Abid Iqbal

14031233

abid2014@namal.edu.pk

12/2/2017

Abstract

The work consists of the experiments based on the scientific research methodology of hypothesis creation and is applied using artificial neural networks on a problem of breast cancer detection. The motivation was to speed up the process of cancer detection for a quick response.

Breast cancer analysis

Using Artificial Neural Networks

Contents

[1. **Introduction**: 2](#_Toc500363633)

[**2.** **Neural Network Background:** 2](#_Toc500363634)

[**3.** **Main Part:** 3](#_Toc500363635)

[**3.1** **Description of dataset:** 3](#_Toc500363636)

[**3.2** **Pre Processing:** 4](#_Toc500363637)

[**3.3** **Network Architecture:** 4](#_Toc500363638)

[**4.** **Experimental Results and Analysis:** 4](#_Toc500363639)

[**4.1** **Data Distribution:** 4](#_Toc500363640)

[4.1.1 Hypothesis: 4](#_Toc500363641)

[4.1.2 Experimental Results: 5](#_Toc500363642)

[4.1.3 Analysis: 5](#_Toc500363643)

[**4.2** **Activation functions in hidden and output layers:** 5](#_Toc500363644)

[4.2.1 Hypothesis: 5](#_Toc500363645)

[4.2.2 Experimental Results: 5](#_Toc500363646)

[4.2.3 Analysis: 5](#_Toc500363647)

[**4.3** **Learning Rate:** 6](#_Toc500363648)

[4.3.1 Hypothesis: 6](#_Toc500363649)

[4.3.2 Experimental Results: 6](#_Toc500363650)

[4.3.3 Analysis: 6](#_Toc500363651)

[**4.4** **Training Functions:** 6](#_Toc500363652)

[4.4.1 Hypothesis: 6](#_Toc500363653)

[4.4.2 Experimental Results: 6](#_Toc500363654)

[4.4.3 Analysis: 7](#_Toc500363655)

[**5.** **Conclusion:** 7](#_Toc500363656)

[**6.** **Bibliography and Citations:** 7](#_Toc500363657)

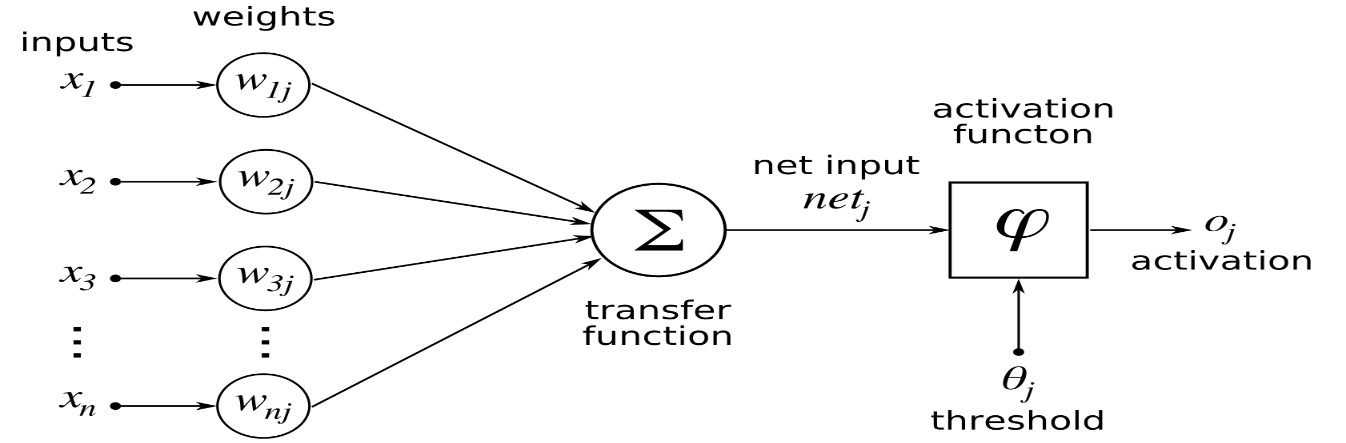
# **Introduction**:

Breast cancer is the most common disease in woman worldwide. According to a survey more than 40,000 women died because of this in Pakistan[1]. There has been an alarming increase in the cases of breast cancer and it was responsible for almost 30.8 % of all cancer deaths among woman in Pakistan. The seriousness of the situation requires a quick diagnosis of the disease for a quick treatment. Keeping in view the severity of the situation, many efforts were put to diagnose and analyze this disease in the past. With the evolution of technology and a considerable progress in the field of artificial intelligence and machine learning, it is the need of the time to use the field of AI for this problem. Use of this technology in field of health will help to analyze any disease with fewer chances of errors. So for this issue, neural networks have been used in past few years. This report also contains the analysis on the data set of **Wisconsin Diagnostic Breast Cancer (WBDC)** using neural networks.

# **Neural Network Background:**

The idea of neural networks is inspired by human brain. The inner working of human brain are often modeled around the concept of neurons and the network of neurons is known as biological neural network[2].

Following diagram shows the structure of a single neuron.



As it can be seen that it is a network of inputs and outputs. Each input has its corresponding weight associated with it and then with the influence of activation function some output is produced. With change in the weights and bias values the output of the neuron will be changed.

Following diagram shows the structure of a neural network.



A neural network has four layers. Input layer represents the input data being provided to the network hidden layer has a network of the neurons and output layers represents output of the network. Using this output new set of weights is calculated by calculating the error in the output and then network is again trained with new weights until network reached required goal or specified number of iterations is completed.

Hidden layer can have as many neurons as possible because there is no fix limit of neurons for a particular problem. Increase in the number of neurons can lead to more training time and more complexity for the system. So optimal number of hidden layers should be picked for a particular problem using trial and error method[1].

# **Main Part:**

This part contains information about the dataset being used, conditions to be set before processing and the architecture of the network being used for these experiments.

## **Description of dataset:**

As already mentioned the data set used in the analysis is of **Wisconsin Diagnostic Breast Cancer (WBDC).** This dataset is publicly available in the internet and is provided by university of Wisconsin hospital, Madison from Dr. William H. Wolberg[4]. This database has 699 instances and 11 attributes including the class attribute. The 1st attribute is of patient’s id, from 2nd to 10th the values of all the major factors involved in cancer analysis and the 11th column contains the information about the class of cancer (benign or malignant). This column has two possible values 2 and 4. According to the class distribution in the dataset 458 or 65.5% instances are Benign and 241 or 34.5% instances are Malignant.

Table 1 provides the attribute information[4].

|  |  |  |
| --- | --- | --- |
| **Serial No.** | **Attribute Name** | **Possible values** |
|  | Id | Integers |
|  | Clump thickness | 1-10 |
|  | Uniformity of cell size | 1-10 |
|  | Uniformity of cell shape | 1-10 |
|  | Marginal adhesion | 1-10 |
|  | Single epithelial cell size | 1-10 |
|  | Bare nuclei | 1-10 |
|  | Bland chromatin | 1-10 |
|  | Normal nucleoli | 1-10 |
|  | Mitosis | 1-10 |
|  | Class | 2 for benign , 4 for malignant |

## **Pre Processing:**

The available dataset has some entries in which there was not a valid value based on the Table 1. Those entries have “?” placed as these values were missing. If we train and test our network by replacing them with any random number then there would be slight chance of bad classification because a random number is not a valid representation of such data which is of a very dangerous disease. Our goal is to design a network with less error percentage in it. So to resolve the issue of missing values, I have skipped those 16 rows from my dataset. So I am using a dataset of 683 rows whereas original dataset is of 699 rows.

The software used for neural network analysis is MATLAB.

## **Network Architecture:**

I will be using a backpropogation network for analysis. Using the newff function the network will be created. The division of data depends upon the experiment. Initially number of neurons is set to 8 and the activation function used in both hidden and output layer is tansig, training function is trainr and learning function is learngd. Mean square error is used to compute error for updating weights. The goal for the network is set to 0.01 and the network will run at max for 100 epochs.

# **Experimental Results and Analysis:**

In this section I will discuss the experiments performed on the neural network using breast cancer dataset. These experiments were performed on the basis of data, activation functions in hidden and output layers, learning rate and training functions.

## **Data Distribution:**

Following set of experiments were performed on different data distribution in training and testing.

### Hypothesis:

For this experiment it is my hypothesis that with increase in the ratio of training data, the percentage accuracy of the network will be increased and vice versa.

### Experimental Results:

Following table shows the experimental results based on the data distribution.

|  |  |  |  |
| --- | --- | --- | --- |
| **Training Data (%)** | **Testing Data (%)** | **Percentage Accuracy** | **Time of Training** |
| 80 | 20 | 99.27 % | 1 min 58 sec |
| 70 | 30 | 98.53 % | 1 min 34 sec |
| 60 | 40 | 97.89 % | 1 min 22 sec |
| 50 | 50 | 97.66 % | 1 min 7 sec |
| 40 | 60 | 97.31 % | 53 sec |
| 30 | 70 | 96.44 % | 40 sec |
| 20 | 80 | 96.70 % | 27 sec |

### Analysis:

By analyzing the experimental data we can figure out that our hypothesis was right. As the amount of training data has increased the network has got less data to be trained and more data as testing which was unseen for the network. As a result the percentage accuracy was decreased. One more point to look here is the time taken for training. When data was more, more time was taken by the network to complete 100 iterations through all data. As the amount of data has decreased, the time has also decreased.

## **Activation functions in hidden and output layers:**

These experiments were performed on the set of activation functions used in hidden and output layers.

### Hypothesis:

For different combinations of functions the percentage accuracy and time used in training will be changed as compared to first experiment.

### Experimental Results:

Following results show the actual experimental output of the network for different combinations of activation functions used in hidden and output layer.

|  |  |
| --- | --- |
| **Activation functions used in hidden and output layer** | **Average percentage accuracy** |
| (tansig,tansig) | 97.68% |
| (logsig,logsig) | 26.70% |
| (logsig,tansig) | 97.69% |
| (tansig,logsig) | 26.70% |

### Analysis:

After experimental results it is clear that using tansig function in output layer has produced maximum accuracy in both cases. It was noted that the avarage time required for training in all of the above experiments was almost same but the accuracy was the main factor which has changed with change in the functions of different layers. So the first part of my hypothesis is proven right but second part is proven wrong.

## **Learning Rate:**

Learning rate is the speed at which the network gets trained on the provided data[5]. It plays a key role in the accuracy of the network.

While performing these experiments the transfer functions were both tansig, epochs were set to 100 and data was divided into 50% training and 50% testing.

### Hypothesis:

If the learning rate is low the training will be better and the percentage accuracy will be high and vice versa.

### Experimental Results:

Following is the result of the experiments performed on the learning rate.

|  |  |
| --- | --- |
| **Learning rate** | **Percentage accuracy** |
| 0.001 | 97.95% |
| 0.01 | 97.36% |
| 0.1 | 96.19% |

### Analysis:

If the learning rate is made too large, the algorithm becomes unstable. If the learning rate is set too small, the algorithm takes a long time to converge[6]. According to these experiments when the learning rate is increased the percentage accuracy is decreased but on the other hand the time of training has also reduced with higher learning rate. The reason being that more the learning rate is, the step will be bigger and as a result it will take less time to complete required iterations in data.

## **Training Functions:**

The function used in training of network is also very important while analyzing accuracy. On different training functions the accuracy is different.

While performing these experiments the transfer functions were both tansig, epochs were set to 100, learning rate was set to 0.001 and data was divided into 50% training and 50% testing.

### Hypothesis:

Accuracy of the network and time required for learning will change with change in the learning functions.

### Experimental Results:

Following is the experimental output of the network.

|  |  |  |
| --- | --- | --- |
| **Learning Function** | **Percentage Accuracy** | **Time of Training** |
| trainr | 97.95% | 1 min 3 sec |
| trainlm | 97.07% | Within 1 sec |
| trainbfg | 98.53% | Within 1 sec |
| trainrp | 98.53% | Within 1 sec |
| traingd | 97.66% | 11 sec |

### Analysis:

All of the training functions have got good accuracy percentage for this dataset. But the factor which is important here is time. trainr takes much time for training because it uses random data picking strategy to train network. So for larger dataset the required time will be more. All other training functions have good behavior in terms of time but trainlm function requires more memory to complete. So all of these constraints are in parallel. One has to pick one based on his dataset and requirements.

# **Conclusion:**

The purpose of this study was to discover new techniques and analyze different parameters of neural networks to improve the accuracy and efficiency of breast cancer analysis system. In this study I have focused on four parameters and having experiments on those, I have discussed optimal values for the given data to produce more accurate results in less time. Those experiments included data distribution, activation functions, learning rate and training functions. It is observed in experiments that less data results in less accuracy but activation functions are very important to be chosen carefully for good results. On the other hand learning rate and training functions play their major part in efficiency of the system.

# **Bibliography and Citations:**

[1] DAWN NEWS, “‘Over 40,000 die of breast cancer every year in Pakistan.’” [Online]. Available: https://www.dawn.com/news/1319675. [Accessed: 03-Dec-2017].

[2] {innoarchitech}, “Artificial Intelligence, Deep Learning, and Neural Networks Explained.” .

[3] M. Chen, “DESIGN OF EXPERIMENTS ON NEURAL NETWORK ’ S PARAMETERS OPTIMIZATION FOR TIME SERIES FORECASTING,” pp. 369–393, 2013.

[4] H. Thazin, T. Thein, K. Mo, and M. Tun, “A N A PPROACH F OR B REAST C ANCER D IAGNOSIS,” vol. 6, no. 1, pp. 1–11, 2015.

[5] P. P. Balestrassi, E. Popova, A. P. Paiva, and J. W. M. Lima, “Neurocomputing Design of experiments on neural network ’ s training for nonlinear time series forecasting,” vol. 72, pp. 1160–1178, 2009.

[6] H. Demuth, “Neural Network Toolbox.”