BREAST CANCER DIAGNOSIS USING NEURAL NETWORKS

Submitted to

Sir Junaid Akthar

Submitted by

Aqsa Samreen

14031250

1. Introduction

20 October is the awareness day for breast cancer worldwide But this battle against this disease is far from finished. Breast cancer is a common problem in both developed and under developed counties. It is evaluated that more than 508000 women died in 2011 due to breast cancer (Global Health Estimates, WHO 2013).Developed countries faces almost 50% breast cancer cases and 58% deaths occur due to this disease in under developed counties (GLOBOCAN 2008)[1].Artificial intelligence is now widely used in many real world problems including breast cancer classifiers and systems. This report explains that how neural network classify the dataset of breast cancer to identify either patient is facing cancer or not. This report covers all the steps from basic structure of the network to the analysis of the results.

2. Background

A neuron is the unit of a neural network. Inputs of the dataset first passed with defined weights to a transfer function, which generate an output. Output values are positive if the values generated by the transfer function are greater than threshold of the neuron.

There are several activation function used to train a network for example an activation function known as 'tansig' which returns a bipolar value as a result of operating on the given inputs.

Iranpour, et al. discussed the application of Support Vector Machines (SVM), Radial Basis Function (RBF) networks for breast cancer detection and obtained an accuracy of 98.1% which is compared favorably to accuracies obtained in other studies like linear SVM classifier (94%), fuzzy classifiers (95.8%), and edited nearest neighbor with pure filtering (95.6%). (Menaka & Karpagavalli, 2013).

Bat optimization algorithm is proposed for the selection of appropriate features from the WDBC dataset. This is one of the optimization algorithms that optimize the features of the breast cancer dataset to increase the accuracy of final results.[2]

3. Preprocessing

Wisconsin Diagnostic Breast Cancer (WBDC) dataset contains 699 breast cancer cases with their diagnosis as benign or malignant. The dataset has been divided into two datasets one for the input, which is passed to back propagation neural network and other one, is output dataset. Total 11 columns are in dataset. The first column has id numbers. Input is followed from the column 2 to column 10 in the dataset. Last column is output column having two type of values i-e 2 or 4,2 is for benign and 4 for malignant.

Column 6 of the input has 16 missing values are represented as '?' in the dataset. To solve this problem '?' is replaced with the median of that column which is '1'.

Matlab is used as platform for creating and training a neural network. Neff is used as function for creating the neural network. Preprocessing is completed.

4. Network Structure

Backpropogation network is created by using newff function as describe before. Input and output data pass to newff with 20 hidden layers in network. Tansig,trainr,learngd and mse used as activation functions are set with default values. In the next step net made by newff passed to train function. Use the activation functions of training to set the number of generations to 100 and goal value 0.01. After training use function y=net(input) which return the y as output. Set threshold for converting the output to 2 and 4.After that find accuracy using y=minus (a, b) where a is output dataset and b is dataset obtained by function net(input).

5. Experimental Results and Analysis

For the purpose of experiment, separate the data into different number of proportions like 80 by 20, 60 by 40, 50 by 50, 40 by 60 and 20 by 80. These proportions of input and output data used for the different experiments and for the results to the above described network.

i) Hypothesis 1

- **Hypothesis:**

If we reduce the number of data for training, it will reduce the accuracy of the network and vice versa.

- **Effects of data distribution to accuracy and error:**

Training Data (%)	Testing Data (%)	Accuracy (%)	Error
80	20	98.5714	1.4286
60	40	97.4910	2.509
40	60	96.8974	3.1026
20	80	95.5277	4.4723

- **Analysis:**

Above table shows that as we are reducing the data for training, accuracy is

also reducing and error is increasing which means that if the network is trained on a large dataset than mean square error will be low. Network will be able to detect more data precisely.

- ## Result:

After the analysis, it is proved that first hypothesis is right.

ii) Hypothesis 2

- ### Hypothesis:

If network is trained on more data and tested on data less than trained data then Accuracy for trained data will be less than tested data.

- ### Effects of data distribution to accuracy:

Training data (%)	Testing Data (%)	Accuracy (%)
80	80	96.2433
80	20	98.5214
60	60	95.2267
60	40	97.4910

- ### Analysis

Above table shows the result of different experiments to prove the hypothesis. If the training data is 80% and testing data is also 80% then its accuracy is less than when the training data is 80% and testing data is 20%.Because the network is trained on a large data and it is easy for the network to detect the right values for a small number of testing data as compare to 80% dataset for testing which is equal to the number of trained data set. Therefore, accuracy of detecting the right values for less testing data is greater than more testing data.

- ### Result

Results of the accuracy show that the hypothesis is correct for this system.

iii) Hypothesis 3

- ### Hypothesis

  Increase in learning rate will lead to the decrease in accuracy.

- ### Effects of Learning rate on accuracy:

Learning Rate	Accuracy (%)	Error
0.05	97.28	2.72
0.04	94.99	5.01
0.03	93.84	6.16

- ### Analysis

  From the above table as we are increasing the learning rate accuracy is decreasing and error is increasing. Its means that by deceasing the learning rate lead to decrease the accuracy.

- ### Result

  Above analysis show that our hypothesis is not correct.

6. References

[1]http://www.who.int/cancer/detection/breastcancer/en/index1.html

[2]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5555532/

[3] Iranpour M, Almassi S and Analoui M, “Breast Cancer Detection from fna using SVM and RBF Classifier”, In 1st Joint Congress on Fuzzy and Intelligent Systems, 2007.

[4]Menaka, K., & Karpagavalli, S. (2013). Breast Cancer Classification using Support Vector Machine and Genetic Programming. *International Journal of Innovative Research in Computer and Communication Engineering*, 1410–1417.