Classification Of Breast Cancer Using Neural Networks

S. Radharani[1], P.Keerthana[2], P.T.Vanathi[3], K.Gunavathi[4]

ABSTRACT

Early detection is an important and promising medical activity to improve the chances of survival of the patients. Classification of Breast Cancer using various Neural Networks is performed and their results are compared in order to identify the network suitable for distinguishing between a benign and a malignant one. A Probabilistic Neural Network (PNN) for breast cancer classification producing accuracies up to 98% and a Back Propagation Neural Network (BPNN) having two output neurons are proposed and their accuracies are being compared with existing BPNN having one output neuron, the Radial Basis Function (RBF) Networks, the Learning Vector Quantization (LVQ) networks and also with Adaptive Neuro Fuzzy Inference System (ANFIS). Wisconsin Breast Cancer Diagnosis (WBCD) dataset is used for training and testing of the proposed neural networks.

Index Terms—BPNN, PNN, Breast Cancer Classification.

I. INTRODUCTION

REAST CANCER is the fifth most common cause of cancer death and is one of the major problems faced during the medical diagnosis. In 2009, approximately 40,000 women are expected to die from breast cancer, while roughly 192,000 women are expected to be diagnosed with the disease. Mammography being one of the widely used techniques causes possible threat in increasing cancer risk due to exposure to low-dose ionizing radiation.

Early diagnosis of this disease is an important and promising medical activity to improve the chances of survival of the patients. But the problem in medical science is that the diagnosis of disease is based upon various tests performed upon the patient which in turn leads to several test results. When several tests are involved, the ultimate diagnosis may be difficult to obtain, even for a medical expert. This has given rise to computerized diagnostic tools intended to aid the physician in making sense out of the confusion of data.

Breast cancer diagnosis has been a typical machine learning benchmark problem for many years and has been dealt using various machine learning algorithms. Artificial Neural Networks are computer algorithms that are typically employed to classify a set of patterns into one of several classes. The classification rules are not written into the algorithm, but are learned by the network from examples [4, 6-8]. Different Neural Network algorithms were proposed as a solution, a few decades before.

In the existing approach, the LVQ and the ANFIS used for testing WBCD dataset yield an accuracy lesser than 98%. The BPNN with only one output neuron produces better accuracies only for some particular combinations of BPNN architectures [1]. In this paper, our work is to propose neural networks suitable for classification of breast cancer producing accuracies better than the previous works. PNN and BPNN with two output neurons proposed in our work produces reasonable and better accuracies when compared to the existing approaches.

2. TECHNIQUES

A. Basic Architecture

The basic architecture for classification of breast cancer is shown in figure 1.

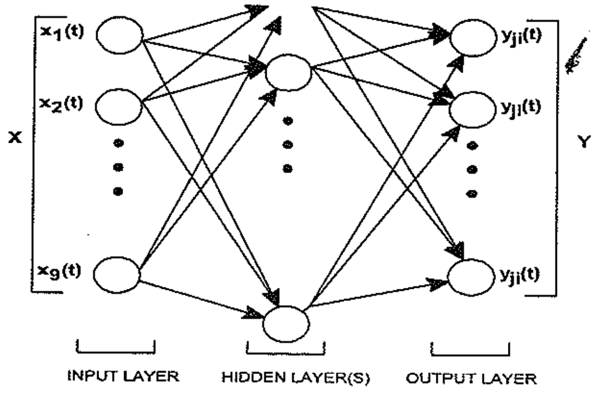

Figure 1 : Basic Architecture for Identification of Breast Cancer

In this architecture there are three layers namely input layer, hidden layer(s) and output layer. The PNN used for classification contains only one hidden layer. The BPNN consists of two hidden layers with four different hidden neurons in hidden layer 1 and 2[1].

B. Probabilistic Neural Networks

A Probabilistic Neural Network (PNN) is predominantly a classifier and can map any input pattern to a number of classifications. Probabilistic Neural Network is trained and tested for the WBCD dataset with different spread values. The dataset is trained by having benign as class 1 and malignant as class 2. The testing results obtained after applying the testing dataset to the PNN should also have benign as class 1 and malignant as class 2.

C. Back Propagation Neural Networks

In the Back Propagation algorithm with one output neuron, the dataset is being trained by having output neurons with values 2 for benign and 4 for malignant. The testing results are thus obtained finally by setting a threshold value of 3.

If $Y_i(t) < 3$, *it is benign*
$Y_i(t) > 3$, *it is malignant*

where, $i = 1 to 200.$ (1)

The proposed Back Propagation algorithm with two output neurons have one of its output neuron trained with value 2 for benign and 0 for malignant and after testing, the first output neuron for benign is obtained by having value 1 as threshold.

If $Y_{1i}(t) > 1$, *it is benign*

where, $i = 1 to 200.$ (2)

The second output neuron is trained with 0 for benign and 4 for malignant. The second testing output is classified by setting threshold value of 2 and second output neuron is obtained for malignant.

If $Y_{2i}(t) > 2$, *it is malignant*

where, $i = 1 to 200.$ (3)

In our proposed network, the two output neurons either have 0 or a value 2or 4 for benign and malignant respectively. In the existing network, the output neuron has 2 for benign and 4 for malignant where the range of values between 2 and 4 is less. Whereas in our proposed architecture, the range of values between 0 and 2, and the range of values present between 4 and 0 are more resulting in better accuracy for classification.

3. Dataset

A. Collection of Dataset

The source of data is Wisconsin Breast Cancer Diagnosis (WBCD). The data set consists of 699 instances, of 699 clinical cases. There are 9 attributes per instance plus the class attribute and the sample code number. Each instance contains 11 attributes of which we use only 9 attributes as input to training and testing. Each of the instances has to be categorized into either of the two categories: Benign or malignant. This categorization is done by the value in the class attribute, 2 for benign and 4 for malignant.

B. Removal of instances with missing attributes

Out of the 699 instances, value of one attribute is missing in 16 instances. Therefore 16 instances have been left out while using this data set. So, 683 instances have been used out of which 483 have been used for training the networks and 200 instances have been used for testing purposes. Of the 699 instances, we have 458 benign and 241 malignant instances and in these there are 14 and 2 missing instances respectively resulting in 444 benign and 239 malignant instances which yield 683 instances when summed up.

C. Data Splitting into Training and Testing dataset

After the removal of instances with missing attributes the 444 benign instances can be segregated as 314 training and 130 testing instances i.e., 70.72% of 444 yields 314 training and 29.28% of 444 yields 130 testing instances.

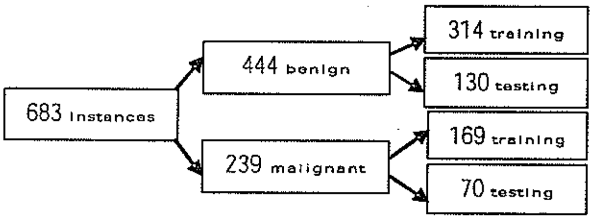

Figure 2 : Data Splitting

The 239 malignant instances can be segregated as 169 training and 70 testing instances i.e., 70.72% of 239 yields 169 training and 29.28% of 239 yields 70 testing instances. The data splitting is represented in figure 2.

4. Parameters

A. Parameter for PNN

The parameter, spread value is being varied in this Probabilistic Neural Network. The range of variation is from 0.4 to 2.0. The variation is made in steps of 0.01 and their accuracies are noted.

B. Parameters for BPNN

The BPNN is performed mainly with 4 combinations of number of hidden layers as 15(10+5), 20(10+10), 25(17+8) and 30(18+12). Then for each architecture the learning rate is varied from 0.05 to 0.07 and for each learning rate the momentum factor is varied from 0.2 to 0.8.

V. Simulation Results

A. Simulation Results for PNN and its Comparison with LVQ and ANFIS

The parameter spread value is varied from 0.1 to 10 in incremental steps of 0.1. We say this as first depth process of PNN. Then the accuracies obtained for first depth is given in table I.

Then parameter spread value is varied from 0.4 to 2 in different incremental steps to calculate better accuracy for classification. We say this as second depth process of PNN. Then the accuracies obtained for second depth is given in table II.

Table 1 : Results for PNN (First Depth)

Spread Value	Accuracy in percentage
0.1	72.5
0.2	95.5
0.3	97.5
0.4	97.5
0.5-1.8	98.0
2	97.5
3	97.5
4	97.0
5	95.5
6	95.5
7	95.0
8	93.5
9	91.0
10	88.5

From table I, it is noticed that the PNN gives 98% accuracy for spread values ranging from 0.5 to 1.8.

Table 2 : Results for PNN (Second Depth)

Spread Value	Accuracy in percentage
0.41-0.48	97.5
0.494	97.5
0.495	98.0
0.5-1.86	98.0
1.864	98.0
1.865	97.5
1.87-2	97.5

From table II, the value at which the accuracy changed from 97.5% to 98% is significant.

Table 3 : Comparison of Results for PNN with LVQ and ANFIS

Network	Accuracy in percentage
PNN	98.0
LVQ [1]	97.0
ANFIS [1]	89.5

The accuracy obtained for breast cancer identification using our proposed PNN is better than the accuracy of existing LVQ and ANFIS networks and is represented in the table III.

B. Simulation Results for BPNN having two output neuron and its Comparison with BPNN having one output neuron

The simulation results of BPNN with one output neuron are compared with results of BPNN with two output neurons. The algorithms used for simulation are traingdm, trainbfg and traincgp. The simulation is performed by applying the testing dataset to the BPNN, whose weight matrices are averaged weight values obtained after 100 iterations of training. The simulation results for BPNN using traingdm algorithm consists of four different combinations of hidden neurons. Each combination is varied for various learning rate and momentum factor. The simulation results for 15 hidden neurons with traingdm are presented in table IV.

Table 4 : Results for BPNN with Traingdm for 15 Hidden Neurons

Learning Rate	Momentum Factor	Accuracy in percentage for BPNN with 1 output neuron	Accuracy in percentage for BPNN with 2 output neurons
0.05	0.7	99.0	99.5
0.05	0.8	94.0	98.0
0.06	0.2	98.5	98.0
0.06	0.3	96.0	94.5
0.06	0.4	97.0	96.5
0.06	0.5	98.5	97.0
0.06	0.6	97.0	98.0
0.06	0.7	89.0	98.0
0.06	0.8	99.0	99.0
0.07	0.2	99.0	97.5
0.07	0.3	95.0	98.0
0.07	0.4	99.0	99.0
0.07	0.5	100	98.5
0.07	0.6	96.5	99.0
0.07	0.7	100	91.5
0.07	0.8	97.5	99.0

From table IV, we see that the average percentage of accuracy obtained from the 16 different combinations of learning rate and momentum factor for BPNN with one output neuron is 97.2% whereas we get 97.6% for BPNN with two output neurons.

Table 5 : Results for BPNN with Traingdm for 20 Hidden Neurons

Learning Rate	Momentum Factor	Accuracy in percentage for BPNN with 1 output neuron	Accuracy in percentage for BPNN with 2 output neurons
0.05	0.7	65.0	96.0
0.05	0.8	65.0	95.5
0.06	0.2	82.0	99.5
0.06	0.3	99.5	99.0
0.06	0.4	99.5	97.0
0.06	0.5	98.5	96.5
0.06	0.6	75.0	100
0.06	0.7	97.0	99.0
0.06	0.8	99.5	98.5
0.07	0.2	98.5	97.0
0.07	0.3	94.5	99.0
0.07	0.4	65.0	99.0
0.07	0.5	99.5	95.5
0.07	0.6	65.0	99.0
0.07	0.7	98.5	96.0
0.07	0.8	100	98.0

The simulation results for 20 hidden neurons with traingdm are presented in table V. From table V, we see that the average percentage of accuracy obtained for BPNN with one output neuron is 87.6% whereas we get 97.8% for BPNN with two output neurons.

From table V, it is clear that the accuracy obtained for classification of breast cancer using BPNN with two output neurons is better when compared to the results obtained from BPNN with one output neuron.

The simulation results for 25 and 30 hidden neurons with traingdm are presented in tables VI and VII respectively.

Table 6 : Results for BPNN with Traingdm for 25 Hidden Neurons

Learning Rate	Momentum Factor	Accuracy in percentage for BPNN with 1 output neuron	Accuracy in percentage for BPNN with 2 output neurons
0.05	0.7	98.5	97.5
0.05	0.8	98.0	89.5
0.06	0.2	96.0	95.5
0.06	0.3	99.0	98.0
0.06	0.4	97.0	98.5
0.06	0.5	96.0	98.5
0.06	0.6	97.0	97.0
0.06	0.7	96.5	99.0
0.06	0.8	97.5	98.5
0.07	0.2	91.5	96.5
0.07	0.3	98.0	99.0
0.07	0.4	95.5	96.5
0.07	0.5	95.0	95.0
0.07	0.6	98.0	98.0
0.07	0.7	98.5	93.5
0.07	0.8	100	96.5

From table VI, the average percentage of accuracy obtained for both BPNN with one output neuron and two output neurons is 97%.

Table 7 : Results for BPNN with Traingdm for 30 Hidden Neurons

Learning Rate	Momentum Factor	Accuracy in percentage for BPNN with 1 output neuron	Accuracy in percentage for BPNN with 2 output neuron
0.05	0.7	99.0	95.0
0.05	0.8	65.0	95.5
0.06	0.2	97.5	94.0
0.06	0.3	95.5	98.5
0.06	0.4	98.0	98.0
0.06	0.5	93.5	98.5
0.06	0.6	98.5	97.5
0.06	0.7	65.0	94.5
0.06	0.8	40.0	95.5
0.07	0.2	98.5	95.0
0.07	0.3	98.5	97.5
0.07	0.4	98.5	96.5
0.07	0.5	92.5	96.5
0.07	0.6	96.5	94.5
0.07	0.7	99.0	97.0
0.07	0.8	65.0	98.0

From table VII, the average percentage of accuracy obtained for BPNN with one output neuron is 87.5% and for two output neurons is 96%.

The simulation results for four combinations of hidden neurons using trainbfg and traincgp algorithms are presented in tables VIII and IX respectively.

From table VIII, it is observed that the accuracies for the classification of breast cancer using BPNN with two output neurons are either increased or maintained when compared to the BPNN with one output neuron due to the flexibility in range of output neuron values.

The average accuracy obtained for BPNN with two output neurons using trainbfg algorithm is 98% which is better than 89.2% accuracy obtained for BPNN with one output neuron.

Table 8 : Results for BPNN with Train BFG

Number of Hidden Neurons	Accuracy in percentage for BPNN with 1 output neuron	Accuracy in percentage for BPNN with 2 output neurons
15	98.5	99.0
20	65.0	97.5
25	97.0	97.5
30	96.5	98.0

Table 9 : Results for BPNN with Train CGP

Number of Hidden Neurons	Accuracy in percentage for BPNN with 1 output neuron	Accuracy in percentage for BPNN with 2 output neurons
15	97.5	97.5
20	99.0	98.0
25	96.5	98.5
30	95.5	98.0

Using traincgp algorithm the average accuracy obtained for BPNN with two output neurons is also 98% which is better than 97.1% accuracy obtained from BPNN with one output neuron. This is shown in table IX.

6. Conclusion

The neural network based classification of breast cancer is more effective. On an average, the accuracy for classification of breast cancer using BPNN with one output neuron is 92.5%. For the same classification PNN gives 95.8% and BPNN with two output neurons gives 97.1%. Thus we conclude that, BPNN with two output neurons using traingdm algorithm outperforms BPNN with one output neuron, PNN, LVQ and ANFIS. In particular, BPNN with two output neurons having 20 hidden neurons is best suitable for classification of breast cancer.

REFERENCES

[1] *Anupam Shukla, Ritu Tiwari, Prabhdeep Kaur G. O. Young, "Knowledge Based Approach for Diagnosis of Breast Cancer"* IEEE International Advance Computing conference (IACC), 2009, pp. 6-12.

[2] Richard Dybowski, Vanya Gant: "Clinical applications of artificial neural networks", Cambridge University Press, 2001, pp. 1-20.

[3] Drasko Furundzic, Miodrag Djordjevic, Ana Jovicevic Bekic: "Neural networks approach to early breast cancer detection", Journal of Systems Architecture 44, 1998, pp. 617-633.

[4] Rudy Setiono: "Extracting rules from pruned neural networks for breast cancer diagnosis",Artificial Intelligence in Medicine 8, 1996, pp. 37-51.

[5] Rudy Setiono: "Generating concise and accurate classification rules for breast cancer diagnosis", Artificial Intelligence in Medicine 18, 2000, pp. 205-219.

[6] Rudy Setiono, Huan Liu: "NeuroLinear: From neural networks to oblique decision rules", Neurocomputing 17, 1997, pp. 1-24.

[7] Xin Yao, Yong Liu: "Neural Networks for Breast Cancer Diagnosis", IEEE, 1999, pp. 1760-1767.

[8] Carey E. Floyd, Jr., Joseph Y.Lo, Joon Yun, Daniel C. Sullivan and Phyllis J. Kornguth: "Prediction of Breast Cancer Malignancy Using an Artificial Neural Network", CANCER, 1994, vol.74, No.11, pp. 2944-2948.

Author's Biography

S Radharani received the B.E degree in Electronics and Communication Engineering, the M.E degree in Applied Electronics in 1985, 1991 respectively, from PSG College of Technology, Coimbatore, Tamil Nadu, India. Her research interests include Digital signal processing and Image processing. She is currently working as a senior grade lecturer in the ECE department of PSG College of Technology, Coimbatore, Tamil Nadu, India. She has around 11 years of teaching and research experience.

P.Keerthana received the B.E degree in Electronics and Communication Engineering from Vivekananda College of Engineering for Women in 2008. She is currently pursuing M.E degree in Communication Systems from PSG College of Technology.

Ponnusamy Thangapandian Vanathi received the B.E degree in Electronics and Communication Engineering, the M.E degree in Computer Science and Engineering and the PhD in 1985, 1991and 2002 respectively, from PSG College of Technology, Coimbatore, Tamil Nadu, India. Her research interests include Soft computing, Digital signal processing and Image processing. She is currently working as an assistant professor in the ECE department of PSG College of Technology, Coimbatore, Tamil Nadu, India. She has around 18 years of teaching and research experience. She is a life member of ISTE. She has published in 10 national and international journals and 50 national and international conference publications.

Kandasamy Gunavathi received the B.E degree in Electronics and Communication Engineering, the M.E degree in Computer Science and Engineering and the PhD in 1985, 1989and 1998 respectively, from PSG College of Technology, Coimbatore, Tamil Nadu, India. Her research interests include Neural Networks, Digital signal processing and Image processing. She is currently working as a professor in the ECE department of PSG College of Technology, Coimbatore, Tamil Nadu, India. She has around 20 years of teaching and research experience. She is a life member of ISTE. She has published in 20 national and international journals and 60 national and international conference publications.