AN APPROACH FOR IRIS PLANT CLASSIFICATION USING NEURAL NETWORK

Madhusmita Swain[1], Sanjit Kumar Dash[2], Sweta Dash[3] and Ayeskanta Mohapatra[4]

[1, 2] Department of Information Technology, College of Engineering and Technology, Bhubaneswar, Odisha, India
[3]Department of Computer Science and Engineering, Synergy Institute of Engineering and Technology, Dhenkanal, Odisha, India
[4]Department of Computer Science and Engineering, Hi-tech Institute of Technology, Bhubaneswar, Odisha, India

[madhusmita39, sanjitkumar303, swetadash123, ayeskantamohapatra]@gmail.com

ABSTRACT

Classification is a machine learning technique used to predict group membership for data instances. To simplify the problem of classification neural networks are being introduced. This paper focuses on IRIS plant classification using Neural Network. The problem concerns the identification of IRIS plant species on the basis of plant attribute measurements. Classification of IRIS data set would be discovering patterns from examining petal and sepal size of the IRIS plant and how the prediction was made from analyzing the pattern to form the class of IRIS plant. By using this pattern and classification, in future upcoming years the unknown data can be predicted more precisely. Artificial neural networks have been successfully applied to problems in pattern classification, function approximations, optimization, and associative memories. In this work, Multilayer feed- forward networks are trained using back propagation learning algorithm.

KEYWORDS

IRIS dataset, artificial neural networks, Back-propagation algorithm

1. INTRODUCTION

Bioinformatics [7] is a promising and novel research area in the 21st century. This field is data driven and aims at understanding of relationships and gaining knowledge in biology. In order to extract this knowledge encoded in biological data, advanced computational technologies, algorithms and tools need to be used. Basic problems in bioinformatics like protein structure prediction, multiple alignments of sequences, phylogenic inferences, etc are inherently non-deterministic polynomial-time hard in nature. To solve these kinds of problems artificial intelligence (AI) methods offer a powerful and efficient approach. Researchers have used AI techniques like Artificial Neural Networks (ANN), Fuzzy Logic, Genetic Algorithms, and Support Vector Machines to solve problems in bioinformatics. Artificial Neural Networks [1, 11] is one of the AI techniques commonly in use because of its ability to capture and represent complex input and output relationships among data. The purpose of this paper is to provide an overall understanding of ANN and its place in bioinformatics to a newcomer to the field.

DOI : 10.5121/ijsc.2012.3107

Classification [6] is one of the major data mining processes which maps data into predefined groups. It comes under supervised learning [10] method as the classes are determined before examining the data. All approaches to performing classification assume some knowledge of the data. Usually, a training set is used to develop the specific parameters required. Pattern classification aims to build a function that maps the input feature space to an output space of two or more than two classes. Neural Networks (NN) [1, 10] are an effective tool in the field of pattern classification. Neural networks are simplified models of the biological nervous systems. An NN can be said to be a data processing system, consisting of a large number of simple, highly interconnected processing elements(artificial neurons), in an architecture inspired by the structure of the cerebral cortex of the brain. The interconnected neural computing elements have the quality to learn and thereby acquire knowledge and make it available for use. NN are an effective tool in the field of pattern classification.

This paper is related to the use of multi layer feed-forward neural networks (MLFF) [1,7,10] and back propagation algorithm[10]towards the identification of IRIS plants [8] on the basis of the following measurements: sepal length, sepal width, petal length, and petal width. A variety of constructive neural-network learning algorithms have been proposed for solving the general function approximation problem. The traditional BP algorithm typically follows a greedy strategy where in each new neuron added to the network is trained to minimize the residual error as much as possible. This paper also contains an analysis of the performance results of back propagation neural networks with various numbers of hidden layer neurons, and differing number of cycles (epochs).

2. RELATED WORK

There are some experts that understand the IRIS dataset very well. There are few experts that have done research on this dataset. The experts have mentioned that there isn't any missing value found in any attribute of this data set. The data set is complete.

Satchidananda Dehuri and Sung-Bae Cho [3] presented a new hybrid learning scheme for Chebyshev functional link neural network (CFLNN); and suggest possible remedies and guidelines for practical applications in data mining. The proposed learning scheme for CFLNN in classification is validated by an extensive simulation study. Comprehensive performance comparisons with a number of existing methods are also presented.

Saito and Nakano proposed a medical diagnosis expert system based on a multilayer ANN in [7]. They treated the network as a black box and used it only to observe the effects on the network output caused by change the inputs.

Mokriš I. and Turčaník M. [9] focussed on analysis of multilayer feed forward neural network with sigmoidal activation function, which is used for invariant pattern recognition. Analysed invariance of multilayer perceptron is used for the recognition of translated, rotated, dilated, destroyed and incomplete patterns. Parameters of analysis are the number of hidden layers, number of neurons in hidden layers and number of learning cycles due to Back-Propagation learning algorithm of multilayer feed forward neural network. Results of analysis can be used for evaluation of quality of invariant pattern recognition by multilayer perceptron.

Fernández-Redondo M. and Hernández-Espinosa C. reviewed two very different types of input selection methods: the first one is based on the analysis of a trained multilayer feed forward neural network (MFNN) and the second ones is based on an analysis of the training set. They also present a methodology that allows experimentally evaluating and comparing feature selection methods.

Two methods for extracting rules from ANN are described by Towell and Shavlik in [12]. The first method is the subset algorithm [5] which searches for subsets of connections to a node whose summed weight exceeds the bias of that node. The major problem with subset algorithms is that the cost of finding all subsets increases as the size of the ANNs increases. The second method, the MofN algorithm [13] is an improvement of the subset method that is designed to explicitly search for M-of-N rules from knowledge based ANNs. Instead of considering an ANN connection, groups of connections are checked for their contribution to the activation of a node, which is done by clustering the ANN connections.

3. IRIS PLANT DATASET

One of the most popular and best known databases of the neural network application is the IRIS plant data set which is obtained from UCI Machine Learning Repository and created by R.A. Fisher while donated by Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov) on July, 1988

The IRIS dataset [2, 4, 8] classifies three different classes of IRIS plant by performing pattern classification [8]. The IRIS data set includes three classes of 50 objects each, where each class refers to a type of IRIS plant. The attributed that already been predicted belongs to the class of IRIS plant. The list of attributes present in the IRIS can be described as categorical, nominal and continuous. The experts have mentioned that there isn't any missing value found in any attribute of this data set. The data set is complete.

This project makes use of the well known IRIS dataset, which refers to three classes of 50 instances each, where each class refers to a type of IRIS plant. The first of the classes is linearly distinguishable from the remaining two, with the second two not being linearly separable from each other. The 150 instances, which are equally separated between the three classes, contain the following four numeric attributes:

1. sepal length – **continuous**
2. sepal width - **continuous**
3. petal length - **continuous**
4. petal width – **continuous** and

the fifth attribute is the predictive attributes which is the class attribute that means each instance also includes an identifying class name, each of which is one of the following: IRIS Setosa, IRIS Versicolour, or IRIS Virginica.

The expectation from mining IRIS data set would be discovering patterns from examining petal and sepal size of the IRIS plant and how the prediction was made from analyzing the pattern to form the class of IRIS plant. By using this pattern and classification, the unknown data can be predicted more precisely in upcoming years. It is very clearly stated that the type of relationship that being mined using IRIS dataset would be a classification model. This can classify the type of IRIS plant by examining the sizes of petal and sepal. Sepal width has positive relationship with Sepal length and petal width has positive relationship with petal length. This pattern is identified with bare eyes or without using any tools and formulas. It is realized that the petal width is always smaller then petal length and sepal width also smaller then sepal length.

4. ARTIFICIAL NEURAL NETWORK

An Artificial Neural Network [10] is a computational model inspired in the functioning of the human brain. It is composed by a set of artificial neurons (known as processing units) that are interconnected with other neurons. Each connection has a weight associated that represents the influence from one neuron on the other. The first formal model of a artificial neuron was proposed in 1943 by McCulloch and Pitts. They show that such model was able to perform the computation of any computable function using a finite number of artificial neurons and synaptic weights adjustable.

The word network in Neural Network refers to the interconnection between neurons present in various layers of a system. Every system is basically a 3 layered system, which are Input layer, Hidden Layer and Output Layer. The input layer has input neurons which transfer data via synapses to the hidden layer, and similarly the hidden layer transfers this data to the output layer via more synapses. The synapses stores values called weights which helps them to manipulate the input and output to various layers. An ANN can be defined based on the following three characteristics:

1. The Architecture indicating the number of layers and the no. of nodes in each of the layers.
2. The learning mechanism applied for updating the weights of the connections.
3. The activation functions used in various layers.

There are several possible activation functions. The choice of a suitable function depends on the problem that the neuron is intended to solve. Here we have used the sigmoid as the activation function.

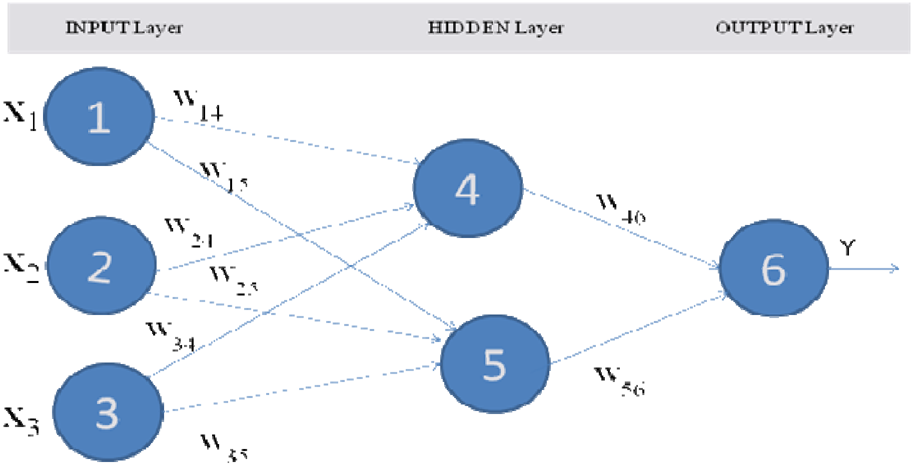

Figure 1: A simple neural network model

Using this formal model of artificial neuron, several ANN models have been proposed, including Multilayer feed-forward (MLFF). This is perhaps the most used model in a wide variety of applications.

4.1. Multilayer Feed Forward Network

Feed-forward ANNs [1, 7] allow signals to travel one way only; from input to output. There is no feedback (loops) i.e. the output of any layer does not affect that same layer. Feed forward ANNs tend to be straight forward networks that associate inputs with outputs. This type of organization is also referred to as bottom-up or top-down. They are extensively used in pattern classification [8].

This network is made of multiple layers. It possesses an input and an output layer and also has one or more intermediary layers called hidden layers. The computational units of the hidden layer are known as the hidden neurons or hidden units. The hidden layer aids in performing useful intermediary computations before directing the input to the output layer. The input layer neurons are linked to the hidden layer neurons and the weights on these links are referred to as input-hidden layer weights. Similarly, the hidden layer neurons are linked to the output layer neurons and the corresponding weights are referred to as hidden-output layer weights.

4.2. Training of Neural Network

Every NN model must be trained with representative data before using. There are basically two types of training, supervised and unsupervised [10].The basic idea behind training is to pick up set of weights (often randomly), apply the inputs to the NN and check the output with the assigned weights. The computed result is compared to the actual value. The difference is used to update the weights of each layer using the generalized delta rule [6, 10]. This training algorithm is known as 'back propagation'. After several training epochs, when the error between the actual output and the computed output is less than a previously specified value, the NN is considered trained. Once trained, the NN can be used to process new data, classifying them according to its required knowledge. When using supervised training it is important to address the following practical issues [7].

Over training- This is a serious problem where NN reduces error to an extent that it simply memorizes the data set used in training. Then it will not be possible to categorize the new data set and the generalization is not possible, which is not the required behaviour of the NN.

Validation set- To prevent over-training, validation of dataset is done. As the training precedes the training error will decreases and the result of applying the validation set improves.

Test data- Test data is a separate dataset which is used to test the trained NN to determine whether the NN has generalized the training data set accurately.

Data preparation- Sometimes it is useful to scale data before training. This will improve the training process.

5. PROPOSED MODEL

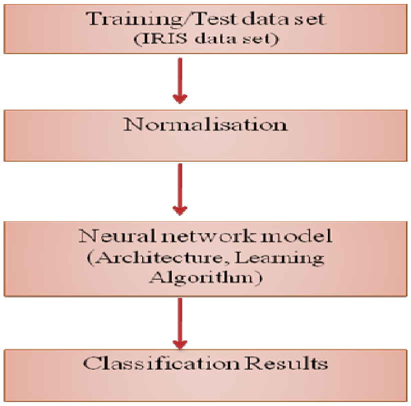

Figure 2: Proposed Model

4.1 Dataset Construction

This projects initial activity was towards the separation of the consolidated data set, into a training set, and a testing set. There are 150 instances with 25 in each one of three classes. Existing 4 numerical attributes for identification of the classes are sepal length, sepal width, and petal length and petal width in centimetres. The classes are IRIS Setosa, IRIS versicolor, IRIS virginica. In this application 75 data of each class are used for training and 75 are separated for test purpose.

4.2 Normalization

Second step is normalization. Since we are using NNs, it only demands numerical inputs. Using the following formula we have converted the value in to appropriate range.

$$x_{ij} = (x_{ij} - colm_{min}) /(colm_{max} - colm_{min})$$

- Calculate the $colm_{min}$ minimum value of the column
- Subtract the $colm_{min}$ from each of the data
- Calculate the value of ($colm_{max}$ - $colm_{min}$)
- Calculate the new value by dividing the two previous result

Where

$colm_{min}$ = minimum value of the column
$colm_{max}$ =maximum value of the column
x_{ij} = x(data item)present at ith row and jth column

Third step is NN model. In this work, we have used MLFF and learning algorithm as back propagation algorithm.

5.3 Backpropagation Algorithm

Back-propagation [10] training algorithm when applied to a feed forward multi-layer neural network is known as Back propagation neural network. Functional signals flows in forward direction and error signals propagate in backward direction. That's why it is Error Back Propagation or shortly Back Propagation network. The activation function [10] that can be differentiated (such as sigmoid activation function) is chosen for hidden and output layer computational neurons. The algorithm is based on error-correction rule. The rule for changing values of synaptic weights follows generalized delta rule.

Steps:

Initialize all weights in network
//Propagate the inputs forward

- For each input layer unit *j*
- $O_j = I_j$ //output of an input unit its actual input value
- For each hidden or output layer unit *j*
- $I_j = \sum_i w_{ij} O_i$ // the net input of unit *j*
- $O_j = 1/(1+e^{-Ij})$ // the output of each unit *j*
- //Back propagate the errors

- For each unit j in the output layer
- $Err_j = O_j\,(1\text{-}O_j)(T_j - O_j)$ // the error
- For each unit j in the hidden layer, from the last to the 1st hidden layer
- $Err_j = O_j\,(1 - O_j) \sum_k Err_k\, w_{jk}$ //error with respect to the
- Next higher layer , k
- for each weight w_{ij} in network
- $\Delta w_{ij} = (\eta)\, \mathrm{Err}_j\, O_i$ //weight increment
- $W_{ij} = w_{ij} + \Delta w_{ij}$ //weight update

5. Simulation Results

The simulation process is carried on a computer having Dual Core processor and 2 GB of RAM. The MATLAB version used is R2010a. The IRIS dataset (downloaded from the UCI repository, www.ics.uci.edu, which is a 150×4 matrix, is taken as the input data. Out of these 150 instances, 75 instances were used for training and 75 for testing. Under supervised learning, the target of the first 25 instances is taken as 0, for the next 25 instances as 0.5 and for the last 25 instances as 1. The network architecture taken was 4×3×1, i.e, the input layer has 4 nodes, the hidden layer has 3 nodes and the output layer has 1 node. The tolerance value taken was 0.01.

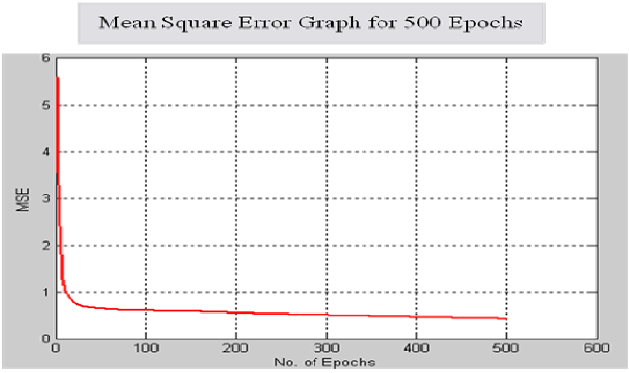

Figure 3: For 500Epochs vs. MSE at η=0.7

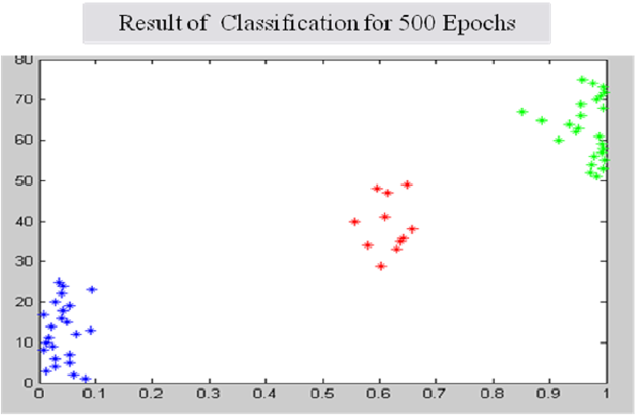

Figure 4: Result of classification for 500 epochs at η=0.7

Table 1: Test data classification for 500 Epochs

IRIS Plant	Total	Classified	Not Classified
Setosa	25	24	1
Versicolor	25	11	13
Virginnica	25	25	0

In 500 iterations, out of 25 instances of Setosa class only 24 are classified, out of 25 instances of Versicolor class only 11 are classified and out of 25 instances of Virginica class all instances are classified. Hence accuracy rate is 83.33%.

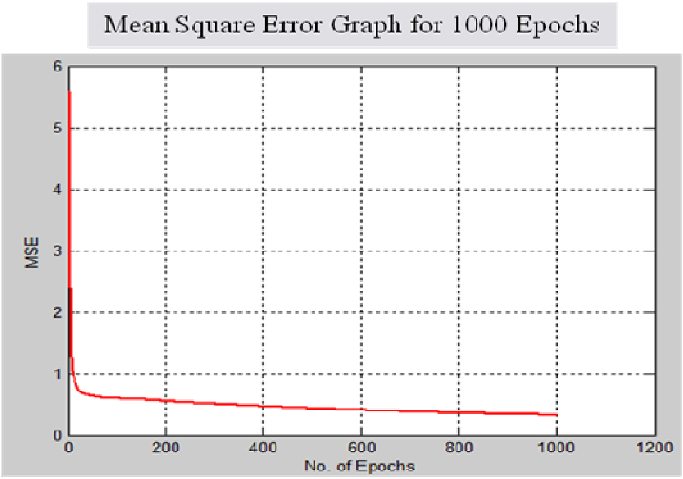

Figure 5: For 1000 Epochs vs. MSE at η=0.7

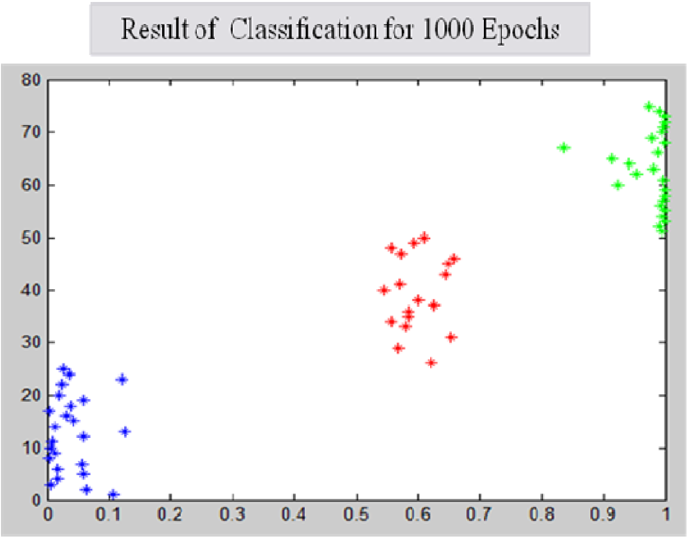

Figure 6: Result of classification for 1000 epoch at η=0.7

Table 2: Test data classification for 1000 epochs

IRIS Plant	Total	Classified	Not Classified
Setosa	25	24	1
Versicolor	25	18	7
Virginnica	25	24	1

In 1000 iterations, out of 25 instances of Setosa class only 24 are classified, out of 25 instances of Versicolor class only 18 are classified and out of 25 instances of Virginica class 24 instances are classified. So accuracy rate = 87.33%

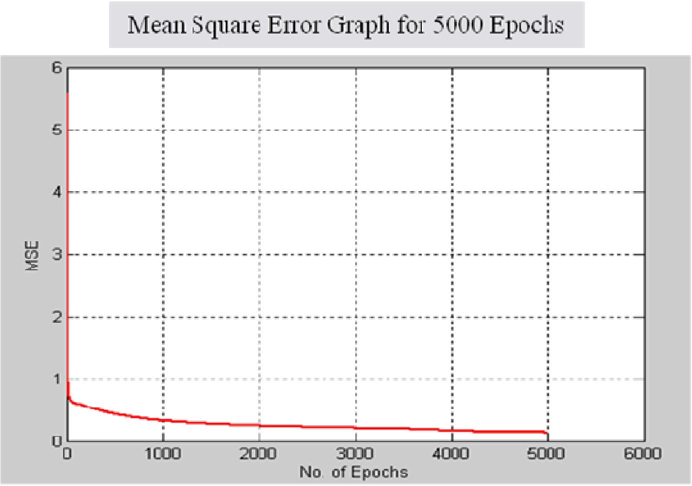

Figure 7: For 5000 Epochs vs. MSE at η=0.7

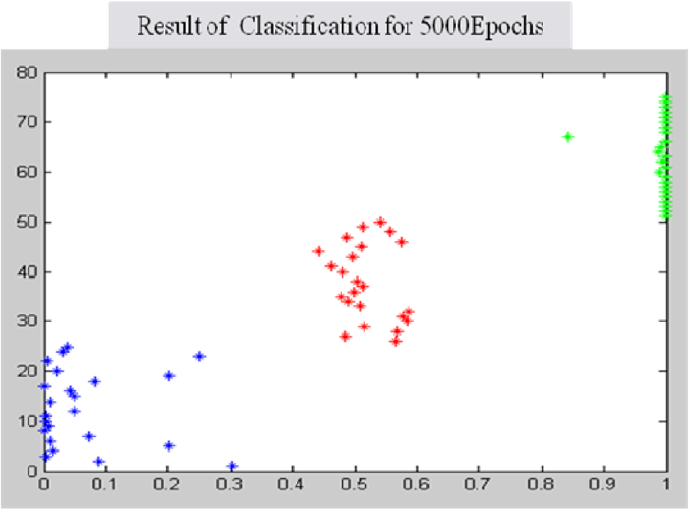

Figure 8: Result of classification for 5000 epochs at η=0.7

Table 3: Test data classification for 5000 epochs

IRIS Plant	Total	Classified	Not Classified
Setosa	25	23	2
Versicolor	25	22	3
Virginnica	25	25	0

In 5000 iteration, out of 25 instances of Setosa class only 23 are classified, out of 25 instances of Versicolor class only 22 are classified and out of 25 instances of Virginica class 25 instances are classified. So accuracy rate = 96.66%

6. CONCLUSIONS

The Multi Layer Feed Forward Neural network gives us a satisfactory result, because it is able to classify the three different types of IRIS of 150 instances with just few errors for the other one. From the graphs we observe that Back propagation Algorithm gives the best accuracy .The no. of epochs required to train the neural network range from 500 to 50000 and the accuracy ranges from 83.33% to 96.66%.

From the above results, graphs and discussion, it is concluded that Multi Layer Feed Forward Neural Network (MLFF) is faster in terms of learning speed and gave a good accuracy, i.e., has the best trade-off between speed and accuracy. So, for faster and accurate classification, Multi Layer Feed Forward Neural Networks can be used in many pattern classification problems.

REFERENCES

[1] Aqel, M.M., Jena, R.K., Mahanti, P.K. and Srivastava (2009) 'Soft Computing Methodologies in Bioinformatics', European Journal of Scientific Research, vol.26, no 2, pp.189-203.

[2] Avcı Mutlu, Tülay Yıldırım(2003) 'Microcontroller based neural network realization and IRIS plant classifier application', International XII. Turkish Symposium on Artificial Intelligence and Neural Network

[3] Cho, Sung-Bae.and Dehuri, Satchidananda (2009) 'A comprehensive survey on functional link neural network and an adaptive PSO–BP learning for CFLNN, Neural Comput & Applic' DOI 10.1007/s00521-009-0288-5.

[4] Fisher, A. W., Fujimoto, R. J. and Smithson, R. C.A. (1991) 'A Programmable Analog Neural Network Processor', IEEE Transactions on Neural Networks, Vol. 2, No. 2, pp. 222-229.

[5] Fu, L.(1991) ' Rule learning by searching on adapted nets. In Proceedings of National Conference on Artificial Intelligence' Anaheim, CA, USA, pp. 590-595.

[6] Han, J. and Kamber, M. (2000)' Data Mining: Concepts and Techniques' , 2nd ed. Morgan Kaufmann.

[7] Dr.Hapudeniya, Muditha M. MBBS(2010),'Artificial Neural Networks in Bioinformatics' Medical Officer and Postgraduate Trainee in Biomedical Informatics , Postgraduate Institute of Medicine , University of Colombo ,Sri Lanka, vol.1, no 2,pp.104-111.

[8] Kavitha Kannan ‘Data Mining Report on IRIS and Australian Credit Card Dataset’, School of Computer Science and Information Technology, University Putra Malaysia, Serdang, Selangor, Malaysia.

[9] Marček D., ‘Forecasting of economic quantities using fuzzy autoregressive models and fuzzy neural networks’, Vol.1, pp.147-155.1

[10] Pai, G. V and Rajasekaran, S, (2006), ‘Neural Networks, Fuzzy Logic and Genetic Algorithms Synthesis and Applications’, 6th ed, Prentice Hall of India Pvt. Ltd.

[11] Rath, Santanu and Vipsita, Swati (2010) ‘An Evolutionary Approach for Protein Classification Using Feature Extraction by Artificial Neural Network’, Int’l Conf. on Computer & Communication TechnologyıICCCT’10.

[12] Towell, G.G. and Shavlik, J.W. (1993), ‘Extracting refined rules from knowledge-based neural networks’ Mach. Learn, Vol.13, pp.71-101.

[13] Towell, G.G; Shavlik, J.W. (1994) ‘Knowledge-based artificial neural networks’, Artif. Intell. Vol.7, pp.119-165.