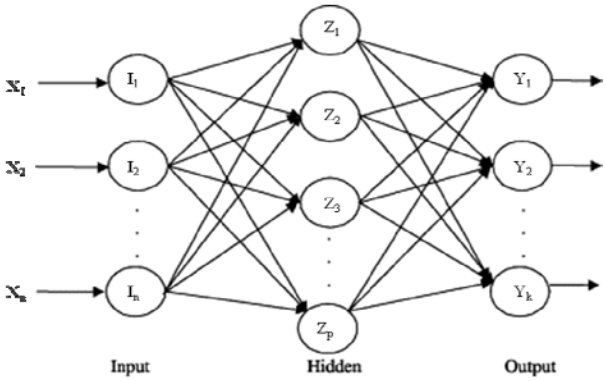

Figure 1: Basic structure of BPN Classifier

The pseudo-code for backpropagation classification is as follows:

Step 1: Initialize weights to small random values.

Step 2: while stopping condition is false, do steps 3-10.

Step 3: For each training pair do steps 4-9

Step 4: Each input unit receives the input signal x_i and transmits the signal to all units in the hidden layer.

Step 5: Each hidden unit (Z_h, h=1...p) sums its input signals

$$Z_{inj}=V_{oj}+\sum (x_i V_{ij})$$

By applying activation function $Z_j=f(Z_{inj})$ and send this signal to all units in output layer.

Step 6: Each output unit (Y_k, k=1...m) sums its weighted input signals.

$$Y_{ink}=W_{ok}+ \sum (Z_j W_{ij})$$

And supplies its activation function to calculate the output signal

$$Y_k= f (Y_{ink})$$

Step 7: Each output unit (Y_k, k=1...m) receives a target pattern corresponding to input pattern, error information term is calculated as

$$\delta k= (t_k-Y_k) f (Y_{ink}) \text{ and } \delta_{inj} = \sum \delta_j W_{jk}$$

Step 8: Each hidden unit (Z_j, j=1...n) sums its delta inputs from unit layers above. The error information term is calculated as:

$$\delta_j =\delta_{inj} f(Z_{inj})$$

Step 9: Each output unit (Y_k, k=1… m) updates its weight (j=0… n)

The weight correction term is given by $\Delta W_{jk} = \alpha \delta_k Z_j$

And bias correction term is given by $\Delta W_{ok} = \alpha \delta_k$

Therefore $W_{jk(new)} = W_{jk(old)+} \Delta W_{jk}$ and

$W_{ok\ (new)} = W_{ok\ (old)\ +} \Delta W_{ok}$

Each hidden unit (Z_j, j=1… p) updates its weights (i=0…n)

The weight correction term is $\Delta V_{jk} = \alpha \delta_j X_j$

And bias correction term is $\Delta V_{oj} = \alpha\delta_j$

Therefore $Vij_{(new)} = V_{ij(old)+} \Delta V_{ij}$ and

$V_{oi(new)} = V_{oi\ (old)\ +} \Delta V_{oi}$

III. CLASSIFICATION WITH FLANN MODEL

Single layer neural networks are suitable for solving linear problem and also they are simple and easy to implement. FLANN is a mathematical model or computational model that is inspired by structural and/or functional aspects of biological neural networks [4, 5]. It consists of an interconnected group of artificial neurons and it processes information using a model. FLANN can be treated as a single layer neural network but with the ability to solve non-linear problems. Here, additional input data are generated off-line using non-linear transformations. In a FLANN, the input layer is expanded to a layer of functional units, which consists of higher order combinations of the input units. Each functional unit is fully connected to next layer. The addition of higher order combinations of inputs artificially increases the dimension of input space and hence the hyper planes generated by the FLANN provide greater discrimination capability in the input pattern space. Both non-hidden layer and enhanced input pattern space make this network high precision of function approximation. There are three approaches of input layer expansion:-functional expansion, combined convolution and the combination of both. Here, we employ the functional expansion. There are three different polynomials for functional expansion of input pattern in FLANN. They are Chebyschev, Legendre and Power series. It is important to choose suitable functional expansion.

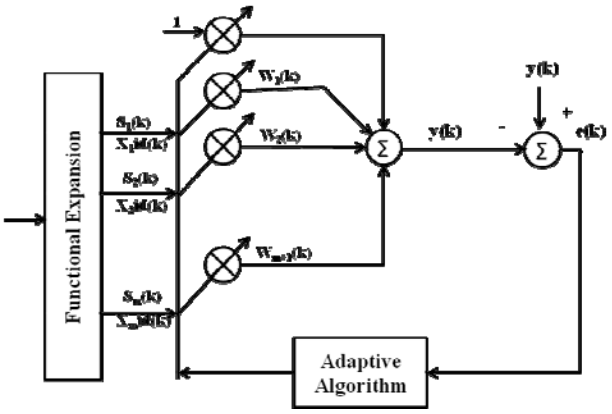

Figure 2: Basic Structure of FLANN Classifier

Training process constitutes following steps:

Step 1: Read the data D= {D_1, D_2… D_N} where each Di is a vector of length M and the class label of order N * 1
Step 2: Normalize the Data.
Step 3: Take 80 % of data for training in DTr and 20% of data for testing as DT.
Step 4: Randomly assign the wt for each FLANN
Step 5: for i=1: maxIter do
Step 6: δw = wt;
Step 7: for j=1: N do
Step 8: evaluate the output of flann v= DTr (j, :)* δw
Step 9: evaluate error e (j) = v – expected_out_put
Step 10: Update δw= δw – 2 * μ * e (j);
Step 11: end //end of step 5
Step 12: er (i) = Mean (e);
Step 13: wt = wt - δw;
Step 14: end //end of step 7

After the training is over, testing process starts. During testing, the prediction capability of the network is tested.

IV. CLASSIFICATION WITH PSO-FLANN MODEL

The proposed hybrid FLANN is based on evolutionary algorithm Particle Swarm Optimization (PSO) [9]. FLANN suffers from following problems:

1. FLANN is more sensitive to error in training pattern.
2. Failure to converge due to insufficient structure, lockup, tracking to local optima, limited cycles
3. Shows good performance on training set and poor performance in testing set due to the problem of training set not representing variations in patterns and too strict in tolerance.

To avoid such problems we have proposed an evolutionary technique based model which is able to classify the patterns in a better way.

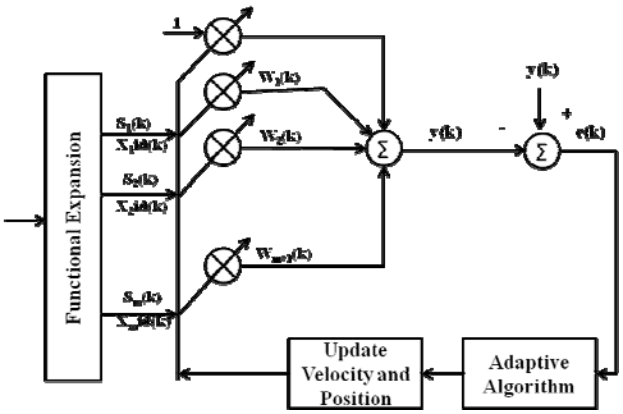

Figure 3. Basic Structure of FLANN with PSO Classifier

We can harness the power of Particle Swarm Optimization [9] for training the FLANN to reduce the local optimality and speed up the convergence. PSO algorithm can be applied to three main attributes of neural networks: connection based, network architecture and network learning algorithm. Here we have focused on connection based. PSO simulates the behavior of bird flocking. It learns from the scenario and uses it to solve the optimization problem. IN PSO, each single solution is a "bird" in the search space. We call it as "particle". All particles have fitness values which are evaluated by fitness function to be optimized and have velocities which direct the flying of particles. The particles fly through the problem space by following the current optimum particles.

Training process constitutes following steps:

Step 1: Read the data D= $\{D_1, D_2 \ldots D_N\}$ where D_i is a vector of length M and the class label of order N * 1

Step 2: Normalize the Data.

Step 3: Take 80 % of data for training in DT_r and 20% of data for testing as DT.

Step 4: Calculate the fitness for each swarm for every DT_i

Step 5: Randomly assign the wt for each FLANN

Step 6: initialize the parameter pbest, c1 and c2 for swarm, Velocity V_{id} and position Updation X_{id}.

Step 7: for i=1: maxIter do

Step 8: $\delta w = wt$;

Step 9: for j=1: N do

Step 10: evaluate the output of flann v= Xid (j)* Mean (DTr (j, :))*DTr (j, :)* δw

Step 11: evaluate error and pbest of each swarm

e(j)= v – expected_out_put

pbest (j)=max(pbest(j), v)

Step 12: Update $\delta w = \delta w - 2 * \mu * e(j)$;

Step 13: update new velocity and new position for swarm

$$V_{id}^{j+1} = wV_{id}^{j} + C_1 * rand_1 () * (pbest - X_{id}) + C_2 * rand_2 () * (gbest - X_{id})$$

Where $X_{id}^{j+1} = X_{id}^{j} - V_{id}^{j}$

Step 14: end //end of step 9

Step 15: er(i) = Mean(e);

Step 16: wt = wt - δw;

Step 17: calculate gbest=max (v)

Step 17: end //end of step 7

The above FLANN with PSO algorithm is used to train and test the breast cancer dataset. Here the datasets are taken and first trained with FLANN with PSO model by providing random weights, position value, the value of random constants c_1 and c_2 and pbest, gbest value. Then error will be checked by expected output and resulted output, according to that weight and position will be modified.

V. PERFORMANE ANALYSIS

The proposed technique for bio-medical data classification has been implemented in the working platform of MATLAB (version 7.8). For evaluating the proposed technique, we have utilized the microarray gene samples of human breast cancer dataset. The data set is collected from UCI machine learning repository (http://archive.ics.uci.edu). The BPN, FLANN and PSO-FLANN have been trained by breast cancer dataset having class label two. Table 1, 2 and 3 shows the performance of different classification techniques used by us.

From the Table 3, it can be seen that the PSO-FLANN has provided more accuracy and less error rate rather than the BPN and FLANN-based gene classification techniques. More accuracy and less error rate leads to effective classification of the given microarray gene data to the actual class of the gene. From the above training and testing phases of BPN, FLANN and PSO-FLANN, we observed that FLANN with PSO provides good results than BPN and FLANN.

TABLE I. CONFUSION MATRIX FOR BREAST CANCER DATA SET USING BPN

		Predicted	
		Class 1	**Class 2**
Actual	**Class 1**	62	15
	Class 2	46	16

TABLE II. CONFUSION MATRIX FOR BREAST CANCER DATA SET USING FLANN

		Predicted	
		Class 1	**Class 2**
Actual	**Class 1**	56	21
	Class 2	30	32

TABLE III. CONFUSION MATRIX FOR BREAST CANCER DATA SET USING PSO-FLANN

		Predicted	
		Class 1	**Class 2**
Actual	**Class 1**	70	7
	Class 2	4	58

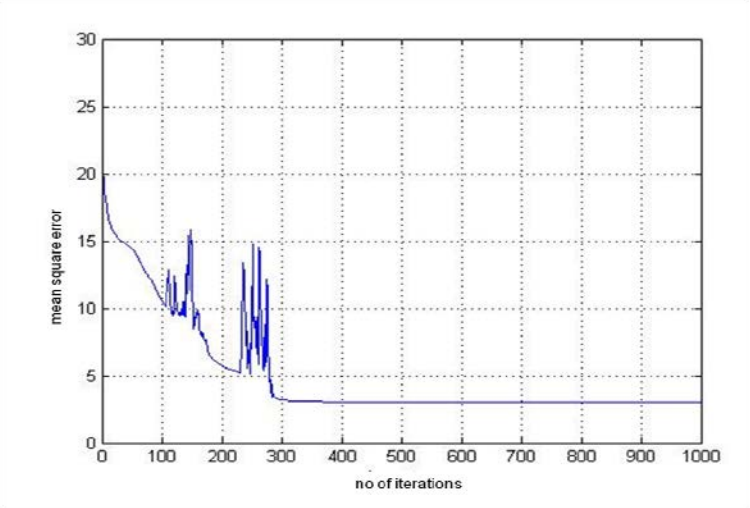

Figure 4. Error Graph using BPN

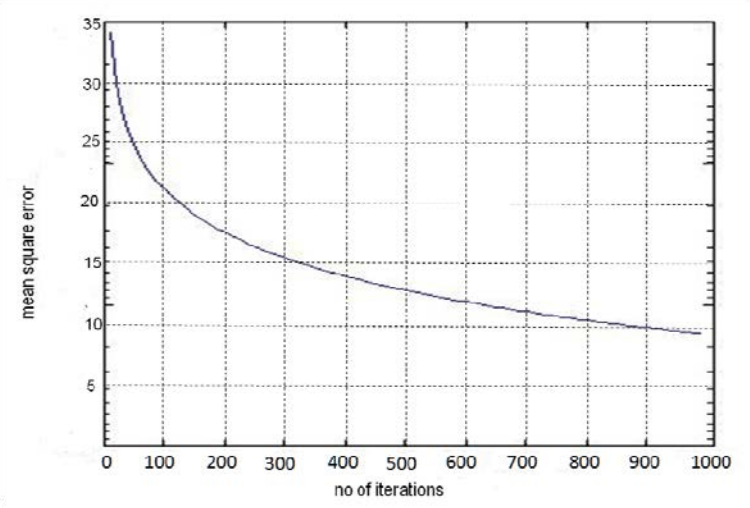

Figure 5. Error Graph using FLANN

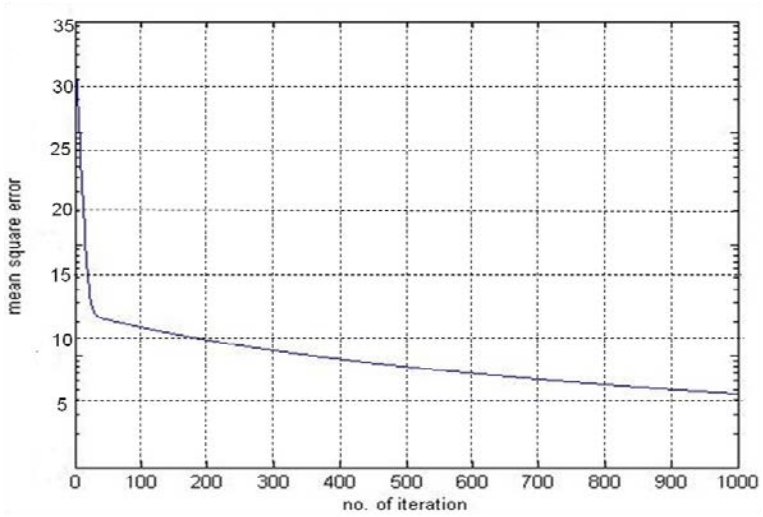

Figure 6. Error Graph using PSO-FLANN

TABLE IV. COMPARISON OF CLASSIFICATION RESULTS

Name of data set	Percentage of Accuracy		
	BPN	**FLANN**	**PSO_FLANN**
Breast Cancer	56.12%	63.34%	92.36%

VI. CONCLUSION

The performance of PSO-FLANN is better for classification as compared to BPN and FLANN. The performance has been checked on breast cancer data set which is obtained from the UCI machine learning repository website. This method overcomes the nonlinearity of the classification problem. The FLANN with PSO architecture, because of its simple architecture and computational efficiency may be conveniently employed in other tasks of data mining and knowledge discovery in databases such as clustering, feature selection, feature extraction, association rule mining, regression, and so on. However, a necessary property of algorithms which are capable of handling large and growing datasets is their scalability or linear complexity with respect to the data size. The extra calculation generated by the higher order units can be eliminated, provided that these polynomial terms are stored in memory instead of being recalculated each time the FLANN trained.

REFERENCES

[1] A M. Sarhan; Cancer Classification Based on Microarray Gene Expression Data using DCT and ANN; Journal of Theoretical and Applied Information Technology, pp. 208-216, 2009.

[2] Patra J. C., Nguyen C. Thanh and Meher Pramod K.; Computationally Efficient FLANN-based Intelligent Stock Price Prediction System; Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, June 14-19, 2009.

[3] Enrique Alba, JoseGarcıa-Nieto, Laetitia Jourdan and El-Ghazali Talbi; Gene Selection in Cancer Classification using PSO/SVM and GA/SVM Hybrid Algorithms.; Author manuscript, published in "Congress on Evolutionary Computation, Singapor : Singapore (2007)..

[4] Dehuri S., Cho S.B.; A Comprehensive Survey on Functional Link Neural Networks and an Adaptive PSO-BP Learning for CFLNN; Department of Computer Science, Yonsei University, 262 Seongsanno, Seodaemun-gu, Seoul 120-749, Korea.

[5] Leandro dos Santos Coelho, Cezar Augusto Sierakowski ; A software Tool for Teaching of Particle Swarm Optimization Fundamentals; Advances in Engineering Software, Volume 39, Issue 11, November 2008, Pages 877-887 .

[6] Frank Rapaport, Marie Dutreix, Emmanuel Barillot, Jean-Philippe Vert; Classification of Micro Array Data using Gene Networks; February 1, 2007; BMC Bioinformatics.

[7] F. Pauline; Classification of Breast Cancer by Comparing Backpropagation Training Algorithms; IJCSE; Vol-3, No.1; Jan 2011; ISSN: 0975-3397.

[8] A. Gupta, M. Shreevastava; Medical Diagnosis using Backpropagation Algorithm; International Journal of Emerging Technology and Advanced Engineering; Vol.1, Issue 1, November 2011, Page 55-58 ; ISSN: 2250-2459.

[9] Chow, M.-Y., Sharpe, R. N., & Hung, J. C. (1993). On the Application and Design of Artificial Neural Networks for Motor Fault Detection. IEEE Transactions on Industrial Electronics, 40(2), 181-188.

[10] S.K. Gruvbergek, H.E. Cuniliffe, K.M. Carr and I.A. Hedenfalk; Microarrays in Breast Cancer Research and Clinical practice-The Future Lies Ahead; An Official Journal of Society for Endocrinology and European Society of Endocrinology; Dec 1, 2006.