Naeem Rashid

Dr. Junaid Akhtar

Neural Networks and Fuzzy System

2 December 2017

Breast Cancer Classification using Neural Network Approach

Breast Cancer is widely spread disease among women. Early detection of malignant tissues using Neural Networks and Image Classification e.g Image Classification using SVM(Support Vector Machine) classifier[1], Linear Programming based Machine Learning technique are used to identify the presence of breast cancer among women. This report addresses the solution of same problem with Feed Forward Neural Network using data provided by UCI Machine Learning dataset repository.

Methodology adopted for this procedure includes creation of feed-forward back-propagation Neural Network in Matlab, Training and testing of network and comparison of results with provided output data to measure the accuracy of the trained network. Hypothesis and their respective results are analyzed to enhance the efficiency and accuracy of trained network during the Training and Testing sessions. Results includes recommendations to train Neural Network for Breast Cancer classification to achieve optimal results.

1. INTRODUCTION

Neural Networks techniques are widely used in medical field for diagnosis of diseases[2]. Artificial Neural Networks(ANNs) computing systems are inspired by the way human brain functions. Learning from previous knowledge is the key part of an ANN. Most of the neural networks are based on "Hit and Trial methods" like the way humans learn from mistakes.

Definition of Artificial Neural Network from his founder Dr. Hecht-Nielson

"*...a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs.*"[3].

1.1 COMPONENTS OF ARTIFICIAL NEURAL NETWORK

A typical ANN consists of three parts.

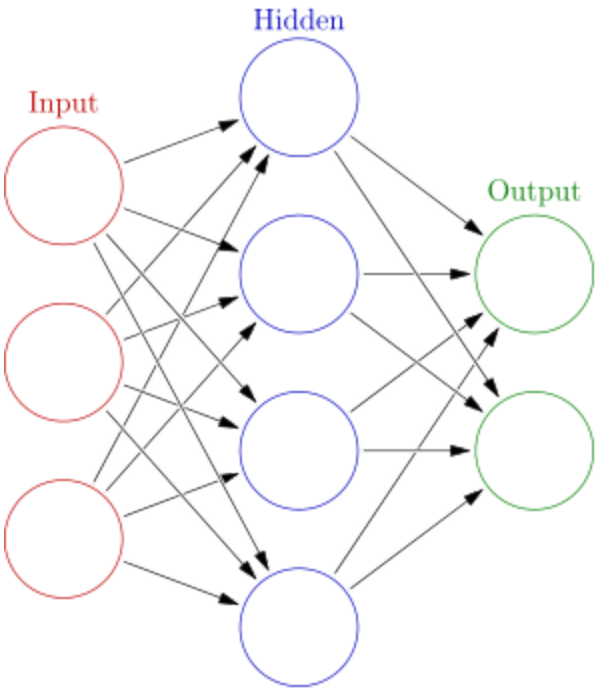
- Neuron
- Input Layer
- Hidden Layer(Optional)
- Output Layer

1.1.1 NEURON

A neuron is the key part of an ANN with a role of central processing it has an input which is numerical representation of information, a weight which depicts the strength of the connection between two neurons and an output which after implementation of various functions to the input e.g Transfer function, Unit step function, sigmoid [4] etc. has a meaningful representation of it.

1.1.2 LAYER

Layers are consists of interconnected neurons. Each layer has a specific operation assigned to it. Input layer receive information in numerical representation, hidden layer implements ANN's deep learning and output layer represent the result of the processed input.



graphical representation of ANN[5]

1.2 TYPES OF LEARNING AND BACK-PROPAGATION ARTIFICIAL NEURAL NETWORK

Two major types of learning strategies are involved in training of an ANN.

- Supervised Learning
- Un-supervised Learning

1.2.1 SUPERVISED LEARNING

This type of learning takes place when the result of the dataset is known to the trainer e.g pattern recognition in image processing[7], classification problems.

1.2.2 UNSUPERVISED LEARNING

Un-supervised learning is used when dataset with known answer is not available e.g clustering, association problems[6].

1.2.3 BACK-PROPAGATION NEURAL NETWORK

Back Propagation algorithm calculates the error function at each steps and reduced the error function using activation function, usually, gradient descent at each iteration by adjusting the weights(randomly assigned at start) to achieve the goal output[8]. Back Propagation networks are basic networks in machine learning and are ideal for simple mapping and pattern recognition tasks.

2 METHODOLOGY

2.1 DATASET DESCRIPTION

Breast Cancer data is available at UCI Machine Learning dataset repository.Dataset includes file named breast-cancer-wisconsin.data having 11 columns. First column includes id of the patient and last one includes the description whether patient has a breast cancer or not with values 2 and 4 i.e 4 being malignant and 2 being benign. Column 2 to 10 with total of 699 rows are used for ANN training and testing. Results are compared with eleventh column to calculate the accuracy.

Dataset includes some missing values with "?" which are replaced by the mean value of that column by replacing all values having (?) with 0 and then using Matlab helper function *mean* to replace 0 with column's mean value.

2.2 NETWORK TRAINING

Using Matlab helper function *newff* a feed-forward back-propagation network is built. Different optional parameters of *newff* includes learning rate, learning epochs, error tolerance, activation function etc. Once network created it was sent to the training using *train* function. Once completion of training network is tested with testing data. Last step include the accuracy measurement of trained network by comparing ANN's results with given results.

3. EXPERIMENTS, RESULTS AND ANALYSIS

In order to detect optimal settings for ANN trained to classify Breast Cancer different hypothesis were made and experiments are conducted to verify those hypothesis. Following are some hypothesis with their experimental results and analysis.

3.1 DATASET DISTRIBUTION

Dataset distribution plays vital role in neural network training. Better results can be achieved with balanced dataset. Before running the tests dataset having benign cases is separated from malignant ones resulting two separate datasets one holding benign cases and other malignant. While test data includes equal amount of both sets. Dataset includes 458 sets of benigns cases and 241 sets of malignant ones. Half of the data is used for testing and other half for training. Following are the experiments performed with different dataset distributions.

Malignant Data(%)	Benign Data(%)	Result (Accuracy %)
1	99	74.4990
10	90	84.24
30	70	97.13
50	50	98.00
70	30	98.85
90	10	97.13
99	1	35.53

As it is clear from the results that an ANN trained on equal distribution of dataset gives best accuracy. Since the benign cases are almost double of malignant ones neural network trained with 70 % of malignant and 30% of benign gives the best accuracy.

3.2 TRAINING AND TESTING DATASET RATIO

In order to stabilize data distribution different ratios of training and testing data is experimented. Generally increase in dataset used to train neural network increases overall accuracy rate but has impact on neural network learning time. Increase in examples increases diversity and generality of trained network. Following results proves the point.

Training Data	Testing Data	Accuracy
90%	10%	100%

80%	20%	99.2754%
70%	30%	98.0770%
60%	40%	98.5612%
50%	50%	97.9885%
40%	60%	96.8900%
30%	70%	96.9263%
20%	80%	96.5950%
10%	90%	96.9745%

Results of the experiment shows that with increase in training data accuracy get increased. But with decrease in training data accuracy of ANN get decreased and results get unpredictable as show from set 40-60 to 30-70 and 20-80 to 10-90.

3.3 HIDDEN LAYERS

Ideally increase in hidden layer would increase the accuracy rate. As increase of hidden layer increase the learning time and it is hypothesized that accuracy of ANN would also have been increased. With 50% of training and testing data and constant number of neurons hidden layers are increased and results are calculated as shown below.

No. of Hidden Layers	**Accuracy**
1	97.9885%
5	98.2759%
10	98.8506%
15	98.2759%

20	98.5632%

As shown from above ANN with one hidden layer has the highest accuracy but with increase in hidden layers effect on accuracy is not so marginal. Most of the ANN gets desired result with one hidden layer where small amount of data is involved.

3.4 EPOCHS

Number of times training vectors are used to update the weights are called epoch.

Generally lesser the epoch will give less performance following are the details for different experiments ran to find best epochs for required ANN.

Number of Epochs	**Accuracy(%)**
10	97.9885
20	98.5632
30	97.9885
40	97.7012
60	98.2759
80	97.7012
100	98.5632

Results shown above are quite unpredictable one possible reason could be that weights are assigned randomly at runtime which sometimes increase learning rate and vise versa.

4 CONCLUSION

Neural Networks are already being used for classification of breast cancer. The purpose of this report was to identify various factors that can increase or decrease the performance and accuracy of a neural network. There are infinitely many factors that can be involved in neural network performance but our observations includes data distribution, data quantity, hidden layers and epochs. With the results of experiments it is concluded that following factors results in enhancement of neural network performance.

- Equal data distribution
- Larger training data results good accuracy
- Increasing hidden layers does not necessarily increase in overall ANN's performance
- Lower epoch results low accuracy

Works Cited

1. Anna Rejani, Yi., & Selvi, D. (2009). Early Detection of Breast Cancer Using Svm Classifier Technique. International Journal on Computer Science and Engineering, 1(3), 127–130.
2. Kajan, S., Pernecký, D., & Goga, J. (n.d.). Application of Neural Network in Medical Diagnostics.
3. Artificial Intelligence Neural Networks. (n.d.). Retrieved December 2, 2017, from https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_neural_networks.htm

4. Artificial Neural Network Structure and Functions. (n.d.). Retrieved December 2, 2017, from http://www.saedsayad.com/artificial_neural_network.htm
5. By Glosser.ca - Own work, Derivative of File:Artificial neural network.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24913461
6. supervised and unsupervised learning. (n.d.). Retrieved December 2, 2017, from https://stackoverflow.com/questions/26182980/can-anyone-give-a-real-life-example-of-supervised-learning-and-unsupervised-lear
7. Le Cun Jackel, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D., Cun, B. Le, Denker, J., & Henderson, D. (1990). Handwritten Digit Recognition with a Back-Propagation Network. Advances in Neural Information Processing Systems, 396–404. https://doi.org/10.1111/dsu.12130
8. Benvenuto, N., & Piazza, F. (1992). The backpropagation algorithm. IEEE Transactions on Signal Processing, 40(4), 967–969. https://doi.org/10.1109/78.127967