|  |
| --- |
|  |
| Breast Cancer Diagnosis using Neural Network |
| Coursework-1 |
| 14031273 |
| **Naira Rahim** |
| **12/3/2017** |

|  |
| --- |
| Abstract |

Breast cancer is becoming a serious issue now a days and its diagnosis in time can be very helpful in its control. To help doctors in this diagnosis, a self- learning system is proving very helpful in this regard. Artificial Neural Network is an intelligent system that can help in this diagnosis through learning. For this purpose, the network is first trained on the available data and in return the system learns how to predict in the future on the unseen data. This system can provide an accuracy of more than 90% in most of the cases and can be very accurate.

Contents

[Abstract 1](#_Toc500090963)

[1. Introduction 3](#_Toc500090964)

[2. Background 3](#_Toc500090965)

[3. Data Processing 3](#_Toc500090966)

[4. Experimental Results and Analysis 3](#_Toc500090967)

[4.1 Hypothesis 1 3](#_Toc500090968)

[4.1.1 Experiment 1: 3](#_Toc500090969)

[4.1.1.1Results: 4](#_Toc500090970)

[4.1.2 Experiment 2: 4](#_Toc500090971)

[4.1.2.1 Results: 4](#_Toc500090972)

[4.1.3 Conclusion: 4](#_Toc500090973)

[4.2 Hypothesis 2 4](#_Toc500090974)

[4.2.1 Experiment: 4](#_Toc500090975)

[4.2.1.1 Results: 4](#_Toc500090976)

[4.2.2 Conclusion: 4](#_Toc500090977)

[4.3 Hypothesis 3 4](#_Toc500090978)

[4.3.1 Experiments: 5](#_Toc500090979)

[4.3.1.1 Results: 5](#_Toc500090980)

[4.3.2 Conclusion: 5](#_Toc500090981)

[4.4 Hypothesis 4 5](#_Toc500090982)

[4.4.1 Experiment 1: 5](#_Toc500090983)

[4.4.1.1 Results: 5](#_Toc500090984)

[4.4.1.2 Conclusion: 5](#_Toc500090985)

[4.4.2 Experiment 2: 5](#_Toc500090986)

[4.4.2.1 Results: 6](#_Toc500090987)

[4.4.2.2 Conclusion: 6](#_Toc500090988)

[5. Bibliography 6](#_Toc500090989)

# Introduction

The importance of AI in today’s world can never be denied. Since the very beginning researchers are finding new ways to develop systems that are intelligent. Learning is an important aspect of intelligent systems (Peace and Ita, 2015). Neural networks are now being used in most of the fields of science and are proving fruitful in solving and diagnosing many common problems faced by people around the globe.

# Background

Breast cancer is a disease which is affecting women and is becoming common among them in this time. The battle against this cancer is long way from finish(Tike Thein and Mo Tun, 2015). As per the results showed by the American Cancer Society (ACS), breast cancer makes up to 25% of all the cancer being diagnosed in women all across the globe (*Worldwide statistics on breast cancer: Diagnosis and risk factors*, no date).

While keeping in mind the importance of AI in the medical field, neural network is being used as a power tool in the diagnosis of breast cancer. This report gives an understanding of how we can utilize a neural system for grouping instances of breast cancer by designing our network and then analyzing the results obtained from this network.

The data used for this classification of breast cancer is obtained from Wisconsin Diagnostic Breast Cancer (WBDC), which have been used by many others in the development of their networks as well.

# Data Processing

The data contains a total of 699 cases with their diagnosis as benign or malignant along with other features which distinguish them both. Initially we found some missing information in our data and a “?” was present instead of any value. To make those rows more meaningful we replaced these “?” with -1 so that we may start our processing on our dataset. The dataset contain 11 columns, the first of which is the ID for uniquely identifying the data and the columns from 2 to 10 contains the symptoms of the cancer and the last column is the result columns declaring the symptoms to be of benign or malignant.

# Experimental Results and Analysis

A neural network is thus developed by using the function “newff”. A single hidden layer is used in this network and activation function “tansig” is used in this network. The epochs are set “100” and the goal is set to “0.01” in this network.

## 4.1 Hypothesis 1

The structure of the expected output affects the efficiency of the results.

### 4.1.1 Experiment 1:

Since we are given data in the form of columns and the last column represents whether the cancer is described as benign or malignant represented as 2 and 4 respectively. So we train the data while representing them as 2 and 4. The expected output matrix is a matrix with single column having values 2 for benign and 4 for malignant. Keeping the fact in mind that the data missing in any of the columns in the dataset is replaced by -1 in that column.

### 4.1.1.1Results:

|  |  |  |  |
| --- | --- | --- | --- |
| Training Data (%) | Testing Data (%) | Accuracy (%) | Error (%) |
| 70 | 30 | 97.6077 | 2.3923 |

### 4.1.2 Experiment 2:

Now we will change our expected output matrix to be a 2 dimensional matrix that is if the last column has value 2, our new matrix should have values 1, 0 and when the last column has value 4, our new matrix should have values 0, 1.

### 4.1.2.1 Results:

|  |  |  |  |
| --- | --- | --- | --- |
| Training Data (%) | Testing Data (%) | Accuracy (%) | Error (%) |
| 70 | 30 | 99.5215 | 0.4785 |

### 4.1.3 Conclusion:

The second experiment provides more accurate results which prove that the output matrix should be designed in such a form which should reduce the chances of error and thus is more error free.

## 4.2 Hypothesis 2

Increasing the data in training set will result in a more accurate result when the network is tested.

### 4.2.1 Experiment:

A series of experiments is performed to proof this hypothesis by changing the training and testing data. And the first hypothesis is also kept in mind performing these experiments and the results of matrixes contain the second type of matrix which has 2 columns. This is because the percentage of error was less in this matrix.

### 4.2.1.1 Results:

|  |  |  |  |
| --- | --- | --- | --- |
| Training Data (%) | Testing Data (%) | Accuracy (%) | Error (%) |
| 60 | 40 | 98.2079 | 1.7921 |
| 70 | 30 | 99.5215 | 0.4785 |
| 80 | 20 | 99.2806 | 0.71939 |
| 90 | 10 | 100 | 0 |

### 4.2.2 Conclusion:

The result proves our hypothesis that the accuracy of our system is at its maximum that is 100% accurate when 90% of the data is used for training.

## 4.3 Hypothesis 3

Increasing the size of hidden layers in a network should increase the efficiency of the results.

### 4.3.1 Experiments:

Keeping the training and testing data constant that is 70% and 30% respectively, we performed a series of experiments and only change the size of hidden layer in each experiment.

### 4.3.1.1 Results:

|  |  |  |
| --- | --- | --- |
| Size of Hidden layer | Accuracy (%) | Error (%) |
| 5 | 97.1292 | 1.9139 |
| 10 | 99.0431 | 0.9569 |
| 20 | 99.5215 | 0.4785 |
| 25 | 99.5215 | 0.4785 |
| 30 | 98.5646 | 1.4354 |

### 4.3.2 Conclusion:

By increasing the size of hidden layer the accuracy of the network keeps on increasing until a point where it become maximum (in our case maximum accuracy obtained is 99.5215%) and after that the accuracy decreases gradually after becoming constant.

## 4.4 Hypothesis 4

The column where the data was missing is replaced by the mean value of all the columns in the row and it should result in a more accurate answer because now the data is more meaningful.

### 4.4.1 Experiment 1:

We obtain results by changing the training and testing data.

### 4.4.1.1 Results:

|  |  |  |  |
| --- | --- | --- | --- |
| Training Data (%) | Testing Data (%) | Accuracy (%) | Error (%) |
| 60 | 40 | 98.9247 | 1.7921 |
| 70 | 30 | 98.5646 | 0.4785 |
| 80 | 20 | 100 | 0.71939 |
| 90 | 10 | 100 | 0 |

### 4.4.1.2 Conclusion:

The method of replacing the missing data with the median of the row proves to be more efficient because by increasing the training data the accuracy of the network gradually increases and when the error becomes minimum, which in our case is 0% at 80% training data, it remains 0% when the training data becomes 90%. Median has provided our data an accuracy which was not possible by the value -1 because it was totally out of context.

### 4.4.2 Experiment 2:

Now we will check the effect of increasing the size of hidden layer when the median of the rows are added in place of missing data.

### 4.4.2.1 Results:

|  |  |  |
| --- | --- | --- |
| Size of Hidden layer | Accuracy (%) | Error (%) |
| 5 | 99.0431 | 0.9569 |
| 10 | 99.0431 | 0.9569 |
| 20 | 99.5215 | 0.4785 |
| 25 | 98.5646 | 1.4354 |
| 30 | 98.0861 | 1.9139 |

### 4.4.2.2 Conclusion:

The results are almost the same as that obtained when -1 was present in the place of missing data. Another observation which we can make from the results is that when the size of the hidden layer was small, the accuracy remained constant until it reaches its highest value and after that the usual trend starts that the accuracy keeps on decreasing gradually.

# Bibliography

Peace, I. C. and Ita, S. A. (2015) ‘Effect of Learning Rate on Artificial Neural Network in Machine Learning’, 4(2), pp. 359–363.

Tike Thein, H. T. and Mo Tun, K. M. (2015) ‘An Approach for Breast Cancer Diagnosis Classification Using Neural Network’, *Advanced Computing: An International Journal*, 6(1), pp. 1–11. doi: 10.5121/acij.2015.6101.

*Worldwide statistics on breast cancer: Diagnosis and risk factors* (no date). Available at: https://www.medicalnewstoday.com/articles/317135.php (Accessed: 3 December 2017).