Int J Pharm Bio Sci 2013 Jan; 4(1): (B) 965 - 972

Research Article **Biostatistics**

International Journal of Pharma and Bio Sciences

ISSN 0975-6299

A COMPARATIVE STUDY OF NEURAL NETWORK ARCHITECTURES FOR PREDICTING GENE EXPRESSION IN M. TUBERCULOSIS

VENKATESAN P[1] AND TINTU THOMAS.*[2]

[1]Scientist E and Deputy Director, Department of Statististics, National Institute for Research in Tuberculosis-ICMR, Chennai,India
[2]Department of Community Medicine, C.U.Shah Medical College, Gujarat,India.

ABSTRACT

Identification and classification of differentially expressed genes is a challenging process.The study was performed to find applicability of supervised feed forward neural networks to solve classification problems in microarray data.We used two neural network learning algorithms namely RBF and MLP for the classification of gene expression of mycobacterium tuberculosis.The result showed that MLP and RBF classifier have similar performance and both have given high prediction.

KEYWORDS: microarray, artificial neural network, multilayer perceptron, radial basis function

TINTU THOMAS.
Department of Community Medicine, C.U.Shah Medical College, Gujarat, India.

INTRODUCTION

Microarrays are invaluable to identify gene alteration at the level of m RNA, enabling assessment of thousands of genes at the same time. Microarrays can help in estimating the amount of protein in the cell and lot of information can be derived from this technology. Microarrays give large scale, high through put gene expression of thousands of genes. Determining the expression levels of thousands of genes and classification of expression levels into categories are very essential for supporting clinical decisions. Standard statistical methodologies in classification or prediction do not work well when the number of variables exceeds the number of samples. The machine learning algorithms based tools are very useful for classification and prediction of microarray gene expression data. An artificial neural network is one of the prediction or classification tool. Golub et al. utilized a nearest-neighbor classifier method for the classification of acute myeloid lymphoma (AML) and acute leukemia lymphoma (ALL) in children[1]. Dudoit et al. performed a systematic comparison of several discrimination methods for classification of tumors based on microarray experiments[2]. Wang et al. worked on the classification of genomic data using MLP[3].Machine learning algorithms have been applied to the analysis of microarray data even though the expression profiles generated by experiment are fairly complex. The development of new methodologies is needed for the analysis of microarray data.

Artificial neural network (ANN)

Artificial neural network (ANN) is a mathematical cum computational model, which resembles the structure of biological neural networks. It consists of an interconnected group of artificial neurons which processes information using a connectionist approach and it helps to model complex relationships between input and output or to find patterns in the data. A neural network function f*(x)* is defined as a composition of other functions and $g_i(x)$, This can be represented as a network structure, which allows portraying the dependencies between variables. The nonlinear weighted sum, $f(x) = K\Sigma_i w_i g_i(x)$, where the activation function K is some predefined function, called as the hyperbolic tangent.

Multilayer perceptron (MLP)

Multilayer perceptron is a modified form of standard linear perceptron, which can differentiate data that is not linearly separable[4].Multilayer perceptron has a linear activation function in all neurons that can be easily proven with linear algebra, that any number of layers can be reduced to the standard two layer input output model, which make MLP different from other perceptron. The two main activation functions named as hyperbolic tangent and logistic function can be symbolically represented in equation 1 and 2 respectively. Here y_i is the output of the i^{th} node (neuron) and v_i is the weighted sum of the input synapses. The MLP perceptron consists of an input and an output layer with one or more hidden layers.of nonlinearly activating nodes.Each node in one layer connects with a certain weight w_{ij} to every other node of the following layer.

$$\phi(y_i) = \tanh(vi) \quad (1)$$

and

$$\phi(y_i) = (1 + e^{-vi})^{-1} \quad (2)$$

MLP Learning algorithm

The learning algorithm is the generalization of least square algorithm also known as back propagation algorithm. MLP Learning is a supervised learning and it occurs in the perceptron by changing connection weights. The changes of each weight Δw_{ji} (n) can be calculated by gradient descent as in equation (3).

$$\Delta W_{ji}(n) = -\eta \frac{\partial \varepsilon(n)}{\partial v_j(n)} y_i(n) \qquad (3)$$

Where y_i is the output of the previous neuron and η is the learning rate. This derivative can be simplified as

$$-\frac{\partial \varepsilon(n)}{\partial v_j(n)} = e_j(n)\phi'(v_j(n)) \qquad (4)$$

$$-\frac{\partial \varepsilon(n)}{\partial v_j(n)} = e_j(n)\phi^1(v_j(n))\ \Sigma_k - \frac{\partial \varepsilon(n)}{\partial v_j(n)} w_{kj}(n) \qquad (5)$$

where ϕ^1 is the derivative of the activation function and the error can be find out by

$$e_j(n) = d_j(n) - y_j(n) \qquad (6)$$

where y is the value as result of perceptron and *d* is the target value. The correction to the weights based on those corrections which minimize the error in the entire output $\varepsilon(n)$ is given by equation 7 below

$$\varepsilon(n) = \frac{1}{2}\Sigma e_j^2(n) \qquad (7)$$

This error signal is used to update the weights and thresholds of the hidden layers as well as the output layer. The weights and the thresholds are updated in an iterative method until the error signal becomes minimum and the mean square error (MSE) is taken as a performance measurement.

Radial bases function (RBF)

RBF was first proposed by Powell to solve the real multivariate interpolation problem[5]. But RBF was first used by Broomheadand Lowe[6]. RBF networks were independently proposed and applied by number of researchers[7, 8,9,10,11]. Park and Sanndberg proposed the universal approximation theorem, which played an important role in practical application of RBF[8]. In RBF network, the hidden layer has a nonlinear activation function and a linear output layer. Radial basis function (RBF) networks typically have three layers namely an input layer, a hidden layer and an output layer. In RBF networks, the hidden layer of *J* locally tuned units are fully interconnected to an output layer of *L* linear units. All hidden units simultaneously receive the *n*-dimensional real valued input vector *X.* Each hidden unit output $Z_j$is obtained by closeness of the input *X* to an *n*-dimensional parameter vector μ_j associated with j^{th} hidden unit[11,12]. The response characteristics of the j^{th} hidden unit (*j*= 1, 2, … *J*) is assumed as,

$$Z_j = K\left(\frac{\| X - \mu_j \|}{\sigma_j^2}\right) \qquad (8)$$

Where *K* is a strictly positive radially symmetric function (kernel) with a unique maximum at its centre μ_j and which drops off rapidly to zero away from the centre. The parameter σ_j is the width of the receptive field in the input space from unit *j*. Given an input vector *X*, the output of the RBF network is the *L*-dimensional activity vector *Y*, whose l^{th} component (l = 1, 2 …*L*) is given by,

$$Y_l(X) = \sum_{j=1}^{J} w_{ij} Z_j(X) \qquad (9)$$

From equation (8) and (9), we can observe that, the RBF produced network structure, which is approximating a desired function *f* (*X*) by superposition of non-orthogonal, bell-shaped basis functions. The overall accuracy of RBF networks can be controlled by three parameters, the number of basic functions used, their location and their width.[11,13] If we have assumed a Gaussian basis function for the hidden units given as *Zj* for *j* = 1, 2, ..□*J*,

$$Z_j = \exp\left(-\frac{\| X-\mu_j \|^2}{2\sigma_j^2}\right) \qquad (10)$$

and μ_j and σ_j are mean and the standard deviation respectively, of the j^{th} unit receptive field and the norm is Euclidean.

Types of radial basis function

The basis function ϕis a real function of the distance (radius) *r* from the origin. Any function ϕ that satisfies the property ϕ(x) = ϕ(||x||) is a radial function. The commonly used radial basis functions are Guassian, Multiquadric, Inverse quadratic, Inverse multiquadric and poly harmonic splines.

RBF training algorithm

Training of the RBF neural network is a two-step process. In first, the centers of each of the *J* Gaussian basis functions were fixed to represent the density function of the input space. The second is to determine the weight vector *W* which would best approximate the limited sample data *X*, thus leading to a linear optimization problem that could be solved by ordinary least squares method. Least squares function is the common objective function which helps to select the parameter values that minimize its values. The minimization of the least squares objective function by optimal choice of weights optimizes accuracy of fit. If we have multiple objective like smoothness and accuracy a regularized objective function is preferable. The sum of squared error criterion function can be considered as an error function *E* to be minimized over the given training set. That is, to develop a training method that minimizes *E* by adaptively updating the free parameters of the RBF network. Mainly there are three types of parameters in RBF network that need to be chosen to adapt the network for a particular task. They are receptive field center's μ_j of the hidden layer Gaussian units, the receptive field widths σ_j and the output weights w_{ij}. If we use fully supervised gradient-descent method training, then μ_j,σ_j and w_{ij} are updated as follows:

$$\Delta\mu_j = -\rho_\mu \nabla_{\mu j} E \qquad (11)$$

$$\Delta\sigma j = -\rho_\sigma \frac{\partial E}{\partial \sigma j} \qquad (12)$$

$$\Delta w_{ij} = -\rho_w \frac{\partial E}{\partial w i j} \qquad (13)$$

Application to Mycobacterium tuberculosis Data

Database

We have used gene expression data of mycobacterium tuberculosis from GEO database no. GSD1552 with platform ID GPL278. It is a double channel microarray data of M.tuberculosis. Channel 1 contains H37Rv genomic DNA control which is labeled Cy3, and channel 2 contains CDC 1551 genomic DNA which is labeled with Cy5. We have used normalized log ratio values of each gene. We have used both RBF and MLP network for classifying and predicting over expressed genes and we compared the efficiency of prediction performance of networks.

RESULTS

The feed forward neural network using back propagation for error correction is constructed to model the data using SPSS 19.0. Total data is randomly divided into two third for training and remaining for testing**.** For MLP network architecture, hidden layer with sigmoid activation function was chosen. A back propagation algorithm based on conjugate gradient optimization technique was used to model MLP. The training of the MLP was stopped when no further optimization was possible. The MLP comprised of one hidden layer containing neurons for classifying the original data and the MLP network structure is shown in figure1b. For RBF network, we used gradient descent method and local minima characteristic of back propagation algorithm. The RBF neural network architecture produced a single hidden layer with 7 neurons and the network structure is as shown in figure 1a. The training and testing error sum of squares by MLP networks are 12.871 and 6.033 respectively. Similarly the training and testing error sum of squares given by RBF networks are 12.321 and 8.371 respectively. The overall correct prediction performance of training and testing of MLP and RBF network are shown in the following Table.1.

Table 1
Correct classification by neural network

	MLP	RBF
Training	96.6 %	97.1 %
Testing	97.3 %	95.9 %
Area under curve	0.906	0.925

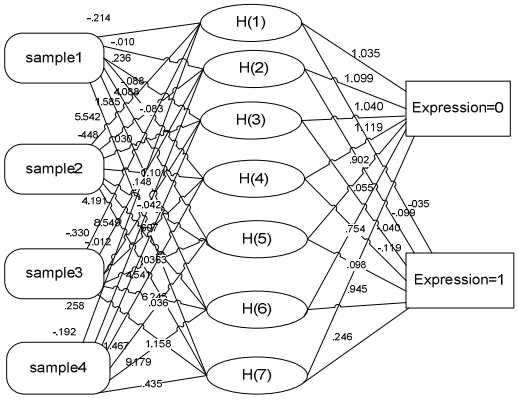


Figure 1a. ***Network Structure of RBF***

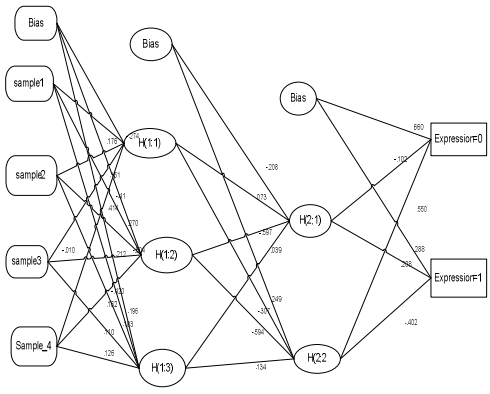


Figure 1b.*Network Structure of MLP*

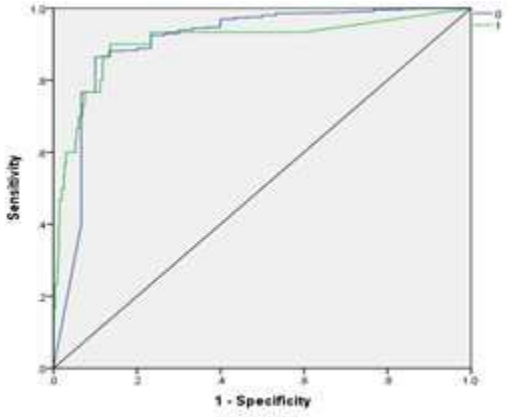
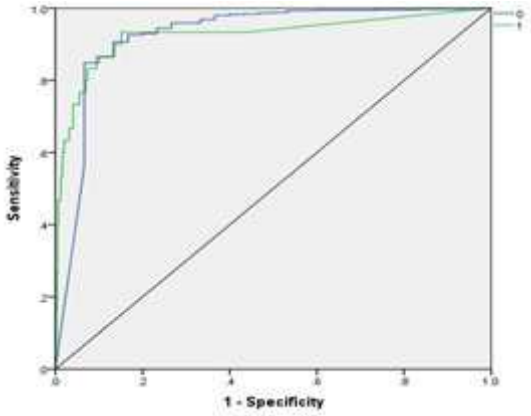


Figure 2a .*ROC curve of RBF Figure*



2b.*ROC curve of MLP*

DISCUSSIONS

Multilayer perceptrons (MLP) and radial basis functions (RBF) networks are the comprehensively used feed forward neural networks. Both differ fundamentally in the way how the hidden units combine values coming from the inputs. The MLP use inner products and the RBF use Euclidian distance. We used both RBF and MLP network algorithms for the classification and prediction of gene expression data of M.tuberculosis. We divided our total data by two third for training and remaining for testing and we noticed that the differentially expressed genes were classified and predicted correctly in training as 97.1% ,95.9% and testing data set using RBF and MLP network as 97.1% , 96.6%and 97.3% respectively. We found that the sensitivity, specificity and area under curve of two networks classifiers are almost same which indicates that both networks have good prediction and classification capabilities for gene expression data. Venkatesan P and Suresh M.L study results show that the accuracy of artificial neural network (ANN) for the breast cancer survival prediction was better than regression based approaches[14]. Burke H.B et al.study concluded that neural networks are more accurate in predicting the breast cancer than LR and CART models for 5th year survival[15.]

HacibT in his paper states that RBF neural network identifies the electromagnetic parameters faster than MLP neural network[16]. Dan Ardelean etal. in their paper reported that in an adequate choices of the training conditions the quality of the RBF model is better than the quality of the MLP[17]. Padmavathi's paper of breast cancer prediction used RBF and MLP, suggested that RBF have good predictive capabilities and time taken was less when compared to MLP[18]. Venkatesan P and Anitha S study indicated that performance of the RBF neural network has a better performance than other models like MLP and classical logistic regression[19]. SerenoF et al. paper entitled **"**Comparative Study of MLP and RBF Neural Nets in the Estimation of the Foetal Weight and Length" concluded with slight confusion regarding prediction performance while comparing the RBF and MLP networks to solve the problem of foetal weight prediction[20]. Many researchers have compared the efficiency of RBF and MLP and majority have recommended that RBF network was better than MLP, and some of them doubt the prediction efficiency. In our study we observed that the prediction and classification performance of RBF and MLP networks were good for gene expression data and both have almost similar performance. Whenever the complex classification problem arises in microarray data analysis, we can use this neural network classifier which will help to minimize crucial classification problems.

CONCLUSION

Based on our study, we concluded that both RBF and MLP neural networks have good classification and prediction capabilities in gene expression data sets. During the study, we have observed that both the neural network classification methods showed almost similar performance in the microarray data sets and the methods also showed good prediction and classification performance. So it is recommended to use both RBF and MLP neuralnetworks methods to solve lassification problems in complex gene expression studies.

REFERENCES

1. Golub, et al.Molecular classification of cancer -class discovery and class prediction by gene expression monitoring .Science,286(5439):531-537(1999).
2. Dudoit, et al.Comparison of discrimination methods for the classification of tumors using gene expression data .Journal of the American Statistical Association,97(457):77-87(2002).
3. Wang et al.Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data ,Bioinformatics,22(6): 755-761(2006).
4. Cybenko G. Approximation by superposition's of a sigmoidal function ,Mathematics of Control Signals and Systems,2(4):303–314(1989)
5. Powell, MJD. Radial basis functions for multivariable interpolation.A review. IMAConference on Algorithms for the Approximationof Functions and Data,1:143-167(1985).
6. Broomhead DS.and Lowe. Multivariable functional interpolation and adaptive networks,Complex Systems,2:321 – 355(1998).
7. Chen S, Cowan CFN and Grant PM. Orthogonal least squares learning algorithm for radial basis function networks,IEE Trans Neural Networks,2:302-309(1991).
8. Park J,Sandberg IW. Universalapproximation using radial-basis-function networks ,Neural computation,3:246-257(1991).
9. Poggio T and Girosi, F.Networks for approximation and learning,Proceedings of the IEE,1481-1497(1990).
10. Liu, G.P., V. Kadirkamanathan and S.A. Billings.Variable neural networks for adaptive control of nonlinear systems,IEE transaction on systems,29:34-43(1999).
11. Ramuhalli P , Udpa L and Udpa SS.Finite element neural networks for solving differential equations, IEEE Trans. Neural Networks,16(6): 1381-1392(2005)
12. Hacib T, Mekideche MR and Ferkha N., Computational Investigation on the Use of FEM and RBF Neural Network in the Inverse Electromagnetic Problem of Parameter Identification,IAENG International Journal of Computer Science,33:2(2009).
13. Haykin S.Neural Networks, A Comprehensive Foundation (2ed.).Prentice Hall (1998).
14. Venkatesan P and Suresh ML.Brest Cancer survival prediction using Artificial Neural Network,IJCSNS International Journal of Computer Science and Network Security ,9(5): 169-174(2009).
15. Burke HB,GoodmanPH and Rosen DB.Comparing the prediction accuracy of artificial neural networks and other statistical model for breast cancer survival ,Advances in Neural information processing system,7:1063-1067(1995).
16. Hacib, T. Mekideche, MR and Ferkha N.Inverse problem methodology for the measurement of the electromagnetic parameters using MLP neural network,Word academy of science engineering and technology,38:608-613(2008).
17. Dan Ardelean,Marius Kloetzer. and Octavian,Pastravanu., RBF neural networks

in nonlinear System identification, Proc 7th International Symposium on Automatic Control and Computer Science,82-86(2001).

18. Padmavathi, J. A comparative study on breast cancer prediction using RBF and MLP ,International Journal of Scientific& Engineering Research,2 (1) :1-5(2011).
19. Venkatesan P and Anitha S.Application of Radial basis function neural network for diagnosis of diabetes mellitus, Current Science,91(9):1195-1199(2006).
20. Sereno F Marques, De SaMatos. Bernardes.A Comparative Study of MLP and RBF Neural Nets in the Estimation of the Foetal Weight and Length Proceedings of RECPAD.11th Portuguese Conference on Pattern Recognition(2000).