2018

Breast Cancer Classification Using Neural Networks

MANZOOR HUSSAIN
15026425

SUBMITTED TO | Dr.Junaid Akhtar

Contents

Abstract

Artificial Neural Networks are statistical learning models which have amazing abilities to get useful information from raw and indefinite data. Therefore the neural network can be used to figure out various trends from input data, which are quite complicated for humans.
This report provides a solution to “Breast cancer classification” using “Feed Forward Neural Network .The results were collected from different architectures of neural networks and lastly a solution is proposed that is able to predict the class of a new case with 99.5% accuracy depending upon multiple factors of network creation.

Introduction

According to the World Health Organization (WHO):
“There are about 1.38 million new cases and 458000 deaths from breast cancer each year.” [1]

Breast cancer is a malignant tumor not only in men but also in women both in the developed and the developing countries.If satisfactory diagnosis and treatments are available on early detection, there is a high chance that breast cancer can be cured while if detected late, however, therapeutic treatment is often no longer an option and eventually can cause death. But, there are many ways of detecting humanity from Breast Cancer nowadays .Feed forward neural network, a Neural Network approach has been used commonly for the breast cancer classification nowadays.
In this report, we will have a brief overview of neural networks, solution of this classification problem, and we will test and analyse different hypotheses relating to performance and accuracy of provided solution
To address that issue, various techniques of data mining and artificial intelligence like artificial neural network, decision tree, logistic regression and genetic algorithms were used [2].

Background

Neural network is mostly contained on individual or interconnected models, usually called neurons or nodes. The neurons are basic building blocks of the neural network. A set of sample input(s) are given to neuron, then these inputs are forwarded to a decision making function, which is called activation function. Then that activation function generates the output based on given input.
A neuron takes one or several inputs and one output that models certain biological properties. Different weights are assigned to inputs, these inputs with their weights are given to transfer function which outputs a numerical value. The output is positive if the value is greater than a specific threshold assigned to this neuron otherwise the value is negative zero or negative depending upon the nature of transfer function. There is a long list of such transfer functions which provide different results on same input data set.
So the simple structure is input layer, hidden layers and output layer. If the input is provided to input layer and then to hidden layer and then to output layer, such single directional neural network is termed as “Feed Forward Neural Network”.

Such a neural network is being used in this solution
A research was done by Dr Medhat and Mushammad Farouq about the abilities of support vector machine (SVM) classifications with tree boost and tree forest in analysing the dataset for the extraction of mass features that discriminates the true and false cases.
Here SV showed better results as compared to those of tree boost and tree forest [5].

Moreover, K. Rajiv Gandhi, Marcus Karnan and S. Kanan used Particle Swarm Optimization Algorithm for the classifications of the breast cancer datasets [5].

Main Part

In this section, we will present the solution step by step, from data gathering to training a neural network, obtaining and compiling results and measuring the solution's accuracy.

Description about Dataset

The major details about the dataset which has been collected from Wisconsin Breast Cancer Database (WBCD) was analysed and following information figured out [6] [7].
After analysing the data, following information were found [8]:

Total no's of samples = 699
Total columns = 11
Missing Attributes = 16
Benign Cases (which has been indicated as 2)
Malignant Cases (which has been indicated as 4)
Column 1 = Id Number
Column 2-10 = Sample Input Data
Column 11 = Output Data (Class Attributes)

Details about Each Attribute of Row

Each row consist on 11 columns. The details about each column value is following [8] :

Sample code number
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Marginal Adhesion
Single Epithelial Cell Size
Bare Nuclei
Bland Chromatin
Normal Nucleoli
Mitoses
Class:

Neural Network Formation

Data Processing

The given dataset for the breast cancer is in raw form. In order to use that dataset in the training of neural network, it is quite necessary to arrange the dataset in proper way. The first column is consist on the Id Number, which is quite unnecessary in the training of neural network. So that column will be skipped.
There are some missing values (indicated with "?"), so we will take mean of the all values of the column(s) in which the values are missing, in order to find out the suitable alternate for missing values. While the time of data loading we manually, replaced "?" with 0. Then that 0 was replaced with 3. Because the mean of col 7 (in which values were missing) is 3.

```
mean(inputData(: , 7))
inputData (inputData == 0) = 3;
```

After pre-processing we placed the refined dataset in the variable named as DATA, which we used in our code.

Experimental Results and Analysis

After setting up a working neural network, different hypotheses are made, tested and analyzed to know and understand the working of neural network and effect of different parameters on results. Few hypotheses are described below:

Effect of Number of Neurons in Hidden Layer

Hypothesis

Increase in the number of neurons in hidden layer should increase the performance.

Experimental Outcomes

No's of neurons on hidden layer	Performance	Accuracy (%)
1	0.4	95.1
3	2.04	96.2
5	0.95	95.8
10	2.2	96.5
15	1.20	94
20	2.6	96.00

Table 3: Effect of neurons in hidden layer

Result

So it can be concluded from above experiment that if we increase the number of neurons, our performance will increase because in that case there will more neurons available for transferring data from hidden layer to output layer The performance increased till 10 neurons, after that it decreased. That means, increasing neurons can be beneficial to some extent (Table 3)

Effect of Learning Rate in the Training

Hypothesis

The selection of learning rate matters a lot in the training of the neural network. It is supposed that for the better performance of the neural network in regard of training and accuracy, the learning rate should be small.

Experimental Outcomes

Learning Rate	Performance	Accuracy (%)
0.0001	2.18	93.7
0.001	1.01	96.5
0.01	1.92	94.01
1	1.54	94.7
2	0.68	95.8
5	2.16	95.5
10	1.57	95.00

Result

The results of the experiment has been shown above. Although there is no major difference in accuracy but still there is variation in the results. Which shows that if learning rate is small the accuracy will be maximum (with learning rate 0.001 in above experiment) and opposite

case with high learning rate. The reason is that with large learning rate the weights proceed too far in right direction, which prevent network from better learning, then ultimately the accuracy will suppress.
So it can be concluded that small learning rate increases accuracy but too small learning rate can adversely affect the accuracy of network [9], which can also be seen in above experiment(With learning rate 0.0001 as compare to 0.01) .

Ratio among training data and testing data

If percentage of training data is higher than testing data, neural network will produce more accurate results. More precisely, 70% training data and 30% testing data will produce more accurate results. Now let's see these results:

Experimental Outcome

Sr.	Training data (%)	Testing data (%)	Accuracy (%)
1	10	90	97.45
2	20	80	96.23
3	30	70	95.90
4	40	60	97.84
5	50	50	97.98
6	60	40	98.20
7	70	30	99.04
8	80	20	99.28
9	90	10	98.55

Result:

As random weights are assigned to inputs in initial iteration of an individual case, so, sometimes, result better and sometimes not. This can be seen in above table, 10% training data is producing better result than 20% training data. However, produced results are expected results except the first three and last one. As training data is increasing, accuracy is increasing. It's because, higher the training data, lesser the unseen data to test so higher the accuracy and vice versa. Ideally 70% training data and 30% testing data is fine, because if we increase training data more than that, then the essence of neural networks is gone and may be It will not so practical in life.

Effect of data distribution

Hypothesis:

With the increase in the training data, accuracy should be improved because the algorithm will learn about more possibilities.
Some tests have performed based on the hypothesis. The experimental results (Table 4) show that the hypothesis was correct to some extent. With increase in the training set, the accuracy increased, but on 80/20 distribution of data the accuracy decreased due to the unfair distribution of Benign and Malignant in the dataset.

Experimental Outcome

Training Data (%)	Testing Data (%)	Accuracy (%)
40	60	92.6
50	50	95.7
60	40	96.4
70	30	99.0
80	20	98.6
90	10	98.6

Table 4: Effect of increasing the data set

Result

Highest accuracy of 99.5% accuracy was obtained using only one hidden layer with 12 neurons, tansig, tansig as transfer function and trainbr as the training function. As the confusion matrix shows that 152 biopsies are correctly classified as benign. This corresponds to 77.6% of all 196 biopsies (30% testing data). Likewise, 43 cases are correctly classified as malignant. This corresponds to 21.9% of all biopsies. 1 of the malignant biopsies are incorrectly classified as benign. Overall, 99.5% of the predictions are correct and 0.5% are wrong classifications

Conclusion

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 as well as testing. It can be concluded that accuracy of the neural network depends on various factors: activation functions, training and learning algorithms and especially on distribution of dataset.

Fine data distribution, like 70-30 percentage ratio, in this problem, produces good results. Lower learning rate produces good results but if it is too lower, then performance issues will occur. Same goes for the number of neurons in hidden layer

Bibliography

[1] "WHO | Breast Cancer Awareness Month In October". Available at: http://www.who.int/cancer/events/breast_cancer_month/en/. [Accessed: 3 December, 2016.]

[2] Der-Ming Liou and Wei-Pin Chang, [Online]. Available: https://link.springer.com/protocol/10.1007/978-1-4939-1985-7_12.

[3] [Online]. Available: https://www.google.com.pk/search?q=Artificial+neuron&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjjkv3F_OvXAhWL8RQKHRq-B0cQ_AUICigB&biw=1366&bih=588#imgrc=y--CpZwRgEVUHM:.

[4] [Online]. Available: https://www.google.com.pk/search?q=neural+network&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjD0sWI--vXAhXOI1AKHS0vDXAQ_AUICigB&biw=1366&bih=637#imgdii=e0jyP8eElwTY9M:&imgrc=EXoHkcJRnWRI3M:.

[5] SHELLY GUPTA , DHARMINDER KUMAR ,ANAND SHARMA , “DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS,” *Shelly Gupta et al./ Indian Journal of Computer Science and Engineering (IJCSE).*

[6] D. W. H. Wolberg. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original).

[7] [Online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/.

[8] [Online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names.

[9] D. Randall Wilson and Tony R. Martinez, “The Need for Small Learning Rates,” in *In Proceedings of the 2001 International Joint Conference on Neural Networks (IJCNN’01), 115-119.*, Utah, USA.