Journal of Computer Science & Systems Biology - Open Access

Research Article JCSB/Vol.2 July-August 2009

Optimizing Number of Inputs to Classify Breast Cancer Using Artificial Neural Network

Bindu Garg[1]*, M.M. Sufian Beg[2] and A.Q. Ansari[3]

[1]Department of Computer Science and Information Technology, Institute Of Technology and Management, Sec-23 A, Gurgaon-122017, India, bindusingla@gmail.com

[2]Department of Computer Engineering, Jamia Millia Islamia, Jamia Nagar, New Delhi-110025 India, mmsbeg@hotmail.com

[3]Department of Electrical Engineering, Jamia Millia Islamia, Jamia Nagar, New Delhi-110025, India, aqansari62@gmail.com

*Corresponding author: Bindu Garg, Department of Computer Science and Information Technology, Institute Of Technology and Management, Sec-23 A, Gurgaon-122017, India, E-mail: bindusingla@gmail.com

Received July 07, 2009; Accepted August 25, 2009; Published August 26, 2009

Citation: Garg B, Sufian Beg MM, Ansari AQ (2009) Optimizing Number of Inputs to Classify Breast Cancer Using Artificial Neural Network. J Comput Sci Syst Biol 2: 247-254. doi:10.4172/jcsb.1000037

Abstract

The Objective of this research work is to prove significant role of each attribute to decide breast cancer type using Computer Aided Diagnosis. One of major challenges in medical domain is the extraction of intelligible knowledge from medical diagnostic data in minimum time and cost This research shows that out of these attributes stated, some attributes can be ignored to decide the type Breast Cancer as if the number of inputs are less then it reduces the time and cost in analyzing the breast cancer. In this paper, significant role of each attribute is proved by experiment in matlab.

Keywords: ANN; Back Propagation; Perceptron; Adaptive; Commutative

Introduction

Breast cancer is an uncontrolled growth of breast cells. Cancer occurs as a result of mutations, or abnormal changes, in the genes responsible for regulating the growth of cells and keeping them healthy. The genes are in each cell's nucleus, which acts as the "control room" of each cell. Normally, the cells in our bodies replace themselves through an orderly process of cell growth: healthy new cells take over as old ones die out. But over time, mutations can change cells by "turning on" certain genes and "turning off" others in a cell. Changed cell gains the ability to keep dividing without control, producing more cells just like it and forming a tumor (World Health Organization International Agency for Research on Cancer, June 2003; a b c World Health Organization, February 2006).

A tumor can be benign (not dangerous to health) or malignant (dangerous to health). Benign tumors are not considered cancerous: their cells are close to normal in appearance, they grow slowly, and they do not invade nearby tissues or spread to other parts of the body. Malignant tumors are cancerous. Less commonly, breast cancer can begin in the stromal tissues (G.L Takkab, 2003) which include fatty and fibrous connective tissues of the breast. It is widely believed that the breast cancer is caused by a genetic abnormality (http://www.merck.com/mmhe/sec23/ch266/ch266a.html). However, only 5-10% of cancers are due to an abnormality inherited from mother or father. About 90% of breast cancers are due to genetic abnormalities that happen as a result of the aging process and the "wear and tear" of life in general. One of the outstanding problems in the present day is to diagnose, define and classify the type of breast cancer using minimum cost. Several Artificial Intelligence (AI) techniques including neural networks and fuzzy logic are successfully ap-

plied to a wide variety of decision making problem in the area of medical diagnosis.

This paper is organized as follows: Section 2 describes Artificial Neural Network. Section 3 describes Wisconsin Breast Cancer Database. Section 4 describes related work to this research. Section 5 describes Algorithm part. Section 6 describes Simulation and results. The final section provides some conclusion relating to the performance of the algorithm when applied to the breast cancer.

Artificial Neural Network

During the recent times, because of their discriminative training ability and easy implementation, the Artificial Neural Networks (ANN) find extensive use in classification of the type of tumor in breast cancer problem. It turns out that the selection of number of nodes for an ANN is an important criterion in breast cancer analysis. However, a large network means more computational expenses, resulting in more hardware and time related cost. Therefore a compact and optimum design of neural network is needed towards real time detection of tumor type in breast cancer analysis.

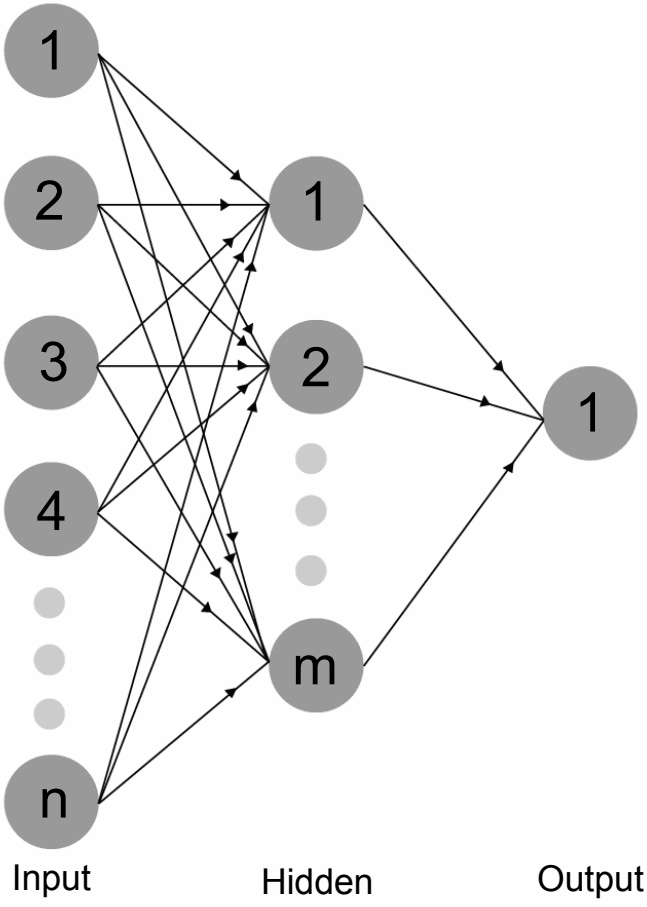

Figure 1: Structure of a Three-Layered Perceptron Type ANN.

An ANN can be defining as a highly connected array of elementary processors called neurons. A widely used model called the multi-layered perceptron (MLP) Figure 1. The MLP type ANN consists of one input layer, one or more hidden layers and one output layer. Each layer employs several neurons and each neuron in a layer is connected to the neurons in the adjacent layer with different weights. Signals flow into the input layer, pass through the hidden layers, and arrive at the output layer. With the exception of the input layer, each neuron receives signals from the neurons of the previous layer linearly weighted by the interconnect values between neurons. The neuron then produces its output signal by passing the summed signal through a sigmoid function.

Training Algorithm

Training may be supervised or unsupervised. Supervised learning refers to the training and synaptic modification of a neural network that is provided with a number of training samples or task examples that construct the training set. Each one training example consists of a unique input signal and the corresponding desired output (response). We can train a neural network to perform a particular function by adjusting the values of the connections (weights) between elements. There, the network is adjusted, based on a comparison of the output and the target, until the network output matches the target. Batch training of a network proceeds by making weight and bias changes based on an entire set (batch) of input vectors. Incremental training changes the weights and biases of a network as needed after presentation of each individual input vector. Incremental training is sometimes referred to as “on line” or “adaptive” training.

The Back-Propagation Learning Algorithm is applied on multilayer feed forward networks, also referred as multilayer perceptrons (MLP). It is based on an error correction learning rule and specifically on the minimization of the mean squared error that is a measure of the difference between the actual and the desired output. Assuming that yj(n) is the actual output of the jth neuron of the output layer at the iteration n and dj(n) is the corresponding desired output, the error signal ej(n) is defined as:

$$e\,j(n) = d\,j(n) - y\,j(n).$$

As all multilayer feed forward networks, the Multilayer preceptrons are constructed of at least three layers (one

input layer, one or more hidden layers and one output layer), each layer consisting of elementary processing units (artificial neurons), which incorporate a nonlinear activation function, commonly the logistic sigmoid function.

There are generally four steps in the training process:

- Assemble the training data
- Create the network object
- Train the network
- Simulate the network response to new inputs

Wisconsin Diagnostic Breast Cancer Data

The Wisconsin Breast Cancer dataset (Merz J, 1996)

Class	Frequency	Percent	Valid Percent	Cumulative Percent
1	357	62.7	62.7	62.7
2	212	37.3	37.3	100
Total	569	100.0	100.0	

Table 1: Statistical Details of Data.

was obtained from a repository of a machine earning Database. The Wisconsin Prognostic Breast Cancer (WDBC) dataset contains 69 instances with 32 attributes each. There are no missing attribute in the dataset. This dataset can be use to predict if breast tumor is benign and malignant.

Several studies have been conducted based on this database. For example, Bennet and Mangasarian (Nett K.P, 1992) used linear programming techniques, obtaining a 99.6% classification rate on 487 cases (the reduced database available at the time) (Ravi Jain, 2003).

Have comparative study of fuzzy classification methods on Breast Cancer Data. There are 9 attributes which are primarily used to classify the breast cancer as either benign or malignant based on cell description gathered by microscopic examination. These are: 1. Clump Thickness 2. Uniformity of Cell Size 3. Uniformity of Cell Shape 4. Marginal Adhesion 5. Single Epithelial Cell Size 6. Bare Nuclei 7. Bland Chromatic 8. Normal Nucleoli and 9. Mitoses.

Breast cancer data has been divided into two categories: one dataset is used for training and other dataset is used for testing. Normalized Data is posted (http://www.scribd.com/doc/14939420/Refrencing-Data.html). Multilayer neural network is trained and tested using 569 Data Samples (http://www.scribd.com/doc/14939420/Refrencing-Data.html). In Training Data (http://www.scribd.com/doc/14939420/Refrencing-Data.html) the first column is ID number, columns 2-10 represent characteristics of cancer cell, and the 11th field denotes the type of cancer (2 denotes malignant and 4 denotes benign).

The type of Breast cancer is decided by all 9 attributes (http://www.scribd.com/doc/14939420/Refrencing-Data.html).

Related Work

The artificial neural network has become a very popular alternative in prediction and classification tasks due to its associated memory characteristics and generalization capability (Shien-Ming Chon, 2004). Thirteen cytology of fine needle aspiration image (i.e. cellularity, background information, cohesiveness, significant stromal component, clump thickness, nuclear membrane, bare nuclei, normal nuclei, mitosis, nucleus stain, uniformity of cell, fragility and number of cells in cluster) are evaluated their possibility to be used as input data for artificial neural network in order to classify the breast precancerous cases into four stages, namely malignant, fibro adenoma, fibrocystic disease, and other benign diseases. An intelligent diagnostic system based on the hybrid multilayer perceptron (HMLP) network to determine the four stages of breast pre-cancerous, namely malignant, fibroadenoma, fibrocystic disease, and other benign diseases was proposed (Nor Ashidi MatIsa, 2007). An automated computerized classification method based on computer vision and artificial neural networks to estimate the likelihood that a mammographic lesion is malignant. The features selected to characterize the mass lesions as well as the classifier used to merge the features are crucial components of the overall classification method. We have shown that our computerized method is robust to case mix and digitization technique (Maryellen L. Giger, 2000). The three neural networks were trained and compared using 1300 data samples. The classification results are indicating that all the networks give good overall diagnostic performance. However, only Hybrid Multilayered Network that provides 100% accuracy, sensitivity and specificity (Mohd Yusoff Mashor, 1996). We have developed a computer vision system that can classify microcalcifications (Yuzheng C. Wu, 1995) objectively and consistently to aid radiologists in the diagnosis of breast cancer. A convolution neural network (CNN) was employed to classify benign and malignant microcalcifications in the radiographs of pathological specimen.

Algorithm

In this research MATLAB is used for implementation. The training data is stored in the file Train.txt which con-

tains training patterns for patients. Firstly, feed forward neural network is trained with 2 to 10 columns as inputs and 11th column as output. Then the network is simulated with input and output. After that testing of network is done by taking Data from test.txt file (contains the testing data). The algorithm is as follows:

```
x=load('train.txt');
p=x(:,2:10);
p=p';
t=x(:,11);
t=t';
network=newff(minmax(p),[9,6,1],
{'logsig','logsig','logsig'},'traingdm');
net.trainParam.lr=0.5; net.trainParam.mc=0.9;
network=init(network);
y=sim(network,p);
figure,plot(p,t,p,y,'o'),title('Before Training');
 network.trainParam.epochs = 100;
network=train(network,p,t);
 y=sim(network,p);
figure,plot(p,t,p,y,'o'),title('after Training');
display('Final weight vector and bias values : \n');
Weights=network.iw{1}; Bias=network.b{1};
Weights
Bias
y=load('test.txt'); p1=y(:,2:10);
p1=p1'; t1=y(:,11); t1=t1';y1=sim(network,p1);
figure,plot(p1,t1,p1,y1,'o'),title('after testing');
```

Description of the Algorithm

Training Data is stored in file Train.Txt. Train.Txt file contain 11 Column. First column ID no is omitted. It is not containing any valuable information. So input for the neural network is 2 to 10 column. Actual output is in 11th column. Feed-forward architecture of neural network is used. There are really two decisions that must be made regarding the hidden layers: how many hidden layers to actually have in the neural network and how many neurons will be in each of these layers.

There are many rule-of-thumb methods for determining the correct number of neurons to use in the hidden layers, such as the following:

- The number of hidden neurons should be between the size of the input layer and the size of the output layer
- The number of hidden neurons should be 2/3 the size of the input layer, plus the size of the output layer.
- The number of hidden neurons should be less than twice the size of the input layer.

So number of hidden neurons are 6

Usually it is recommended to start with only one hidden layer, and if the results are not good, the number of hidden layers will grow up (Hornik K, 1990). The number of neurons of each hidden layers are being calculated automatically (Ward Systems Group Inc., 1999).

Number of iterations has been taken 100. In standard backprop, too low a learning rate makes the network learn very slowly. Too high earning rate makes the weights and objective function diverge, so there is no learning at all. The larger the learning rate the larger the weight changes on each epoch, and the quicker the network learns. However, the size of the learning rate can also influence whether the network achieves a stable solution. If the learning rate gets too large, then the weight changes no longer approximate a gradient descent procedure. (True gradient descent requires infinitesimal steps). By default learning rate is .25 and momentum constant is 0.9 when default learning rate was taken then algorithm was not showing correct Results. Value of Learning Rate is taken by heat and trial in matlab.

The objective of this algorithm is to optimize the number of inputs. Now each attribute is removed one by one and it is checked that which attribute is important to decide tumor type.

Simulation and Results

If the algorithm is executed for all attributes (inputs) in MATLAB, Training is given for all 9 parameters; it gives accurate results as shown in Figure 2 and Figure 3. It means that type of breast cancer is decided accurately. Bottom X axis represents Actual Output. Top X axis Represents Target Output. Left and Right Y Axis Represents Input.

If a particular attribute is removed then dotes (colored) coincides on the point 1 at the top horizontal than output is accurate and this means that particular attribute can be discarded. Sample Data in Table 2 shows the mathematical deviation in accuracy of output after removing an attribute. For complete data refer to reference (http://www.scribd.com/doc/14939420/Refrencing-Data.html).

Removal of Clump Thickness (Without Clump thickness)

If the first attributes clump thickness is removed, there are only 8 inputs to the network. Therefore feed forward neural network is trained with 3 to 10 columns as inputs and 11th column as output. To check whether this attribute plays important role in deciding type of tumor, network is first trained and tested with the sample training data. Subsequently network is tested with the testing data as shown in Figure 4 and Figure 5.

It is proved that even if clump thickness is not considered then also accurate results are obtained. As from Figure 4 dotes are coinciding with the top horizontal. It means without this attribute, type of breast cancer can be de-

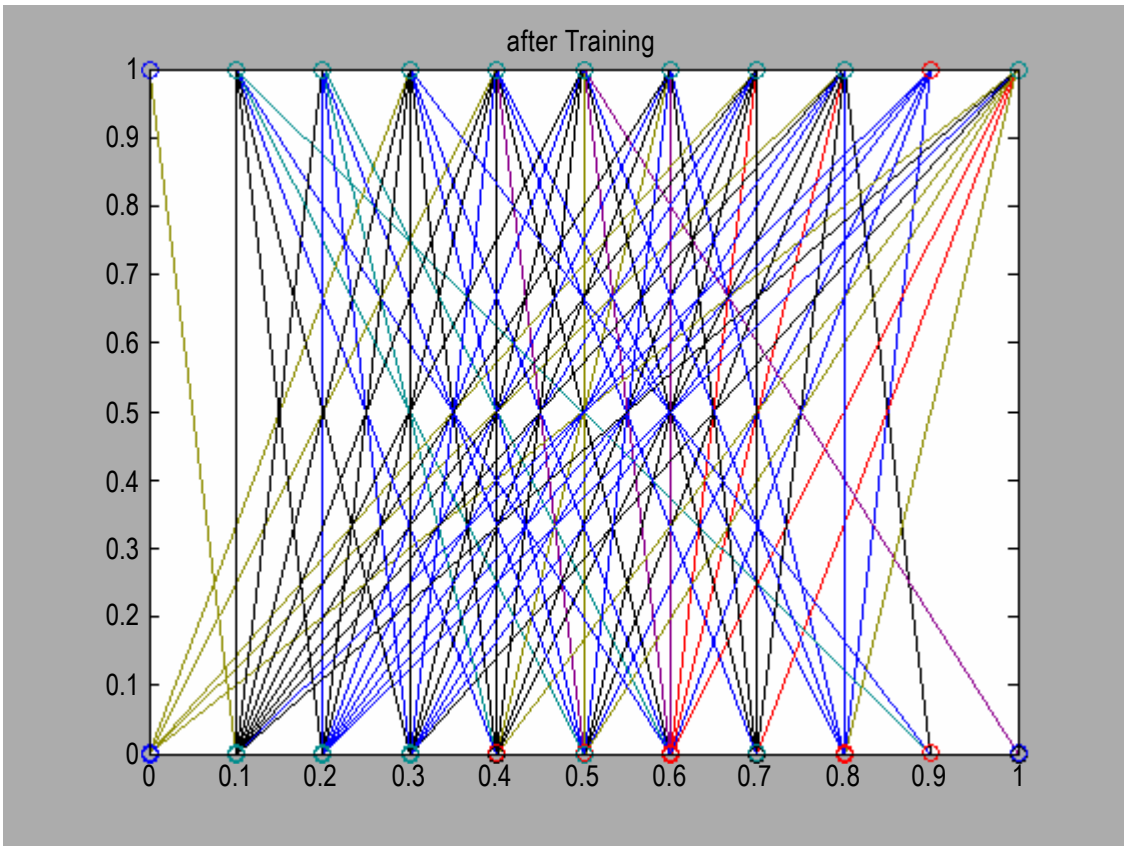

Figure 2: All Parameters (After Training).

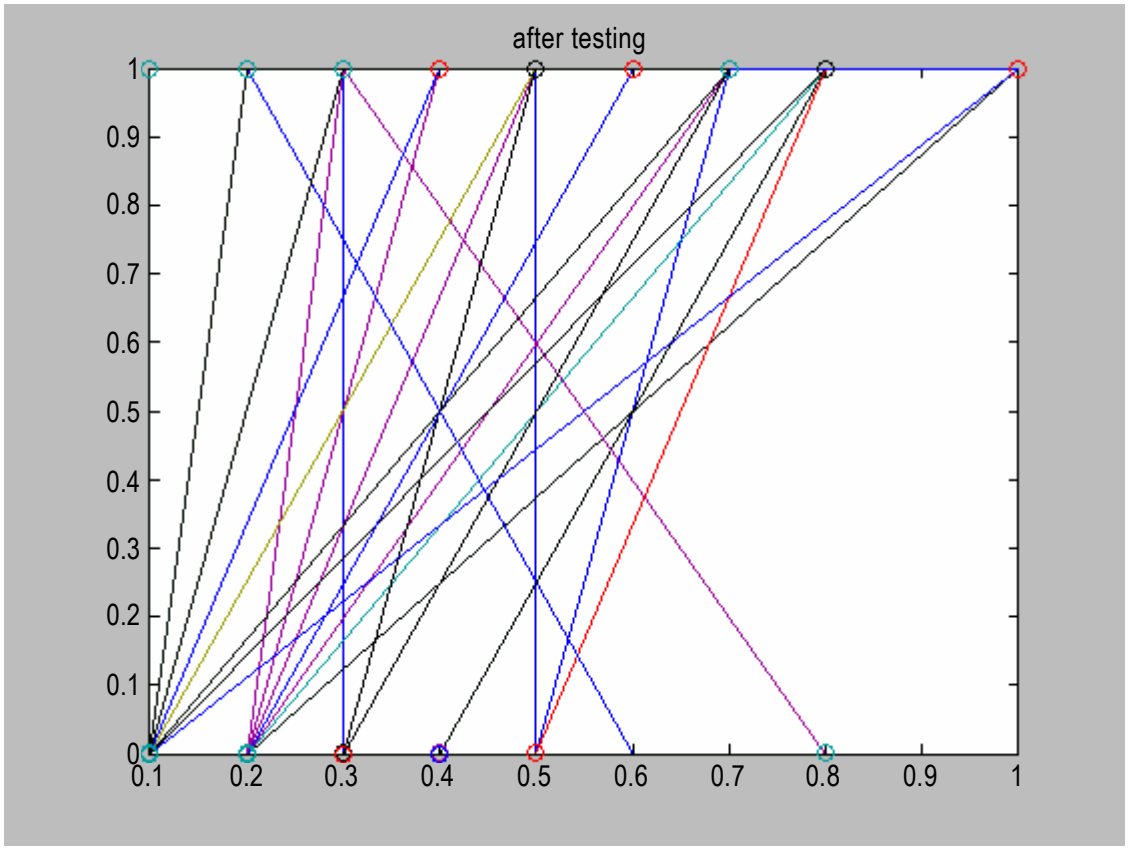

Figure 3: All Parameters (After Testing).

Desired Output	Actual Output	Error	Clump Thickness	Error	Uniformity of Cell Size	Error	Uniformity of Cell Shape	Error	Marginal Adhesion	Error	Single Epithelial Cell Size	Error	Bare Nuceli	Error
0	0.0001	0.0001	0	0	0	0	0	0	0.0024	0.0024	0.0072	0.0072	0	0
0	0.0001	0.0001	0	0	0	0	0	0	0.0024	0.0024	0.0072	0.0072	0	0
0	0.0001	0.0001	0	0	0	0	0	0	0.0024	0.0024	0.0072	0.0072	0	0
0	0.0001	0.0001	0	0	0	0	0	0	0.0024	0.0024	0.0072	0.0072	0	0
0	0.0001	0.0001	0	0	0	0	0	0	0.0024	0.0024	0.0072	0.0072	0	0
0	0.0001	0.0001	0	0	0	0	0	0	0.0024	0.0024	0.0072	0.0072	0	0

Table 2: Sample Data

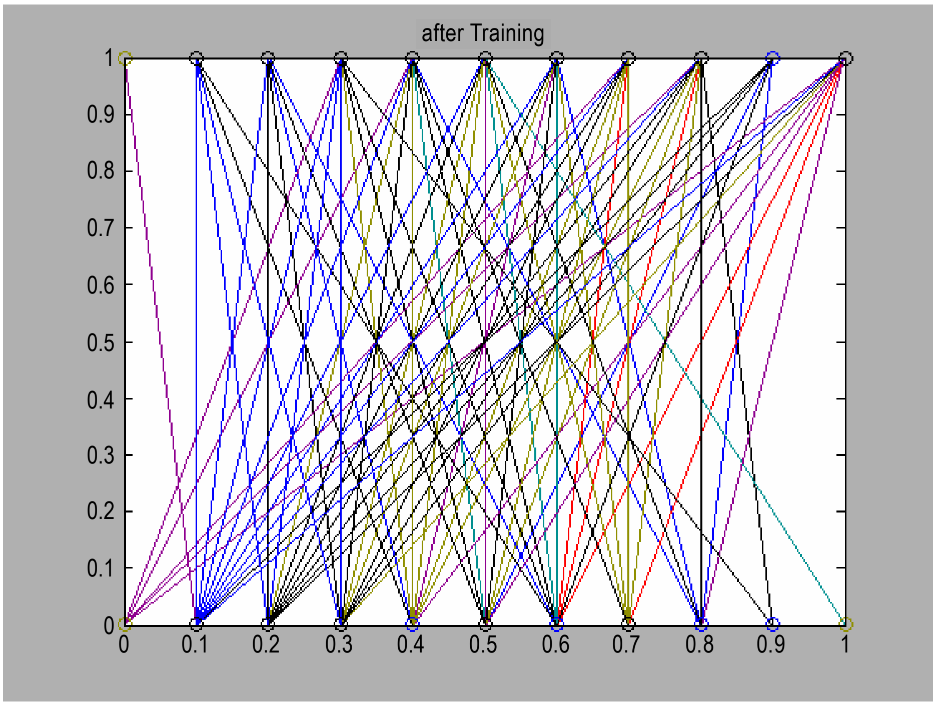

Figure 4: After Training (Clump Thickness).

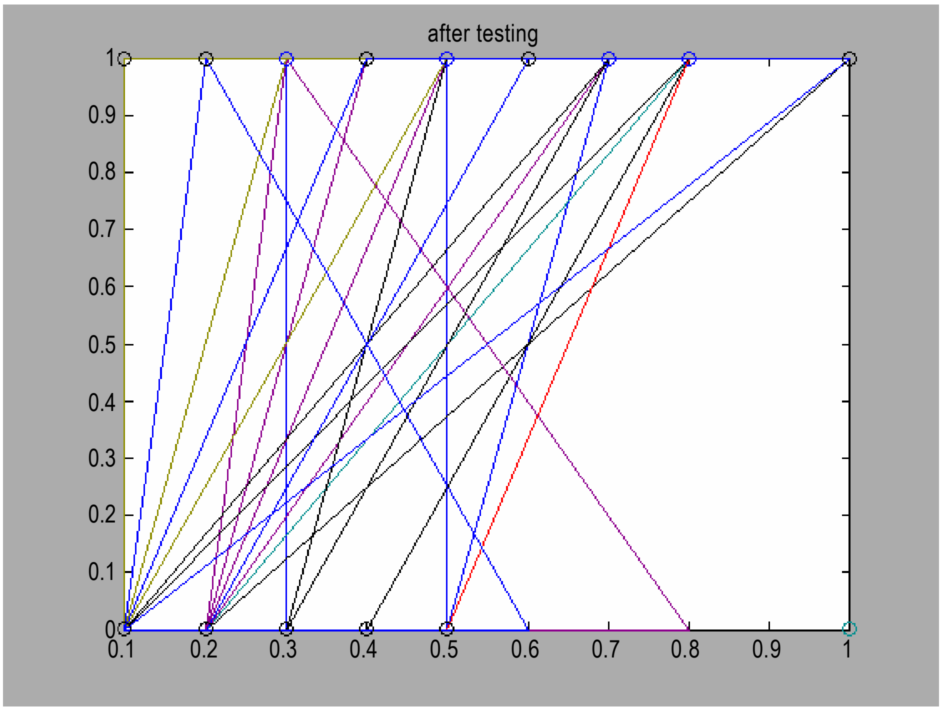

Figure 5: After Testing (Clump Thickness).

cided. Refer Result Data (http://www.scribd.com/doc/14939420/Refrencing-Data.html) for deviation in accuracy of output after removing Clump thickness attribute.

Removal of Uniformity of Cell Size (without uniformity of cell size)

If the second attribute, Uniformity of cell size, is removed, there are only 8 inputs to the network. Here feed forward neural network is trained with 2 and 4 to 10 columns as inputs and 11th column as output. To check whether this attribute plays important role in deciding type of tumor, network is first trained and tested with the sample training data as shown in the figures below.

Figure 6 and figure 7 shows that type of breast cancer cannot be decided without this attribute as the top horizontal is not coinciding. So this attribute cannot be discarded. Refer Result Data (http://www.scribd.com/doc/14939420/Refrencing-Data.html) for deviation in accuracy of output after removing Uniformity of cell size attribute.

Now applying the algorithm on the remaining attributes it is found that the top horizontal coincides on colored dotes on point 1 in case of Uniformity of Cell Shape, Marginal adhesion, Bare Nuclei, Bland Chromatic and Mitoses. Therefore theses attributes can be discarded one at a time only. And it does not coincide on Single Epithelial Cell Size, Normal Nucleoli so these can not be discarded.

Conclusion

In this paper, a case study of attributes is done for real time detection of tumor type in breast cancer analysis. Our study shows that uniformity of cell size, single epithelial cell size and Normal Nucleoli attributes are very important to decide tumor type. But clump thickness, uniformity of cell shape, bare nucleoli, Marginal Adhesion, bland chromatic and mitoses can be ignored. Effectively reducing the time and cost in analyzing tumor type. It can be seen that in previous researches, only the classification of cancer is being studied. In this research, the significant role of each attribute is discussed and studied using matlab. This can help in Hospitals.

The advantage of the proposed algorithm is that it is very simple to implement, neither the trial-and-error process nor prior knowledge about the parameters is required.

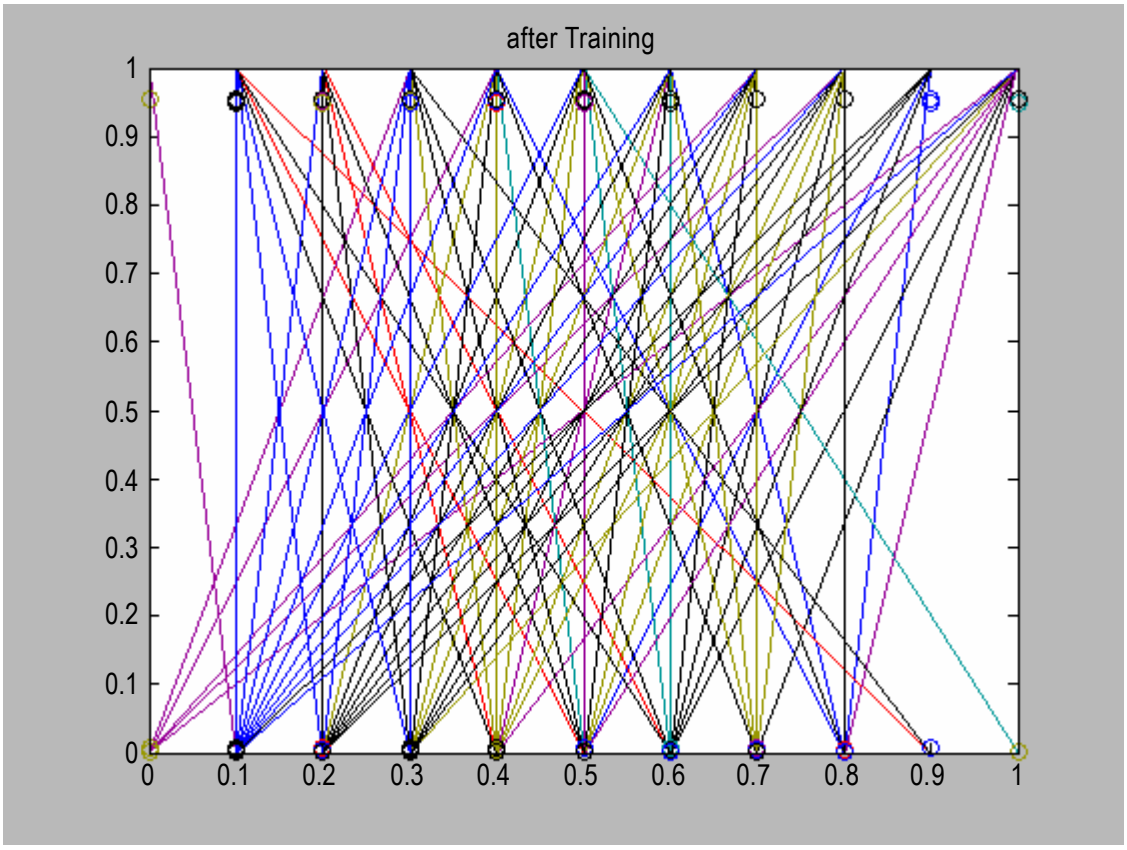

Figure 6: After Training (Uniformity of Cell Size).

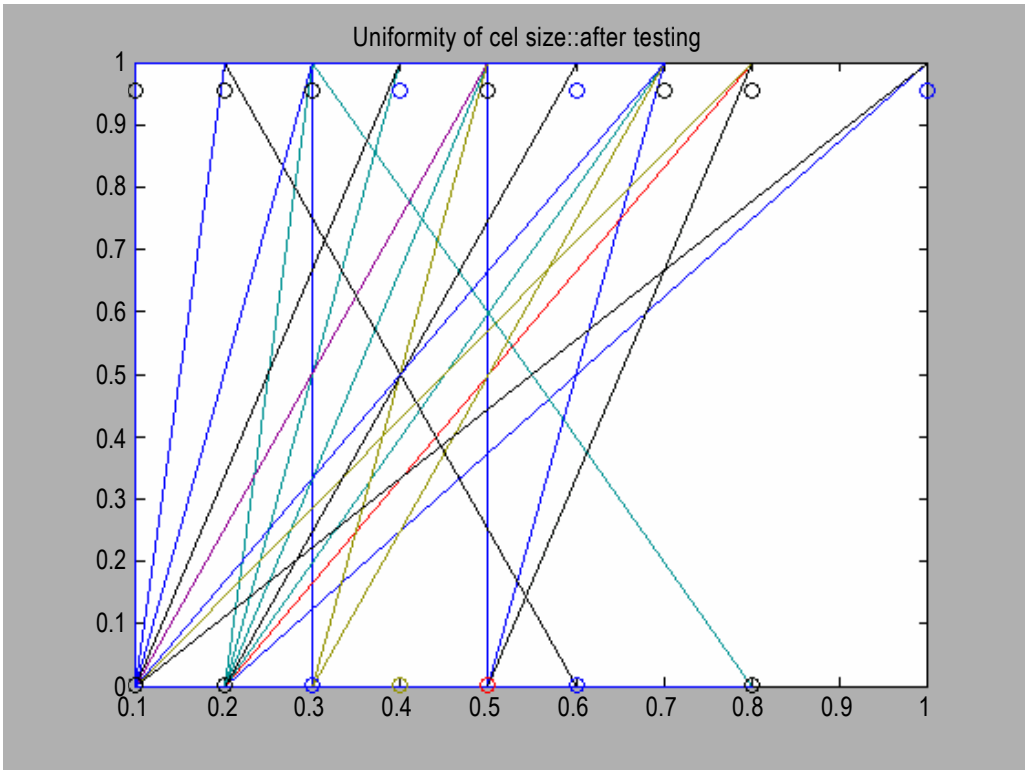

Figure 7: After Testing (Uniformity of cell Size).

References

1. a b c World Health Organization (February 2006) Fact sheet No. 297: Cancer. http://www.who.int/mediacentre/factsheets/fs297/en/index.html.

2. Takkab GL, Kamboj VP (2003) Effect of Altered Hormonal States on Histo-Chemical Distribution of Polysaccharides, Lucknow: Central Drug Research Institute. » CrossRef » Pubmed » Google Scholar

3. http://www.merck.com/mmhe/sec23/ch266/ch266a.html.

4. http://www.scribd.com/doc/14939420/Refrencing-Data.html.

5. Hornik K, Stinchombe M, White H (1990) Neural Networks 2: 359.

6. Giger ML, Huo Z (2000) Artificial Neural Networks in Breast Cancer Diagnosis, Merging of Computer-Extracted Features From Breast Images, IEEE Xplore.

7. Merz J, Murphy PM (1996) UCI repository of machine learning databases. http://www.ics.uci.edu/-learn/MLRepository.html.

8. Mashor MY, Esugasini S, Mat-Isa NA, Othman NH (1996) Classification of Breast Lesions Using Artificial Neural Network. Springer Berlin Heidelberg 15.

9. Nett KP, Mangasarian OL (1992) Neural Network Training via Linear Programming, Advances in Optimization and Parallel Computing. Elsevier Science pp56-67.

10. Mat-Isa NA, Subramaniam E, Mashor MY, Othman NH (2007) Fine Needle Aspiration Cytology Evaluation for Classifying Breast Cancer Using Artificial Neural. American Journal of Applied Sciences 4: 999-1008. » CrossRef » Google Scholar

11. Jain R, Abraham A (2003) A Comparative Study of Fuzzy Classification Methods on Breast Cancer Data. International Conference on Artificial Neural Network IWANN'03, Spain. » CrossRef » Pubmed » Google Scholar

12. Shien MC, Lee TS, Shao YF, Chen IF (2004), Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression spines. Expert System With Application 27: pp133-142.

13. Ward Systems Group Inc. NeuroShel 2 (1999) Ed. Ward Systems Group.

14. World Health Organization International Agency for Research on Cancer (2003) World Cancer Report. http://www.iarc.fr/IARCPress/pdfs/wcr/index.php.

15. Wu YC (1995) Classification of Microcalcification of the Diagnosis of Breast Cancer using Artificial Neural Networks, Page 26, Annual Report 1 Sep 94.