Cat Swarm based Optimization of Gene Expression Data Classification

Amit Kumar[1#], Debahuti Mishra[2]

Department of Computer Applications, ITER, Siksha O Anusandhan University
Bhubaneswar, Odisha

Abstract—**An Artificial Neural Network (ANN) does have the capability to provide solutions of various complex problems. The generalization ability of ANN due to the massively parallel processing capability can be utilized to learn the patterns discovered in the data set which can be represented in terms of a set of rules. This rule can be used to find the solution to a classification problem. The learning ability of the ANN is degraded due to the high dimensionality of the datasets. Hence, to minimize this risk we have used Principal Component Analysis (PCA) and Factor Analysis (FA) which provides a feature reduced dataset to the Multi Layer Perceptron (MLP), the classifier used. Again, since the weight matrices are randomly initialized, hence, in this paper we have used Cat Swarm Optimization (CSO) method to update the weight values of the weight matrix. From the experimental evaluation, it was found that using CSO with the MLP classifier provides better classification accuracy as compared to when the classifier is solely used.**

Keywords—**Classification, Artificial neural network, Multi-layer perceptron, Principal component analysis, Factor analysis, Cat swarm optimization.**

I. INTRODUCTION

Progress in the domain of machine learning and data mining have helped the biomedical researchers to improve the quality of healthcare [1]. Today biomedical informatics has number of applications to solve various real world problems. One such problem is the classification of gene expression data. Classification [2] is defined as the task of identifying the sub-classes to which new observations may belong on the basis of data containing observations whose sub-classes are known. Zuyi wang *et al.* [3] have described diagnostic classification as the task of assigning a particular unknown sample to a known disease class based on the expression levels of gene expression data. Classification techniques on the gene expression data can be implemented by one of the various methods such as decision tree, artificial neural network, rough sets and bayesian methods [2]. In this paper, we have used MLP as the classifier. MLP's are trainable algorithms [4] that can learn to solve complex problems because of their massively parallel processing capability, fault tolerance, self-organisation and adaptive capability, which guarantees of high classification accuracy. The performance of a classifier is highly degraded when applied with gene expression data due to the curse of dimensionality of such datasets. Hence, feature reduction techniques [5] are applied on these datasets which selects the relevant features from the dataset. In this paper, two techniques PCA and FA are used for feature reduction. This feature reduced dataset is used to train the MLP. A typical neural network consist of a couple of hundred of weights whose value must be found to produce an optimal solution. Hence, in this paper we have employed a relatively new bio-inspired optimization technique called cat swarm optimization (CSO) [6] which optimizes the synaptic weights between the neurons. The authors in [6] proposed the CSO algorithm in which they modelled the behavioural attitude of the cats. The algorithm is described by two modes namely *seeking mode* and *tracing mode* seeking mode is used to describe the cat when it is resting and looking around for the next position to move. Whereas, tracing mode describes the cat when it is moving or tracing some targets. In this paper, we have designed a novel CSO based optimized classifier for gene expression data. The layout of this paper is as follows; section II deals with background study, in section III the preliminary concepts of data normalization, feature selection, MLP, and cat swarm optimization are described. In section IV schematic representations of proposed model is given; in section V experimental evaluations and results are described and finally, section VI deals with conclusion and future work.

II. BACKGROUND STUDY

The problem of optimizing a classification technique with suitably high accuracy has always been a challenge and an area of interest for the researchers. Barnaghi *et al.* [2] compared various classification methods. They proved that neural network classifiers method obtained a better result as compared to Bayesian and rough sets. Among the neural network classifiers MLP showed high accuracy as compared to radial basis function. John paul T. Yusiong [4] has optimized the artificial neural network by using CSO. In his work CSO was used as the training algorithm and optimal brain damage (OBD) as the pruning method. His work proved that the network complexity can be reduced by pruning the connection weights among the layers without affecting the classification accuracy. Chu *et al.* [6] have proposed the CSO algorithm in which, they modelled the behavioural attitude of the cats. The algorithm is described by two modes namely seeking mode and tracing mode. Seeking mode is used to describe the cat when it is resting and looking around for the next position to move. This mode is described by four parameters like counts of dimension to change (CDC), seeking range of dimension (SRD), self position consideration (SPC) and seeking memory pool (SMP). Whereas, the other mode i.e. tracing mode describes the cat when it is moving or tracing some targets. Tanwai *et al.* [7] investigated the challenges faced while classifying biomedical datasets. The authors have addressed several issues such as high dimensionality of dataset, missing value and multiple classes. They proposed the guidelines to select machine learning algorithms best suited for a particular dataset. Ling *et al.* [8] worked with different classification methods and provided their performance comparison. In their work they have showed that the nearest neighbour classifier works well with lung cancer and leukaemia dataset and MLP works well with brain tumour dataset. Yang *et al.* [9] applied PCA for feature selection of gene expression data thus reducing the risk of over-fitting. This feature reduced dataset is provided to the neural network and are trained to to learn the relationship between the input pattern and the output with improved accuracy. Ladha *et al.* [10] have proposed the guidelines to select feature selection algorithms. Their work includes the framing of parameters and desirable features that a feature selection algorithm should have. They reviewed various feature selection algorithms and compared their performances. Borges *et al.* [11] applied two methods of feature selection over a gene expression dataset. They compared wrapper approach with sequential search and filter approach combined with dependency evaluation measure. The wrapper approach provided a better classification accuracy but with a high computational cost. Inan *et al.* [12] proposed a new hybrid feature selection technique by combining the apriori algorithm and PCA with the artificial neural network classifier. This new method has helped the classifier to learn fast with a reduced size of dataset with improved accuracy.

III. Preliminary Concepts

A. Data Normalization

Gene expression datasets available at the different dataset repositories are of high range [13], which increases the complexity of computation of any data mining task. In this paper, we have used min-max normalization technique which scales a dataset from high range to low range without affecting the result [14]. Min-max normalization maps a value *v* of attribute A to *v*' in the range [*new-min*$_A$, *new-max*$_A$] by computing (1):

$$V' = \left((v - min_A) / \ (max_A - min_A)\right) \qquad (1)$$

B. Feature Selection

A gene expression dataset has usually less number of samples as compared to the attributes. So, implementing classification techniques over such datasets may decline the generalization ability of a classifier due to over-fitting of the data [15]. In order to overcome the curse of dimensionality, we have employed two feature selection techniques PCA and FA which selects the relevant features from the dataset. PCA [5] uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called as principal component. FA is described as a statistical method which decreases the count of variables by describing the variability among observed correlated variables in terms of lower number of unobserved uncorrelated variables called as factors [16-17].

C. Multi Layer Perceptron (MLP)

Artificial neural networks (ANN) were introduced by Mcculloch and Pits in 1943. ANN has multiple layers of neurons and our aim is to simulate these neurons by varying the coefficients of connectivity called as weight between the neurons. Changing the connection weights among the neurons causes the network to learn the solution to a problem [3]. MLP is a feed forward [7] artificial neural network model which consists of fully connected multiple layers of nodes that can be trained to associate input vectors to specific output vectors. Once the architecture is fixed the network is trained, the network goes on updating mean squared error until a desired value is reached which provides generalization capability to the network [19]. In this paper, we have implemented MLP as classifier. The general architecture of MLP is given in fig.1.

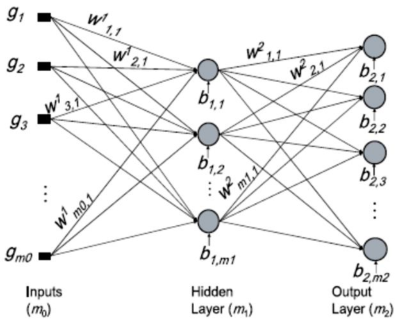

Fig 1: The general architecture of MLP

D. Cat Swarm Optimization (CSO)

It is a relatively new bio-inspired optimization technique based on the behaviour of cats [3]. It describes the behaviour through seeking mode and tracing mode. The process starts with creating copies of the cat, where the number of copy varies depending on the problem statement. Each of these cats is described by its position, which consists of *d* dimensions. Each dimension has its velocity, a fitness value to accept or reject the cat and a flag value which decides the cat will be in seeking mode or tracing mode. While the cat is in different modes following are the various parameters used for defining the behaviour. SMP describe the number of points sought by the cat. Its value depends upon the size of the memory pool. SPC is a value which decides, whether the current position of the cat will be a point to move to. CDC counts the number of dimensions to be changed for a given position of the cat. SRD decide the range within which, a specific dimension of the cat will be changed. The cats are processed in these two modes as described below:

Seeking Mode

Step 1: Make j copies of the present position of cat_k, where j = SMP. If the value of SPC is true, let j = (SMP − 1), then retain the present position as one of the candidates.

Step 2: For each copy, according to CDC, randomly plus or minus SRD percents the present values and replace the old ones.

Step 3: Calculate the fitness values (FS) of all candidate points.

Step 4: If all FS are not exactly equal, calculate the selecting probability of each candidate point by (2), otherwise set all the selecting probability of each candidate point be 1.

Step 5: Randomly pick the point to move to from the candidate points, and replace the position of cat_k.

$$P_i = \frac{FS_i - FS_b}{FS_{max} - FS_{min}}, where\ 0 < i > j \qquad (2)$$

If the goal of the fitness function is to find the minimum solution, $FS_b = FS_{max}$, otherwise $FS_b = FS_{min}$.

Tracing Mode

Step 1: Update the velocities for every dimension (v_k, d) according to (3).

Step 2: Check if the velocities are in the range of maximum velocity. In case the new velocity is over-range, it is set equal to the limit.

Step 3: Update the position of cat_k according to equation (4).

$$v_k, d = v_k, d + r_1 * c_1 * (x_{best}, d - x_k, d), d = 1,2,3, \dots, M \qquad (3)$$

Where x_{best}, d is the position of the *cat*, who has the best fitness value; x_k, d is the position of cat_k, c_1 is a constant and r_1 is a random value in the range of [0, 1].

$$x_k, d = x_k, d + v_k, d \qquad (4)$$

IV. Schematic Representation of Proposed Model

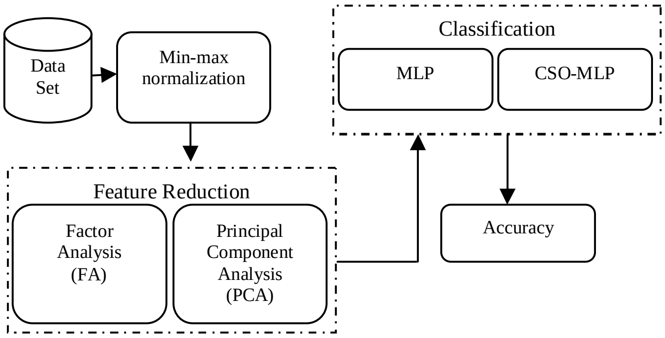

Fig 3: Proposed model

In this proposed model, the gene expression dataset is normalized by using min-max normalization. This normalized dataset is provided to PCA [5] and FA [17] for feature reduction, which reduces the dimension of the dataset. This reduced dataset is provided as input to the two classifiers MLP and CSO-MLP. Finally, the accuracy of the two individual classifiers are measured and compared.

V. Experimental Evaluation and Result Analysis

We have used two benchmarked datasets downloaded from UCI machine learning repository [18] in our experiment which are described in table 1. In this work, we have used MATLAB version 7.10, release name- R2010a. The experiment was carried on Intel core i3 processor, 2.4 GHZ, 32 bit 1GB RAM, 1GB disk space for MATLAB, 3-4 GB for ideal installation.

TABLE I: Description of datasets

Data set Name	Dimension
Breast Cancer	98 * 26
Pima Indian Diabetes	768* 9

The total experimental evaluation has been carried out in the following steps

Step1: Collection of datasets: Two data sets as describe in table 1 has been collected and processed for further processing. Breast cancer data set has 98 instances and 26 samples, whereas, Pima Indian Diabetes data set contains 768 instances and 9 samples as shown in fig.4 and fig.5 for breast cancer and Pima Indian Diabetes respectively.

Step2: Normalization of datasets: Data normalization is an important step in the knowledge discovery process, can be even considered as a fundamental building block of data mining. The attribute data is scaled to fit in a specific range. There are many type of normalization available; we have used one technique called Min Max Normalization here as discussed in section III. The attributes need to scaled to fit in the range [0.0, 1.0]. Applying the min max normalization (1), we get the normalised data set as given in fig. 6 and fig. 7 for breast cancer and Pima Indian respectively.

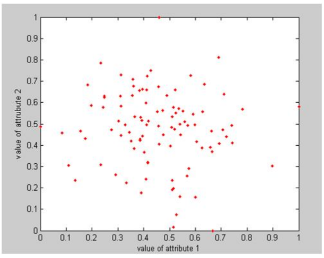

Fig 4: Breast cancer data set (Oiginal)

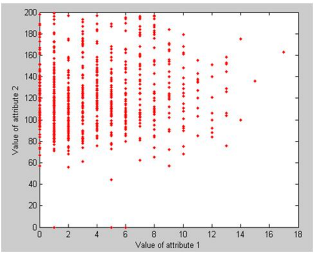

Fig 5: Pima Indian Diabetes data set (Original)

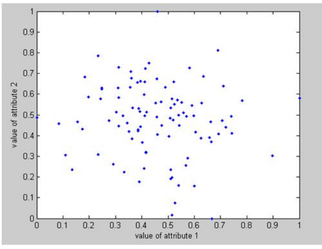

Fig 6: Breast cancer data set (after normalization)

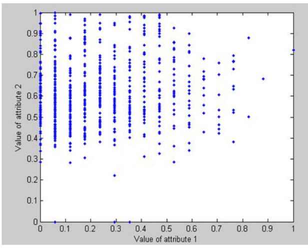

Fig 7: Pima Indian Diabetes data set (after normalization)

Step3: Feature reduction: For high-dimensional datasets dimension reduction is usually performed prior to applying clustering and classification in order to avoid the effects of the curse of dimensionality. Feature reduction is the process of reducing the number of random variables under consideration. Here, we have used both FA and PCA for reduction of features.

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. In other words, it is possible, for example, that variations in three or four observed variables mainly reflect the variations in fewer unobserved variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modelled as linear combinations of the potential factors, plus "error" terms. The information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset. After normalizing both the data sets we have implemented FA for feature reduction and the result is shown in fig. 8 and fig.9.

PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components. The result of feature reduction after applying to both the data sets is given in fig.10 and fig.11.

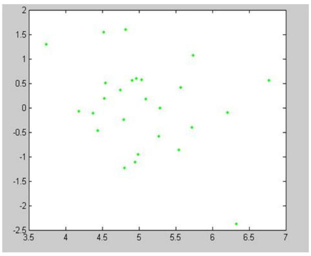

Fig 8: Breast cancer data set (after reduction using FA)

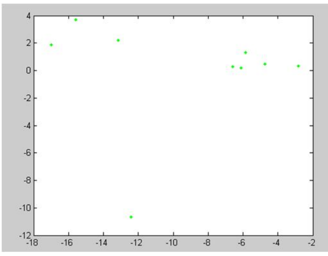

Fig 9: Pima Indian Diabetes data set (after reduction using FA)

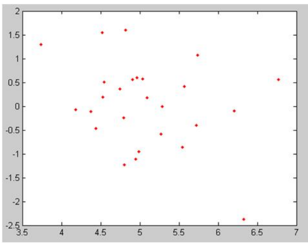

Fig 10: Breast Cancer data set (after reduction using PCA)

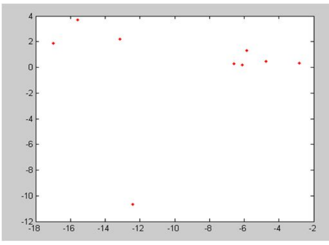

Fig 11: Pima India Diabetes data set (after reduction using PCA)

Step4: MLP for classification: MLP is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back-propagation for training the network. MLP has been applied on the normalized and reduced data sets with the corresponding parameters.

The data set is divided into two parts; out of which 75% is used for training the network and 25% for testing. The parameters of the MPL has been initialized to *Eta*=0.6, *alpha*=0.5. The network has been trained using random weights of *V* and *W*. The output of the network has been calculated using: *OO*=1/tan sigmoid. The error has been computed using: Error (*i*)=(1-*OO*)*(1-*OO*). The weights have been updated using CSO. The CSO updating function is described in the next step. Then, mean square error has been computed and the fig. 12 and fig. 13 shows the error curve of PCA and FA reduced breast cancer data sets whereas, fig. 14 and fig. 15 shows the error curve for FA reduced Pima India dataset respectively.

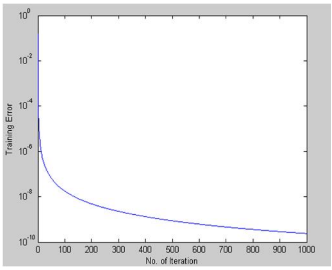

Fig 12: Error curve for PCA reduced breast cancer data set

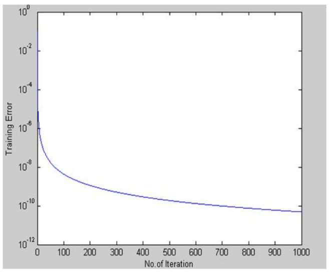

Fig 13: Error curve for FA reduced breast cancer data set

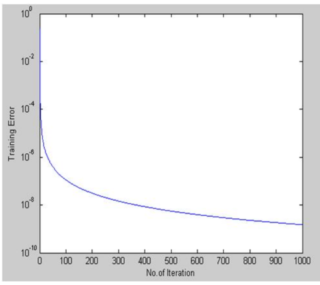

Fig 14: Error curve for PCA reduced Pima Indian Diabetes data set

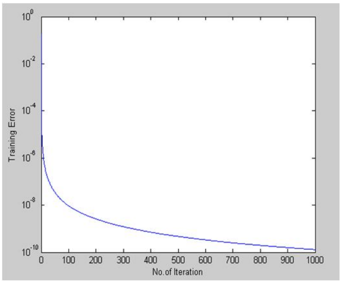

Fig 14: Error curve for FA reduced Pima Indian Diabetes data set

Step5) CSO for updating weights: CSO is one of the new heuristic optimization algorithms which based on swarm intelligence. The following is the steps used for updating the weight of the MLP using cat's behaviour.

Step1) Read the weight matrix v and w and change in weight matrix del$_v$ and del$_w$ from MLP training

step2) Initialize parameters of CSO. SRD=0.5, SMP=50

Step3) Generate copies of cat; if SMP=k, then k number of copies.

Step4) Randomly add or subtract SRD form the value of cat: cat (it)=cat(it)-srd, cat(it)=cat(it)+srd

Step5) Calculate fitness of every cat. fit(it)=1/(1+cat(it)^2)

Step6) Compute maximum and minimum fitness of cat maxf=max(fit); minf=min(fit);

Step7) Find probability of selecting each cat. p= (fit-maxf)/(maxf-minf);

Step8) Update the velocity/weight and position of each cat:
*v(i,j)=v(i,j)+rand(1)*0.6*(cat(q)-x(i,j))*
x(i,j)=x(i,j)+v(i,j)

Step9) Plot the cats

In this experiment, we start with random values of weight matrix in the range of -1 to +1. Then these weights are updated in each of the iteration using CSO algorithm to find out a weight matrix in which the values of the weights are closer to one another. The weight matrix at the cat position 2 consists of weights that are within a small range. Figures presented above are some of the frames which shows how the weights are optimized retaining a suitably high accuracy for the classification of gene expression data.

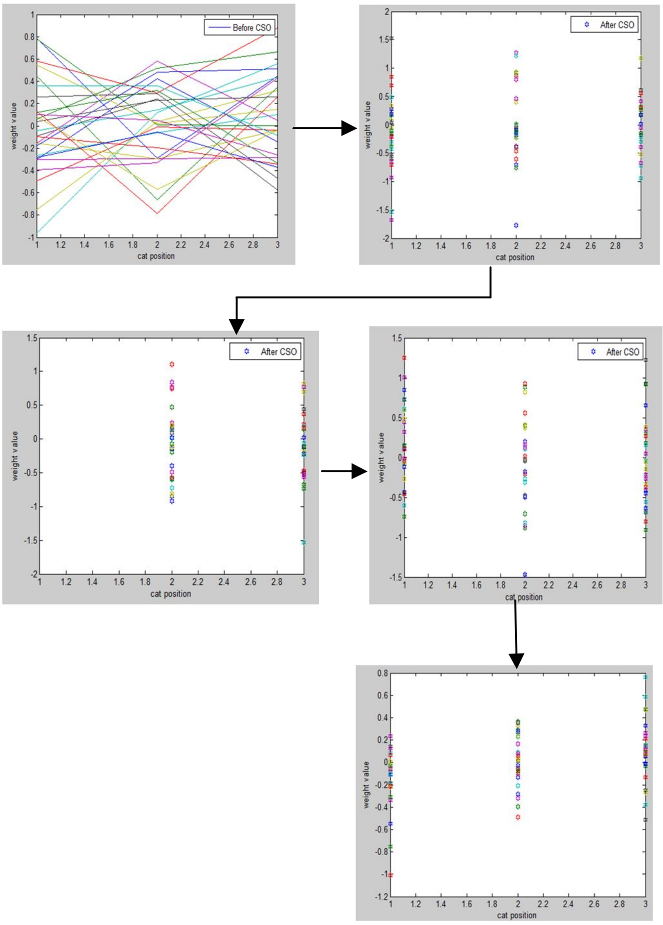

Fig 15: Updation of weights using CSO for PCA reduced Bresat cancer data

VI. Conclusion

In this paper, CSO technique has been proposed which combines the steps of dimensionality reduction to generate an optimized MLP trained network. Using the proposed algorithm a given dataset was classified in such a manner that the accuracy of classification was found to be higher once we optimize our classifier. The experimental results shows that the proposed algorithm provides better accuracy as compared to when the classifier is solely used. Although rate of accuracy achieved is higher but the computational cost involved is an area of worth investigating, therefore future work may try reduce this cost.

References

[1] Shelly Gupta, Dharminder Kumar, Anand Sharma, Performance analysis of various data mining classification techniques on healthcare data, *International Journal of Computer Science & Information Technology (IJCSIT* 2011*)*, Vol 3, pp .155-169.

[2] Peiman Mamani Barnaghi,Vahid Alizadeh Sahzabi,Azuraliza Abu Bakar, A Comparative Study for Various Methods of Classification, *International Conference on Information and Computer Networks (ICICN* 2012*)*, vol. 27, pp. 62-66.

[3] Zuyi Wang1, Yue Wang, Jianhua Xuan, Yibin Dong, Marina Bakay, Yuanjian Feng3 Robert Clarke, Eric P. Hoffman, Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data, *Oxford University Press 2006*, vol. 22, pp.755–761.

[4] John Paul T. Yusiong , Optimizing Artificial Neural Networks Using Cat Swarm Optimization Algorithm, *I. J. Intelligent Systems and Applications*, 2012; pp. 69-80.

[5] Jianwei Niu, Yiling He, Muyuan Li, Xin Zhang, Linghua Ran, Chuzhi Chao, Baoqin Zhang, A Comparative Study on Application of Data Mining Technique in Human Shape Clustering: Principal Component Analysis vs. Factor Analysis, *5th IEEE Conference on Industrial Electronics and Applications, 2010*, pp 2014-2018.

[6] S. C. Chu, and P. W. Tsai, Computational intelligence based on the behaviour of cat, *International journal of Innovative Computing, Information and Control*, 3, vol. 1, pp.163-173, 2007.

[7] Ajay Kumar Tanwani, Jamal Afridi, M. Zubair Shafiq,Muddassar Farooq, Guidelines to Select Machine Learning Scheme for Classification of Biomedical Dataset, *EvoBIO: Springer LNCS* 2009, 5483, pp. 128–139.

[8] Ng Ee Ling, Yahya Abu Hasan, Classification On Microarray Data, *Regional Conference on Mathematics, Statistics and Applications*, 2006; pp.1-8.

[9] Guo-zheng Li , Hua-Long Bu, Mary Qu Yang, Xue-Qiang Zeng, Jack Y Yang, Selecting subsets of newly extracted features from PCA and PLS in microarray data analysis, *IEEE 7th international conference on Bioinformatics and Bioengineering*, 2007,pp.1-15.

[10]L. Ladha, T. Deepa, Feature selection methods and algorithms, *International Journal on Computer Science and Engineering (IJCSE)* 2011, vol. 3 no. 5, pp.1787-1797.

[11] Helyane Bronoski Borges, Júlio Cesar Nievola, Feature Selection as a Pre-processing Step for Classification in Gene Expression Data, *Seventh International Conference on Intelligent Systems Design and Applications*, 2007, pp.157-162.

[12] Onur Inan, Mustafa Serter Uzer, Nihat Yılmaz, A new hybrid feature selection method based on association rules and PCA for detection for breast cancer, *International Journal of Innovative Computing, Information and Control,* February 2013, vol. 9, N0. 2, pp.727-739.

[13]B. Albert, A. Johnson, J. Lewis, M. Raff, K. Roberts, P.Watter, Molecular Biology of Cell, Garland Science",2004.

[14] Luai Al Shalabi, Zyad Shaaban, Basel Kasasbeh, Data Mining: A Pre-processing Engine, *Journal of Computer Science*, 2006, vol. 2, No.9, pp. 735-739.

[15] Qi Shen, Zhen Mei, Bao-Xian Ye, Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification, *Computers in Biology and Medicine*, 2009, vol. 39, pp. 646—649.

[16] Pournamashi Parhi,Debahuti Mishra, Sashikala Mishra,Kailash Shaw, A novel PSO-FLANN framework for feature selection and classification of microarray data, *International conference on modelling, optimization and computing* (ICMOC-2012),vol.38, pp.1644-1649.

[17] Pushpalata Pujari, Guru Ghasi Das, Classification and comparative study of data mining classifiers with feature selection on biomedical dataset, *Journal of Global Research in Computer Science,* 2012 vol. 3, No. 5, pp. 39-45.

[18] http://archive.ics.uci.edu/ml/

[19]Rossi, A. L. D Carv, Bio-inspired parameter tuning of MLP network for gene expression data*, Eighth international conference on Hybrid Intelligent systems*, 2008.

[20] Gilhan Kim, Yeonjoo Kim, Heuiseok Lim, Hyeoncheol Kim, An MLP-based feature subset selection for HIV-1 protease cleavage site analysis, *Artificial Intelligence in Medicine*, 2001, vol. 48, pp.83–89.