NNFS-Coursework-1

Classification of breast cancer diagnosis

November 30, 2017

**Qadeer Ahmad**

**UB: 14031224**

­



*(An Associate College of University of Bradford, UK)*

# Course-Instructor

Dr. Junaid Akhtar

# Reviewer-1

Zulqarnain Haider

# Reviewer-2

Muhammad Zeeshan

Table of Contents

[Classification of breast cancer diagnosis 1](#_Toc500850655)

[Abstract 2](#_Toc500850659)

[Introduction 3](#_Toc500850660)

[Background Information 3](#_Toc500850661)

[Literature Review 3](#_Toc500850662)

[Pre-Processed Data 4](#_Toc500850663)

[Technical Approach 4](#_Toc500850664)

[Hypotheses & Results 5](#_Toc500850665)

[Hypothesis #1: 5](#_Toc500850666)

[Hypothesis #2: 6](#_Toc500850667)

[Hypothesis #3: 6](#_Toc500850668)

[Hypothesis #4: 6](#_Toc500850669)

[Findings & Outcomes 7](#_Toc500850670)

[References 7](#_Toc500850671)

## Abstract

The program or set of instructions to self-learn was the time of era, when the world started turning towards the Automated Systems. The native purpose to evolve & develop such programs was to reduce the human efforts & interactions and to make those systems self-operative. Breast Cancer is one of the increasing issue in the society. It can be identified & solved by the modern technologies. Well, there are many approaches to discover that symptom. One of the best known approach is the Artificial Neural Network that uses automated learning & testing. The network usually takes the set of training data as an input to familiarize itself with that problem. Then, it is tested on a different set of data to validate its performance from the previously learned training data. The result generated by that evolved neural network is more than 90% accurate.

## Introduction

Breast cancer is the most common cancer in women worldwide generating nearly 1.7 million new cases in 2012 and the second most common cancer overall. This amounts to about 12% of all new cancer cases and 25% of all cancers in women **[2]**. Keeping these things intact, it is natural to think that using artificial neural network is one of the most common & widely researched way of finding the solution of breast cancer diagnosis. The basic purpose for using neural networks is that they are relatively faster and accurate in extracting the better results and also good in diagnosing breast cancer patients based on their symptoms. This report gives an insight of how we can use a neural network for classifying cases of breast cancer using a backpropagation neural network, covering all the stages starting from designing the network to analyzing the results obtained from the network.

## Background Information

The basic structure of neural network comprises of neurons with input, hidden and output layer as a building block of a neural network, weights to tune & transfer function. An artificial neuron is a device with several inputs and one output that models certain properties of biological **[1]**. The inputs have a specific set of weights assigned to them and these inputs are passed to a transfer function which generates an output. It will produce the positive output, when the value of that output is greater than the threshold of the neuron. There are several transfer functions which can be applied to a neuron. The transfer function differs with the nature of the output that is to be generated by it. Let’s suppose, an activation function named as ‘tansig’ that returns a bipolar value after set of operations on the given inputs. Another function named ‘logsig’ returns unipolar values instead of bipolar e.g. 0 or 1 or something in between. A set of neurons used to create a neural network can be used to solve a given problem.

## Literature Review

Classification of breast cancer has always been a noticeable problem in the field of computing. The scenario given above has been used in the Wisconsin Diagnostic Breast Cancer (WBDC) dataset as an input for various types of neural networks. Senapati ET. Al (2013) used a linear wavelet neural network in combination with ‘Firefly algorithm’ on the Wisconsin Diagnostic Breast Cancer (WBDC) data set. The results of the experiment were as follows. Overall classification accuracy of the experiment was 98.14% while accuracy for malignant tumors was 99.6% and that of benign tumors was 97.14%. The results were much more accurate and near to truth. Another research was performed in which the researchers used “Association rules” in the preprocessing and then the Neural Networks were used for processing the data which gave an accuracy of 97.4% **[3]**.

El-Sebakhy et al. (2006) used the same dataset but used functional networks as a classifier scheme for breast cancer classification. Their overall accuracy was 96.8%. As we can see that the overall efficiency mentioned above is more than 95%. The chances of identifying the symptoms is very high in it.

## Pre-Processed Data

The dataset used is the same mentioned in the previous section i.e. Wisconsin Diagnostic Breast Cancer (WBDC) dataset. The dataset contains a total of nearly about 700 breast cancer cases with their diagnosis as benign or malignant. The data has been divided into two portions. One division is used as training data to train the backpropagation neural network being used. The rest of the part has been used to test the neural network and record its result for analysis purposes. The dataset contained 11 columns in total. The first was the test case number of a patient followed by 9 symptom columns and a last result column which had just two possible values, 2 for benign and 4 for malignant. The input data sent to the neural network did not contain the first and last column as these were insignificant for the input category. The last column is separated and termed as an output for training the network. There were a few symptom entries in the data set that did not have a value so there was a ‘?’ in that place. To handle that insignificant data, I have removed those test cases from the dataset, in order to avoid errors. The software platform used for creating and testing the network is MATLAB. The function used to create the network is known as ‘newff’. So, in the preprocessing we have separated the input for testing and training.

## Technical Approach

After pre-processing the given dataset, I divided it into two different categories as training and testing datasets, by assigning desired percentage e.g. 30% training & 70% testing etc. The network used is a backpropagation created through the ‘newff’ function. There is one hidden layer which contains 10 neurons and the activation function used for this layer is ‘tansig’. The output layer contains 1 neuron as we only need a single output. The training function used is ‘trainrp’ while the learning function is ‘learngd’. The learning rate for the function is initially set to 0.05, goal is set to 0.01 and the function is set to run for 50 epochs in first hypothesis. In the next hypothesis, the number of hidden layers is increase to 20 and the function is set to run for 250 epochs and the results are mentioned in details in the Figure-1.

The above mentioned solution was achieved by the following way. Firstly, the records of the patients were arranged with the type of disease they have, all the benign in top rows and malignant in bottom rows. Then, it takes the training input equally as benign and malignant e.g. in 20% of training data, 10% would be benign and rest of the 10% would be malignant. Then, we concatenate the training data with the built in function ‘cat’ that merges the data. After that training input & output data is given to ‘newff’ function which has activation & training functions. Those activation functions turn the inputs into more useful data. The functionality of training function updates the weights and bias. Then, it is tested on rest of the patient test cases. After getting & storing the testing outputs from the trained neural network. Then, I have used for loop which starts from 1 to size of that output, each time it compares the results of a neural network with the testing output. At last, I calculated the accuracy of the desired results to check the efficiency of it. It gave me 95%+ accuracy, while keeping in mind my assumptions.

## Hypotheses & Results

### Hypothesis #1:

If the neural network is trained on a large training data set, the accuracy of the net should increase, because the neurons are trained over larger dataset. This experiment takes one hidden layer with 10 neurons, ‘tansig’ and ‘tansig’ as activation functions, and trainr, learngd as training functions.

#### Results

|  |  |  |  |
| --- | --- | --- | --- |
| Training Data % | Testing Data % | Epochs | Accuracy % |
| 90 | **10** | **50** | **98.57** |
| 70 | **30** | **50** | **98.57** |
| 40 | **60** | **50** | **97.61** |
| 20 | **80** | **50** | **96.42** |
| 10 | **90** | **50** | **94.43** |

The results produced in the above figure are somewhat different from the assumed output. So, let’s go for another hypothesis for a better results.

### Hypothesis #2:

In our dataset, 65% of the patient test cases are benign, which will be training data. So, our net is trained more on benign data than malignant data. The remaining unseen data of malignant that is not trained by network, its decision could be right or wrong, well! It depends upon network.

#### Result

According to the experiments based on the 3rd hypothesis. The results produced are 65% accurate and true.

### Hypothesis #3:

If the number of neurons in hidden layer is changed, the accuracy should vary. The number of neurons is kept under the range of input layer and output layer neurons. By decreasing or increasing the neurons should affect the accuracy alike. If the number of neurons are increased from range of inputs, then complex network is created and so the accuracy again will decrease. Activation functions are ‘tansig’, ‘tansig’, and training function is ‘trainr’.

#### Result

|  |  |  |  |
| --- | --- | --- | --- |
| Neurons in Hidden Layers | Training Data % | Testing Data % | Accuracy % |
| 50 | **10** | **90** | **97.9** |
| 50 | **30** | **70** | **97.7** |
| 50 | **70** | **30** | **97.9** |
| 250 | **10** | **90** | **98.2** |
| 250 | **30** | **70** | **95.98** |
| 250 | **70** | **30** | **93.12** |

### Hypothesis #4:

Reducing the training data will also cause a decrease in the accuracy of the results generated by a network. The more cases we provide, the better the network will understand the results and it will also accurately diagnose the case.

#### Result

We can see from the results that our hypothesis was correct. The accuracy does indeed decrease, as we minimize the training data. The obvious reason could be that the network cannot understand/predict all the possibilities of some new cases because of provided training data was not sufficient enough.

## Findings & Outcomes

The problem was to develop the neural network that can detect the breast cancer disease in human. There were two possible types of breast cancer provided in the dataset as benign and malignant. After pre-processing the dataset, a network was developed to validate the accuracy from the network, if a new unseen case was provided to it. Many factors such as the training data, learning rate of network, number of hidden layer neurons affected the outcome of the accuracy. The most accurate result (98.6%) was obtained by using 90% training data to train the neural network, 10, 20 neurons in hidden layer, ‘tansig’ as activation function and ‘trainr, learngd’ as the back propagation function. Another factor which goes hand in hand with this is the number of factors for each input. More factors means the learning rate needs to be lower and vice versa **[4]**. But, for every different problem the settings would also be different. The settings ultimately depend on the inputs and outputs of the provided dataset. Still, the most precise accuracy which gives more than 90% means that neural networks can be one of the best way to detect breast cancer, in our case, the requirement is the Wisconsin Breast Cancer dataset.

## References

|  |  |
| --- | --- |
|  | [1] Globocan.iarc.fr. (n.d.). *Fact Sheets by Cancer*. [online] Available at: http://globocan.iarc.fr/Pages/fact\_sheets\_cancer.aspx [Accessed 27 Nov. 2017].  [2] Swathi, S., Rizwana, S., Babu, A., Kumar, P. and Sarma, P. (2012). *CLASSIFICATION OF NEURAL NETWORK STRUCTURES FOR BREAST CANCER DIAGNOSIS*. [online] http://www.csjournals.com/. Available at: http://www.csjournals.com/IJCSC/PDF3-1/Article\_48.pdf [Accessed 29 Nov. 2017].  [3] M. Karabatak and M. C. Ince. An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications, 36(2):3465{3469, 2009. [Accessed 29 Nov. 2017]. |
|  | [4] Effect of Learning Rate on Artificial Neural Network in Machine Learning. International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181 IJERTV4IS020460 www.ijert.org :Vol. 4 Issue 02, February-2015 [Accessed 30 Nov. 2017]. |