Second International Conference on Computational Intelligence, Modelling and Simulation

Suitable MLP Network Activation Functions for Breast Cancer and Thyroid Disease Detection

I.S.Isa, Z.Saad, S.Omar, M.K.Osman,K.A.Ahmad
Faculty of Electrical Engineering
Universiti Teknologi Mara (UiTM)
Kampus Pulau Pinang, 13500
Pmtg. Pauh, P.Pinang, Malaysia
izasazanita@ppinang.uitm.edu.my

H.A.Mat Sakim
School of Electrical and Electronics Engineering
Universiti sains Malaysia Kampus Transkrian
14300 Nibong Tebal, Pulau Pinang, Malaysia
amylia@eng.usm.my

Abstract –**This paper presents a comparison study of various MLP activation functions for detection and classification problems. The most well-known (Artificial Neural Network) ANN architecture is the Multilayer Perceptron (MLP) network which is widely used for solving problems related to detection and data classifications. Activation function is one of the elements in MLP architecture. Selection of the activation functions in the MLP network plays an essential role on the network performance. A lot of studies have been conducted by reseachers to investigate special activation function to solve different kind of problems. Therefore, this paper intends to investigate the activation functions in MLP networks in terms of the accuracy performances. The activation functions under investigation are sigmoid, hyperbolic tangent, neuronal, logarithmic, sinusoidal and exponential. Medical diagnosis data from two case studies; thyroid disease and breast cancer, have been used to test the performance of the MLP network. The MLP networks are trained using Back Propagation learning algorithm. The performance of the MLP networks are calculated based on the percentage of correct classificition. The results show that the hyperbolic tangent function in MLP network had the capability to produce the highest accuracy for detecting and hence classifying breast cancer data. Meanwhile, for thyroid disease classification, neuronal function is the most suitable function that performed the highest accuracy in MLP network.**

***Keywords* – Activation function; Data classification; Neural network applications, Multilayer perceptron network.**

I. INTRODUCTION

Multilayer perceptron network (MLP) is one of the Artificial Neural Network (ANN) models that has great track of impacts at solving a variety of problems. Due to its robust capability and simple structure design, it has been widely used in many applications. However, in some cases, MLP networks fail to provide a good solution. This happens due to improper selection of architecture, initialization of weights or data selection. Another factor that affects the training process is the selection of transfer function [10].

An activation function in MLP network is typically a non-linear function that transforms the weighted sum of inputs to an output value. An activation function or a transfer function for the hidden nodes in MLP is needed to introduce nonlinearity into the network. The selection of activation function might significantly affect the performance of a training algorithm. Some researchers have investigated to find special activation function to simplify the network structure and to accelerate convergence time [7]. An activation function for MLP network with back propagation algorithm should have several important characteristics. It should be continuous, differentiable and monotonically non-decreasing [1]. Table 1 summarizes various activation functions used in neural network.

Research on effect of activation function for neural network has received a lot of attention in the past literatures. In [14], a Cosine-Modulated Gaussian activation function for Hyper-Hill neural networks has been proposed. The study compared the Cosine-Modulated Gaussian, hyperbolic tangent, sigmoid and symsigmoid function in cascade correlation network to solve sonar benchmark problem. Joarder and Aziz [6] proved that logarithmic function is able to accelerate back propagation learning or network convergence. The study has solved XOR problem, character recognition, machine learning database and encoder problem using MLP network with back propagation learning. Wong *et al* [11] investigated the neuronal function for network convergence and pruning performance. Periodic and monotonic activation functions were chosen for the analyses of multilayer feed forward neural networks trained by Extended Kalman Filter (EKF) algorithm. The study has solved multicluster classification and identification problem of XOR logic function, parity generation, handwritten digit recognition, piecewise linear function approximation and sunspot series prediction. Piekniewski and Tybicki [10] employed different activation functions in MLP networks to determine the visual comparison performance. The study has shown that log-exponential function has been slowly accelerated but it was effective in MLP network with back propagation learning. Barycentric plotting was used as a simple projection scheme to measure the neural network performance.

Some researchers had also proposed special activation function such as logarithmic [4], Adaptive Spline [5], Type-2 Fuzzy [7], Elementary Transcendental [8], Hermite Polynomial [9], periodic and monotonic [11], Novel

978-0-7695-4262-1/10 $26.00 © 2010 IEEE
DOI 10.1109/CIMSiM.2010.93

IEEE computer society

Adaptive [12] and Cosine-modulated Gaussian activation function [14]. However, the most recommended activation function among neural network researchers for MLP network is sigmoid or hyperbolic tangent [6],[10].

TABLE 1: VARIOUS TYPES OF ACTIVATION FUNCTIONS USED IN THE NEURAL NETWORKS

Name	Function	Characteristics
Binary threshold	$\delta(x_j) = \begin{cases} 1 & x_j \geq 0 \\ 0 & x_j < 0 \end{cases}$	Non-differentiable, step-like, $s_j \in \{0,1\}$
Bipolar threshold	$\delta(x_j) = \begin{cases} 1 & x_j \geq 0 \\ -1 & x_j < 0 \end{cases}$	Non-differentiable, step-like, $s_j \in \{-1,1\}$
Linear	$\delta(x_j) = \alpha_j x_j$	Differentiable, unbounded, $s_j \in \{-\infty,\infty\}$
Linear threshold	$\delta(x_j) = \begin{cases} 0 & x_j < 0 \\ \alpha_j x_j & 0 < x_j < x_m \\ 1 & x_j \geq x_m \end{cases}$	Differentiable, piece-wise linear, $s_j \in \{0,1\}$
Sigmoid	$\delta(x_j) = \dfrac{1}{1+e^{-\lambda_j x_j}}$	Differentiable, monotonic, smooth, $s_j \in \{0,1\}$
Hyperbolic tangent	$\delta(x_j) = \tanh(\lambda_j x_j)$	Differentiable, monotonic, smooth, $s_j \in \{-1,1\}$
Gaussian	$e^{-(x_j - c_j)^2 / 2\sigma^2 j}$	Differentiable, non-monotonic, smooth, $s_j \in \{0,1\}$
Stochastic	$\delta(x_j) = \begin{cases} +1 \\ -1 \end{cases}$ $probability : P(x_j)$ $probability : 1 - P(x_j)$	Non-deterministic step-like, $s_j \in \{0,1\}$ or $\{-1,1\}$

II. METHODOLOGY

The objective of this study is to measure performances of MLP networks on classifying diagnosis data of thyroid disease and breast cancer using MLP networks and Levenberg-Marquardt (LM) as the training algorithm. In this paper, LM training algorithm is adopted for updating each connection weights of units. LM algorithm has been used in this study due to the reason that the training process converges quickly as the solution is approached. For this study, sigmoid, hyperbolic tangent, neuronal, logarithmic, sinusoidal and exponential functions are applied in the learning process.

A. Multilayer Perceptron Network

The MLP is probably the most often considered member of the neural network family. The architecture of the MLP is as shown in Fig. 1. It consists of an input layer, one or more hidden layers and an output layer. The number of hidden layers can be changed depending on problem data under training process. The output nodes can also be changed depending on classification of target output. The most common training procedure for this model is by supervised learning specifically the back propagation algorithm. The back propagation algorithm employs the method of gradient descent, which tends to minimize the mean squared error between the output of an MLP network and the desired output.

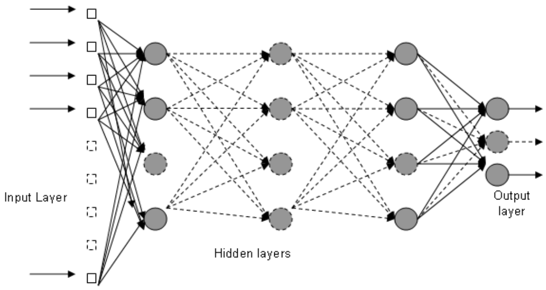

Figure 1: Multilayer perceptron network

The output of network, *y (t)* at output layer *m*, is given by :

$$y_l(t) = h\left(\sum_{j=1}^{n_{m-1}} w_{jl}^m v_j^{m-1}(t)\right);1 \leq l \leq n_m \qquad (1)$$

n_k : number of nodes in the *k*-th layer

n_m : number of nodes in the output layer

w's : weights

h (●): activation function

A common activation function in the nodes of the hidden or output layer is the sigmoid function. However, other functions such as hyperbolic tangent function or quadratic function can also be employed. The activation function can be the same for all the nodes or a different function can be employed for each of the nodes. In some applications, no activation function is employed at the output node. The goal is to train the MLP network to achieve a balance between the ability to respond correctly to the input patterns that are used for training and the ability to provide good response to the input that is similar. Back propagation learning is one of the most important types of learning in feed forward network. It is a systematic method for training a multilayer network such as MLP network.

B. Levenberg-Marquardt Training Algorithm

Levenberg-Marquardt (LM) algorithm is typically the fastest of training algorithms [15]. LM is a Hessian based algorithm for nonlinear least squares optimization. Hessian-based algorithms are used to allow ANNs to learn more suitable features of a complicated mapping [16]. The training process converges quickly as the solution is approached. Due to the reason, LM algorithm has been used in this study. The good aspect of Levenberg-Marquardt (LM) is that the determination of the new points is actually a compromise between a step in the direction of the steepest descent.

C. Activation Functions

The activation functions also known as transfer function are typically a non-linear function that transforms the weighted sum of the inputs (the internally generated sum) to an output value. Sometimes different activation functions [4-14] are acquired for different networks so that it resulting in better performances. An activation function or transfer functions for the hidden nodes in MLP are needed to introduce nonlinearity into the network. There are two types of activation functions that are usually used in neural network, linear and nonlinear activation function. The choice of activation function is important for performance of training algorithm. An activation function for a backpropagation network should have several important characteristics. It should be continuous, differentiable and monotonically non-decreasing [1]. By using backpropagation learning algorithm, the activation function used must be differentiable so that the function is bounded in certain ranges of limits.

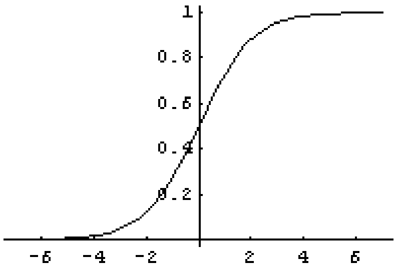

Figure 2: Sigmoid activation function

When designing a network, the initial transfer function is applied to each layer of the network. For most networks, the recommended transfer function is sigmoid or hyperbolic tangent. The most common activation function used in MLP is sigmoid activation function and commonly gives good results. As shown in Fig. 2, sigmoid activation function saturates to 0 or 1, which are the values used to indicate membership in an output class. The input for this function is varied from - ∞ to + ∞ but usually bounded to certain value. The expression of this function is given by (2):

$$f_{sig}(net_i) = \frac{1}{1+e^{-S_s.net_i}} \qquad (2)$$

Where

net_i is the net input to the *i*th neuron

S_s is the slope of the sigmoid function

Another most common activation function used in backpropagation learning is hyperbolic tangent activation function. Hyperbolic tangent activation function is similar to sigmoid activation function in transforming the net input to saturate output class between -1 and +1. Fig. 3 presents the hyperbolic tangent activation function and expression of this function is:

$$f_{th}(net_i) = \frac{e^{net} - e^{-net}}{e^{net} + e^{-net}} \qquad (3)$$

where net_i is the net input to the *i*th neuron

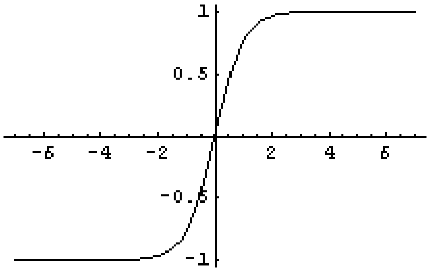

Figure 3: Hyperbolic tangent activation function

A neuronal activation function is a periodic activation function suggested by [11]. This activation function has the combination of sigmoid and sinusoidal function. In Fig. 4, the solid line represents the neuronal function, dashed line represents sigmoid function and dotted line represents sinusoidal function. The expression of this function is given as follows:

$$f_P(net_i) = \frac{1}{1+e^{-S_p.\sin(r_p.net_i)}} \qquad (4)$$

where S_P controls the slope while r_P governs the frequency of the resultant function.

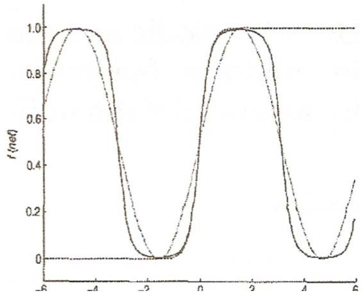

Figure 4: Neuronal activation function

Fig. 5 shows logarithmic function which has been proposed by [4]. This activation function is proposed with an idea to have larger derivatives as the output reaches extreme values. The expression of this function is [6]:

$$f(net) = \begin{cases} \ln(net+1) & net \geq 0 \\ -\ln(-net+1) & net < 0 \end{cases} \qquad (5)$$

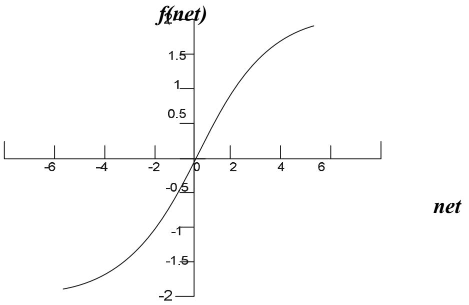

Figure 5: Logarithmic activation function

A sinusoidal function is used instead of sigmoid because it leads to what has been termed a "Generalized Fourier Analysis". The sine function takes the trigonometric sine of the input. Consider a back propagation network with just one output, the learning procedure can be thought of as synthesizing a continuous function $y = g(x)$ by showing it a discrete set of (x,y) pairs. The network configures itself to output a correct value (desired output) for each example input pair. When a previously unseen input pattern is presented to the network, the network in effect performs a non-linear interpolation and produce an output which a reasonable function value.

Fig. 6 shows a sinusoidal activation function. When a sine function is used instead of a sigmoid, the learning procedure seems to perform a mode decomposition where it discovers the most important frequency components of the function described by discrete set of input output examples . The expression of this function is described as follows:

$$f(net) = \sin(net) \qquad (6)$$

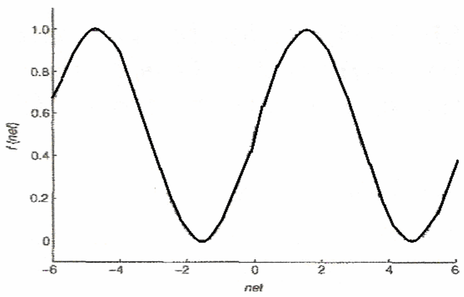

Figure 6: Sine activation function

The exponential function is a function in mathematics. The application of this function to a value *x* is written as *exp (x)*. Fig. 7 shows an exponential function employed for MLP network. This function is ideal in using with a radial unit that models a Gaussian function for radial basis function (RBF) network. However, this function is employed to MLP network to investigate its effect. The expression of exponential function is given as follows:

$$f(net) = \exp(net) \qquad (7)$$

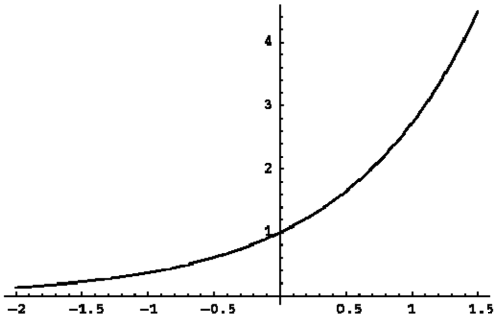

Figure 7: Exponential activation function

III. Proposed System

The purpose of using neural network is to be able to detect and hence classify the data that are too complex for the traditional statistical models. This paper has chosen MLP network as the classifier. The MLP model was simulated using NeuralWare software where the networks are feed forward neural networks with one or more hidden layers. This means that one hidden layer MLP is almost always sufficient to approximate any continuous function up to certain accuracy [2]. The advantages of MLP are, their abilities to learn and give the better performance especially in the case of classification are proven. In additions the construction of MLP is simple. The ability of an MLP network to classify data efficiently and make decisions based on the classification results is one of the distinguishing features that resemble human intelligence. The MLP network has to be train before it able to perform specific task with less error. Theoretically, a single MLP network with the best performance at the fewest number of hidden neurons will be selected as the best ANN to represent a problem [3].

This work uses 246 and 228 dataset of breast cancer disease obtained from the University of Wisconsin Hospital, Madison by Dr. William H. Wolberg for training and testing. This dataset has been used in the research by [17] to analyze the pattern recognition of medical diagnosis. Fig. 8 shows the nuclei cells being detected using image processing analysis to classify between normal cells and cancer cells.

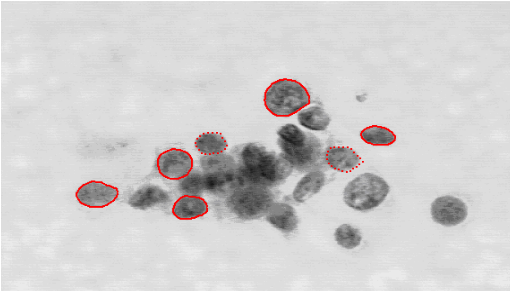

Figure 8: Image detection and classification of nuclei cells (Courtesy from Dept of Surgery, Human Oncology and Computer Sciences, University of Wisconsin Madison, USA)

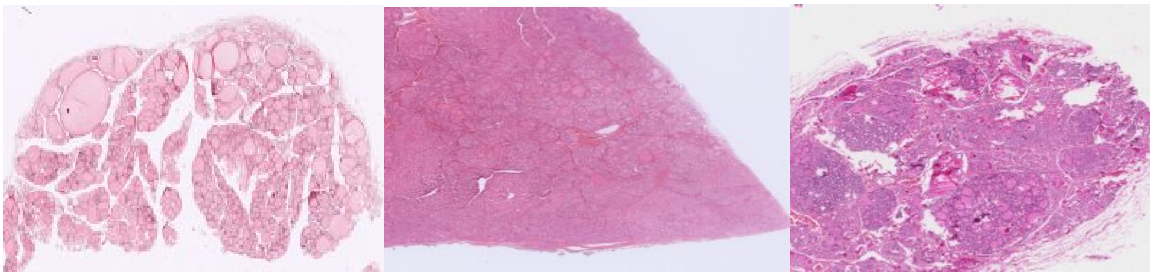

Figure 9: Image detection and classification of thyroid glands a) normal thyroid b) hyper thyroid c) hypo thyroid (Courtesy from Dept of Pathology, James Cook University, Australia)

This paper is using the same dataset to perform new analysis of MLP classification comparing 6 different activation functions. All the 9 attributes were in the range between 2 and 9. Summary of data division is described in Table 2. The breast cancer data had two output classes. The predicted outputs were continuous values and converted to absolute values representation within the threshold ranges.

The other data used was thyroid disease data. The original data thyroid disease came from James Cook University [18] with output target 1, 2 and 3. Three classes of thyroid divisions came from five attributes with continuous values between -0.7 to 141. Fig. 9 present the images of thyroid glands detected using image processing analysis to classify whether normal, hyper or hypo thyroid. The data was divided into 215 for training and 200 for testing. The two samples of data were categorized into three classes. The summary of data division is described in Table 3.

TABLE 2: DATA DIVISION OF BREAST CANCER DISEASE

	Training Data	Testing Data
Normal Cells	123	128
Cancer Cells	123	100
Total Data	246	228

TABLE 3: DATA DIVISION OF THYROID DISEASE

	Training Data	Testing Data
Normal Thyroid	150	150
Hyper Thyroid	35	25
Hypo Thyroid	30	25
Total Data	215	200

In the study, the networks with bias connection were trained and tested with the following parameters as suggested by [19]:

Learning rate, $\eta = 0.3$
Momentum, $\alpha = 0.4$
Gain scale, $\lambda = 1.0$
Tolerance, $\tau = 0.1$

Using the chosen initial weights, MLP networks were trained and tested until the networks were converged. Hidden node that gives the best performance for training and testing was chosen as the best node for the network of each activation function. This method was implemented for training and testing thyroid and breast cancer data set.

The performances of each MLP network will be evaluated in terms of percentages for correct classification, defined as the different between the actual and the simulated results. The performance of correct classification is defined as in Equation 8.

$$\%Correct\ Classification = \frac{Total\ of\ Correct\ Data}{Total\ Number\ Data} \times 100\% \qquad (8)$$

IV. RESULTS AND DISCUSSION

The main objective of this study was to compare the performance of an MLP network by changing different activation functions to the network. The excellent networks, for example, ones that give the highest percentage of correct classification during learning process will be selected as the chosen MLP network for the system. The network would be stopped from training and testing once the accuracy performance shown remained or decreased or converged.. Aaccuracy of MLP networks was analyzed to determine the best function that suited the network in order to perform correct classifying breast cells.

The best hidden neuron that give the best performance of MLP networks of each investigated function were shown in Table 4. The highest testing data accuracy of 97.0% has been achieved by using hyperbolic tangent activation function while sigmoid function achieved the lowest testing accuracy of 95.6%. Table 5 shows performance comparison of thyroid disease classification data for each investigated activation function. The highest testing data accuracy achieved was 94.0% given by neuronal activation function.

TABLE 4: MLP PERFORMANCE OF BREAST CANCER CLASSIFICATION FOR VARIOUS ACTIVATION FUNCTIONS

Functions	Sigmoid	Hyperbolic tangent	Neuronal	Logarithmic	Sine	Exp
No. of Hidden nodes	7	5	25	20	29	4
Training Accuracy (%)	97.6	97.2	97.6	97.2	98	97.2
Testing Accuracy (%)	95.6	97	96	96	96.5	96.5

The results produced from both case studies showed that an MLP network was able to achieve accuracy performance more than 90.0%. The MLP networks trained with the breast cancer data gave best accuracy of 97.0% using the hyperbolic tangent function. However, the highest accuracy of 94.0% for thyroid disease classification was given by the MLP network trained with neuronal function. MLP network that gave highest percentage of correct classification with least number of hidden nodes were chosen as the best performance. Table 6 summarizes the best activation functions for both cases; breast cancer and thyroid diseases.

TABLE 5: MLP PERFORMANCE OF THYROID DIEASES CLASSIFICATION FOR VARIOUS ACTIVATION FUNCTIONS

Functions	Sigmoid	Hyperbolic tangent	Neuronal	Logarithmic	Sine	Exp
No. of Hidden nodes	10	20	9	29	7	23
Training Accuracy (%)	89.3	95.8	94.9	90.7	94	76.7
Testing Accuracy (%)	84	92	94	86	93	63

TABLE 6: THE BEST SELECTED ACTIVATION FUNCTIONS FOR DIFFERENT CASE STUDIES

Case study	Activation functions	No hidden neurons	Training accuracy (%)	Testing accuracy (%)
Breast cancer (two classes output)	hyperbolic tangent	5	97.2	97
Thyroid diseases (three classes output)	neuronal	9	94.9	94

V. CONCLUSION

The activation functions in MLP networks had been investigated to determine the most suitable function to solve classification problems. Results obtained from each network were compared to determine the best function that is one that gave the highest accuracy.

In classifying the breast cancer cells, hyperbolic tangent function had shown the capability of achieving the highest accuracy of an MLP performance. For networks trained with thyroid disease data set, the highest accuracy achieved during testing was 94.0% by neuronal function using 9 hidden neurons. From the findings, it can be concluded that the hyperbolic tangent function is suitable for MLP network to classify data of two classes. Neuronal function is applicable in MLP network for classifying data of three classes. The activation functions can be employed in the future research to obtain higher classification accuracy.

REFERENCES

[1] M. Rosen-Zvi, M. Biehl, and I. Kanter, "Learnability of periodic activation functions: General results", *Physical Review E, 58*, (3), pp. 3606-3609,1998.

[2] M. Y. Mashor, "Performance Comparison Between Back Propagation, RPE And MRPE Algorithms For Training MLP Network." School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 2002.

[3] S. N Sivanandam, S. Sumathi & S. N. Deepa, "Introduction to Neural Network Using Mathlab", India: Mc Graw Hill, 1998.

[4] J. Bilski, "The Backpropagation learning with logarithmic transfer function", *Proceeding of 5th Conf. On Neural Networks and Soft Computing, Poland*, pp. 71-76, Last assessed June 2000.

[5] P. Campolucci, F. Capparelli, S. Guatnieri, F. Piazza and A. Uncini, "Neural networks with Adaptive Spline activation function.", *Proceedings of IEEE Conference on Electrotechnical:, MELECON '96. Vol. 3*, pp 1442-1445, 1996

[6] K. Joarder and S.M.Aziz, "A note on activation function in multilayer feedforward learning", *Proceedings of International Joint Conference on Neural Networks: IJCNN '02.* Vol. 1, pp 519-523, 2002.

[7] M.Karakose and E.Akin, "Type-2 Fuzzy activation function for Multilayer Feedforward neural networks", *Proceedings of International Conference on Systems, Man and Cybernatics. Vol. 4*, pp 3762-3767, 2004.

[8] T.Kim, and T.Adali, "Complex backpropagation neural network using Elementary Transcendental activation functions", *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing: ICASSP '01*. Vol. 2, pp 1281-1284, 2001.

[9] L.Ma and K.Khorasani, "Constructive feedforward neural networks using Hermite Polynomial activation functions", *IEEE Transaction on Neural Networks, Vol.4*, pp 821-832, July 2005.

[10] F.Piekniewski and L.Tybicki, "Visual comparison of performance for different activation functions in MLP networks", *Proceedings of International Joint Conference on Neural Networks: IJCNN '04*. Vol. 4, pp 2947-2952, 2004.

[11] K.W.Wong, C.S., Leung and S.J. Chang, "Use of periodic and monotonic activation functions in multilayer feedforward neural networks trained by extended Kalman filter algorithm*", Proceedings of IEE on Vision, Image and Signal Processing.* Vol. 149, pp 217-224, 2002.

[12] S.Xu and M.Zhang, "A Novel Adaptive activation function", *Proceedings of International Joint Conference on Neural Networks: IJCNN '01*. Vol. 4, pp 2779-2784, 2001.

[13] X.Ying, "Role of activation function on hidden units for sample recording in three-layer neural networks", *Proceedings of International Joint Conference on Neural Networks: IJCNN*, Vol. 3, pp 69-74, 1990.

[14] S.Lee and C.Moraga, "A Cosine-Modulated Gaussian activation function for Hyper-Hill neural networks*", Proceedings of 3rd International Conference on Signal Processing*. Vol. 2, pp 1397-1400, 1996.

[15] M. T. Hagan and M. B. Menhaj, "Training feed forward Networks with the Marquardt algorithm", *IEEE Transaction on Neural Network*, Vol. 5, No. 6, pp. 989-993, 1994.

[16] A.A. Suratgar, M.B. Tavakoli, and A. Hoseinabadi, "Modified Levenberg-Marquardt method for neural networks training", *Proc. World Academy of Science, Engineering and Technology*, pp. 46-48, 2005.

[17] O.L.Mangasarian, R.Setiono, and W.H. Wolberg, "Pattern Recognition Via Linear Programming Theory and Application to Medical Diagnosis*", Technical Reports of CiteSeer, Physical Review E*, pp. 1-10, 1990.

[18] D.Coomans, I.Broeckaert, M.Jonckheer and D.L.Massart, "Comparison of Multivariate Discrimination Techniques for Clinical data – Application to the thyroid Functional State", *Methods of Information in Medicine*, Vol. 22, pp. 93-101, 1983.

[19] S.Kumar, "Neural Networks A Classroom Approach*"*, Mc Graw Hill, International Edition, 2005.