NNFS_CW_1

Submitted by: Najaf Khan

DECEMBER 3, 2017

SUBMITTED TO: SIR JUNAID AKHTER

Abstract

Breast cancer is the second leading cause of cancer deaths worldwide and occurs in one out of eight women. Identification of breast cancer is an open study area which can be solved using artificial intelligence technologies. One of the approaches for solving this problem is by using Artificial Neural Networks. It is one of the most intelligent tools designed for automation. This report provides a solution to “Breast cancer classification” using “Neural Network. 99% accuracy was achieved by this system.

INTRODUCTION

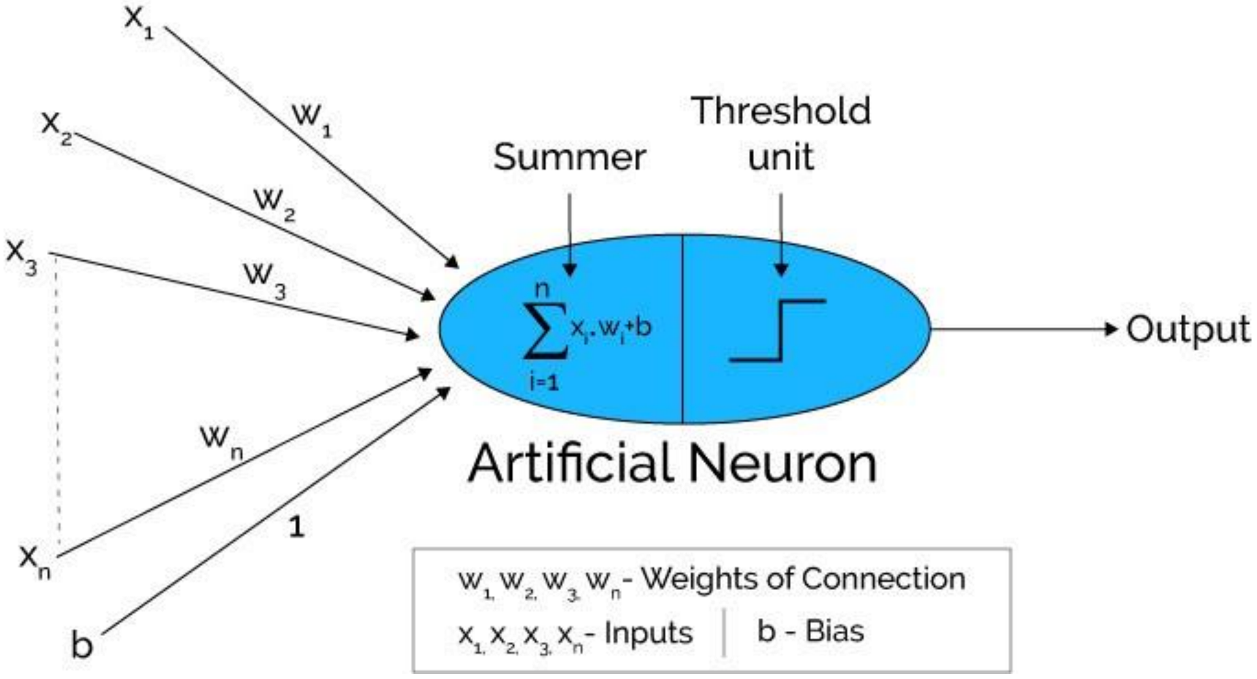
Cancer is a general term that refers to cells that grow larger than 2mm in every 3 months and multiply out of control and spreads to other parts of the body. Breast cancer is the most common cancer in women. In 2012 it resulted in 1.68 million new cases and 522,000 deaths. It is the second most common cancer overall. Even though so much work has been done in health sciences breast cancer remains one of the dangerous disease in women. According to the World Health Organization (WHO): **“There are about 1.38 million new cases and 458000 deaths from breast cancer each year.”** There is still a lot of research going on to detect and control breast cancer and one of the methods is using machine learning and artificial intelligence approach, in which reasonable amount of accuracy has been achieved. This work also shows excellent results using one of the best machine learning technique “Neural Networks”.

BACKDROUND

Basic Concepts

Breast cancer can be classified into different types with respect to its severity and other factors. Some of the well-known types of breast cancer are benign, borderline, and malignant. Here in our data we are processing the two types benign (less dangerous and can be cured) and malignant (Extremely dangerous).

A Neuron is a basic building blocks of the neural networks. Neural networks are typically organized in layers. Layers are made up of several interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to an 'output layer' where the output is computed. The following figure shows the architecture of a neural network.



X_1
w_1
X_2
W_2
X_3
W_3
W_n
X_n
b
1
Summer
Threshold unit
$\sum_{i=1}^{n} x_i \cdot w_i + b$
Output
Artificial Neuron
w_1, w_2, w_3, w_n - Weights of Connection
x_1, x_2, x_3, x_n - Inputs
b - Bias

Related Work

There is a lot of work being done in the field of health sciences especially on detection of breast cancer in early stages as well as controlling it to save lives. There are some previously done researches in this field with some significant results. An approach is proposed by Alolfe and Youssef in 2008 to develop a computer-aided classification system for cancer detection from digital mammograms. M. R. Senapati and A. K. Mohanty proposes to use local linear wavelet neural network for breast cancer recognition by training its parameters using Recursive least square (RLS) approach to improve its performance. Two automated methods are presented to diagnose mass types of benign and malignant in mammograms by Rahimeh Rouhi and Mehdi Jafari. In the first proposed method, segmentation is done using an automated region growing whose threshold is obtained by a trained *artificial neural network* (ANN). In the second proposed method, segmentation is performed by a *cellular neural network* (CNN) whose parameters are determined by a *genetic algorithm* (GA).

MAIN PART

Data Description

Number of Instances: 699

Number of Attributes: 10 plus the class attribute

Attribute Information:

# Attribute	Domain
1. Sample code number	id number
2. Clump Thickness	1 - 10
3. Uniformity of Cell Size	1 - 10
4. Uniformity of Cell Shape	1 - 10
5. Marginal Adhesion	1 - 10
6. Single Epithelial Cell Size	1 - 10
7. Bare Nuclei	1 - 10
8. Bland Chromatin	1 - 10
9. Normal Nucleoli	1 - 10
10. Mitoses	1 - 10
11. Class:	(2 for benign, 4 for malignant)

Missing attribute values: 16 (represented by ”?”)

Class distribution: 1. Benign: 458 (65.5%) 2. Malignant: 241 (34.5%)

Preprocessing

The dataset was taken from UCI Machine Learning dataset repository in which the output values were denoted by 2 for *Benign* while 4 for *Malignant*. The full data consisted of 699 cases. There were some cases where the data was missing so first, missing values were detected in at 16 instances and were replaced by the mean of all the values of that specific column. The first attribute of the dataset was the ID number which was not needed for any processing, so it was removed. Then the data set was divided into two sets in a 70/30 ratio for training and testing purpose respectively. Finally, the values of 2 and 4 were replaced by 0 and 1 for the benign and malignant respectively as it was a binary output and no third output was expected. And it is not necessary to use 0,1 instead of 2,4 for benign and malignant representation.

Network Architecture

For solving the classification problem, last column of the dataset was separated as the desired output and the remaining data was loaded as input data. As there is no specific rule to detect number of neurons in the hidden layers. Hence, for getting the most accurate result, different numbers of neurons in hidden layer were used but mostly 12 were used in experiments. then the data was trained using “newff”. The activation function used was “tansig” while training function “traingdx” was used. In the architecture of neural network one hidden layer and one output layer is used. After training this neural network was tested on different proportions of data and results were achieved. Below we are going to discuss and analyze the results.

EXPERIMENTAL RESULTS AND ANALYSIS:

All results are based on the hypothesis. Moreover, experiments performed with constant weights of neurons, 300 epochs, 0.01 goal, “max_fail” of 100 points and default performance function of mean square error “mse”. the learning rate “lr” was set to the value of 0.02.

Effect of Data Proportion for training and testing

Hypothesis: With the increase in the training data, accuracy should be improved because the algorithm will learn more about the data.

Experiment

- No. of Neurons in Hidden layer = 12
- No. of Hidden Layers = 1
- Activation Functions = ‘tansig’, ‘tansig’
- Training Function = ‘traingdx’
- Learning Function = ‘learngd’
- Number of Epochs = 300

Training Data (%)	Testing Data (%)	Accuracy (%)
10	90	94.753
20	80	95.721

30	70	96.537
40	60	96.912
50	50	97.429
60	40	98.214
70	30	99.047
80	20	100

Result

The experiments were performed based on our hypothesis and the results show that our hypothesis was correct as the higher the data for training higher the accuracy of the system is. The system showed the minimum accuracy when the training data was 10% but it resulted with 100% accuracy when training data was 80%.

Effect of Number of neurons in hidden layer

Hypothesis: Increasing the number of neurons in hidden layer would increase the performance because the system will be using more neurons to learn.

Experiment

- No. of Hidden Layers = 1
- Data Proportion = Training Data (70%), Testing Data (30%)
- Activation Functions = 'tansig', 'tansig'
- Training Function = 'traingdx'
- Learning Function = 'learngd'
- Number of Epochs = 300

No. Of Neurons	Accuracy (%)
1	97.619
2	98.095
4	98.571
8	98.571
12	99.523
24	98.571
48	97.619

Result

Several tests have been experimented based on the above defined hypothesis on 70/30 ratio of training and testing data respectively. As the number of neurons increased in the hidden layer, the accuracy increased. The experiments also show that the hypothesis was correct to some extent because the performance increased till 12 neurons, after that it decreases. That means, increasing neurons can be good till some point after that it causes the reduction in performance because a general rule is that less number of neurons should be used as more neurons cause the overfitting.

Effect of Learning Rate in the Training

Hypothesis: Learning rate defines how slow or fast a neural network would learn so the fast learning rate will result in less accuracy because it can skip somethings, but a slow learning rate may take more time but will result in good accuracy.

Experiment

- No. of Hidden Layers = 1
- No. of Neurons in Hidden layer = 12
- Data Proportion = Training Data (70%), Testing Data (30%)
- Activation Functions = 'tansig', 'tansig'
- Training Function = 'traingdx'
- Learning Function = 'learngd'
- Number of Epochs = 300
- Number of max_fails= 100

Learning Rate	Accuracy (%)
200	94.285
20	96.667
2	97.619
0.2	98.571
0.02	99.523
0.002	99.047
0.00000002	92.857

Result

On the hypothesis above several tests were performed using 70/30 ratio of training and testing data respectively and results were calculated. The results show that our hypothesis is correct till some extent but after that it reduces the accuracy. It is because at too small learning rate neural network learns very slowly and when the number of epochs reaches to it maximum point the neural network is not completely learned which results in a reduced accuracy. The accuracy can be increased by increasing the number of epochs and max_fails if the learning rate is too small.

Effect of Performance Goal

Hypothesis: Performance goal is one of the parameter which results in stoppage of training when reached. So, It seems to me that having lower performance goal results in good accuracy.

Experiment

- No. of Hidden Layers = 1
- No. of Neurons in Hidden layer = 12
- Data Proportion = Training Data (70%), Testing Data (30%)
- Activation Functions = 'tansig', 'tansig'
- Training Function = 'traingdx'
- Learning Function = 'learngd'

- Number of Epochs = 300
- Number of max_fails= 100

Performance Goal	Accuracy (%)
100	77.619
10	84.761
1	86.190
0.1	93.333
0.01	98.571
0.001	98.571
0.00000001	98.571
0.0000000001	98.571
0	98.571

Result

On the above hypothesis multiple experiments were performed and results were calculated. The results show that our hypothesis was correct but results also show that after setting performance goal lower than a specific value in this case which is 0.01 accuracy remains constant. This is because it has been observed by me that accuracy does not solely depend upon performance goal, but others factor too. Neural network stopped training when number of epochs or no of max_fails were reached no matter what the performance goal was, which is the reason why decreasing the performance goal after 0.01 did not affect accuracy.

Conclusion:

Neural networks are quite helpful in classification problem and using neural networks on breast cancer data with excellent accuracy about 99% is a proof of it. This work was focused on the usage of different approaches of neural networks to achieve the best result. It was experimented that the accuracy is dependent on the data distribution for training and testing, training functions, activation functions, number of hidden layers, number of neurons, number of epochs, number of max_fails, learning rate and performance goal etc. At the end it can be concluded that proposed neural network with one hidden layer of 12 neurons using “traingdx” training function, “tansig” activation function and 70/30 ratio of training and testing data respectively can be used to achieve a better classification accuracy of about 99%. But, the overall performance would still be influenced by some other factors as discussed above.

BIBLIOGRAPHY:

1. Janghel, R., Shukla, A., Tiwari, R. and Kala, R. (2010). Breast cancer diagnosis using Artificial Neural Network models. *The 3rd International Conference on Information Sciences and Interaction Sciences*.

2. Neuroph.sourceforge.net. (2017). *Predicting the class of breast cancer with neural networks*. [online] Available at: http://neuroph.sourceforge.net/tutorials/PredictingBreastCancer/PredictingBreastCancer.html [Accessed 1 Dec. 2017].
3. Alolfe, M., Youssef, A., Kadah, Y. and Mohamed, A. (2008). Development of a computer-aided classification system for cancer detection from digital mammograms. *2008 National Radio Science Conference*.
4. Senapati, M., Mohanty, A., Dash, S. and Dash, P. (2011). Local linear wavelet neural network for breast cancer recognition. *Neural Computing and Applications*, 22(1), pp.125-131.
5. Rouhi, R., Jafari, M., Kasaei, S. and Keshavarzian, P. (2015). Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Systems with Applications*, 42(3), pp.990-1002.
6. Mathworks.com. (2017). *Choose a Multilayer Neural Network Training Function - MATLAB & Simulink - MathWorks United Kingdom*. [online] Available at: https://www.mathworks.com/help/nnet/ug/choose-a-multilayer-neural-network-training-function.html [Accessed 1 Dec. 2017].
7. Mathworks.com. (2017). *Train and Apply Multilayer Neural Networks - MATLAB & Simulink - MathWorks United Kingdom*. [online] Available at: https://www.mathworks.com/help/nnet/ug/train-and-apply-multilayer-neural-networks.html [Accessed 1 Dec. 2017].