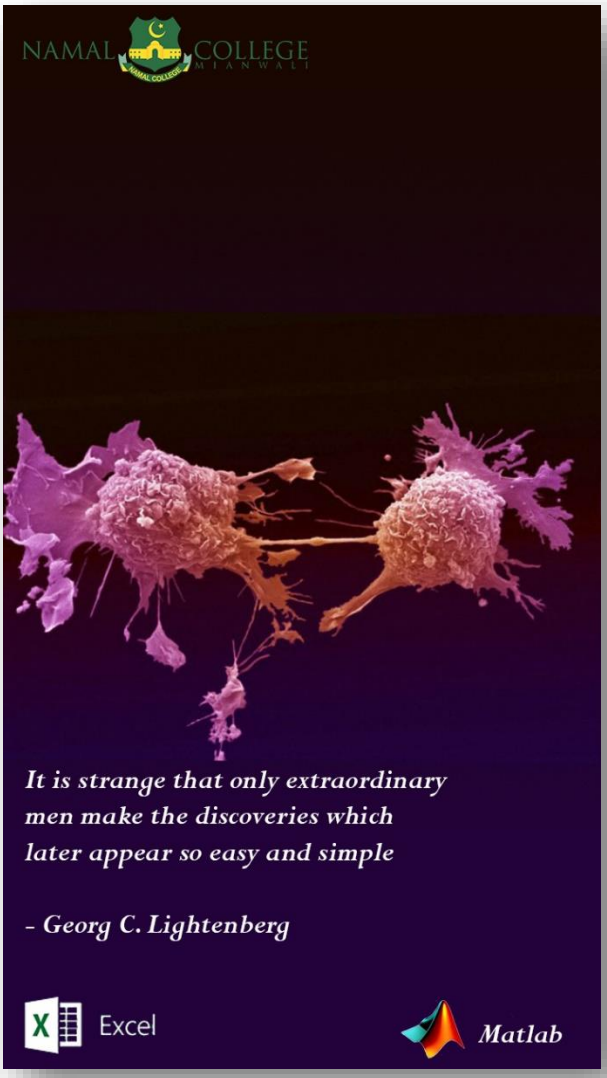


Breast Cancer Classification using Neural Network's Approach

REPORT SUBTITLE

Muhammad Umar | Neural Networks and Fuzzy Systems | 1/12/2017

Introduction

Cancer in few last years has become one of the most effecting disease if not diagnosed at right time or not treated properly, there is ongoing process on how to diagnose it on time and for its cure. Neural networks have gained popularity in past few years as the results generated by this approach are massively successful and accurate, we tried to use the same approach for the diagnosis of breast cancer on having a set of previous cases of diagnosed patients of breast cancer and also the ones without cancer cell.

Background

Breast cancer is one of the most common type of cancer in women, and if not diagnosed at right time, it can lead to even the death of the person. Breast cancer can begin in different areas of the breast like ducts, the lobules or sometimes, the tissues in between (Breastcancer.org, November 9, 2016). We used a dataset of six hundred and ninety nine cases that have been recorded at different periods of time, the dataset used is a thankful contribution of University of Wisconsin Hospital, by Dr. William H. Wolberg.

Sample of the dataset is as:

1. 841769,2,1,1,1,2,1,1,1,1,2
2. 1017122,8,10,10,8,7,10,9,7,1,4

And the classification of the attributes is as follows:

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)

These all datasets are of the tumors, the eleventh attribute defines the tumor as benign or malignant. “A **benign tumor** is a tumor that does not invade its surrounding tissue or spread around the body. A **malignant tumor** is a tumor that may invade its surrounding tissue or spread around the body. ” (Cheprasov, n.d.).

Main Part:

DESCRIPTION OF METHODOLOGY:

We had the dataset of six hundred and ninety nine values, which had some values with not defined result, so those values were removed from the data, then we had six hundred and eighty three value in total, after skimming scamming having eleven attributes, in which, the first one was used for id, likewise the last column was the result on the basis of the other attributes, either the tumor is benign or malignant. The data set got divided into two different matrixes, one; the training data, and the other one was the target result. Training data has nine attributes that are the values of different symptoms on which we train our neural network. Target data was one column of result that is used to check the accuracy percentage at the end, when the trained data generates result.

JUSTIFICATION FOR DESIGN:

In this assignment, it was asked to train neural network using Matlab's neural network simulation. There are many different ways to train dataset on Matlab's simulation, as it was first working experience with neural networks, it was better to get to know about how it works and what are the best conditions and settings to train closer to perfect. Therefore, the dataset was divided into six different combinations by percentage, so that a clear idea of the ratio of the training data and the data being provided to check the performance of the neural network can be generated. The dataset provided is sorted on the basis of the resultant factor, which is from benign factor data to malignant factor data in ascending order. Different combinations of attributes affecting neural network training were tested in favor of hypothesis or in against.

Experimental Results & Analysis:

Before conducting any testing, a hypothesis is generated in whose favor different tests are performed. Likewise, a general hypothesis for neural networks is:

HYPOTHESIS:

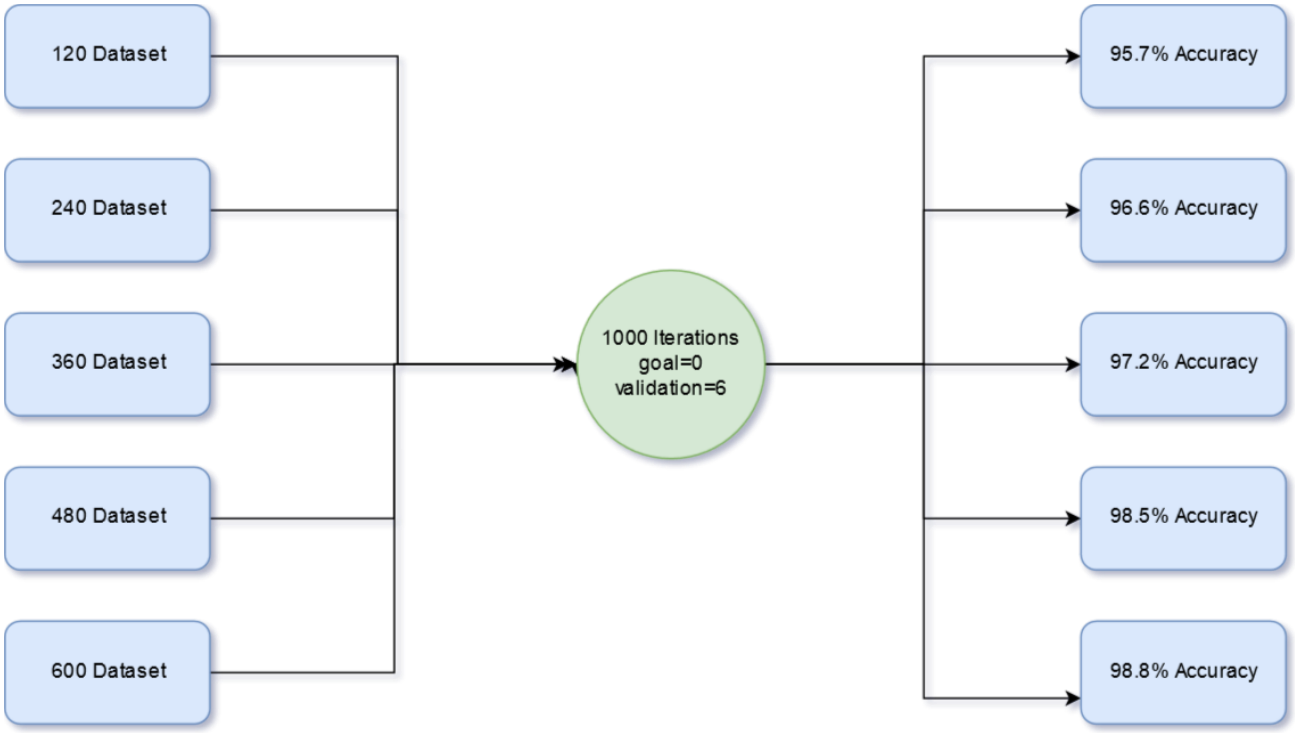
Performance and accuracy of a neural network is directly proportional to the size of the dataset for training a neural network.

Now, to prove hypothesis right, different tests were conducted on different sizes of dataset for training data and likewise for the testing data.
All tests conducted are on same sizes of datasets which are:

1st Dataset		2nd Dataset		3rd Dataset		4th Dataset		5th Dataset	
Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
120	563	240	443	360	323	400	283	600	83

FIRST TRAINING:

The first training for neural network is done using default values with same weights for every set of data, results are as follows:

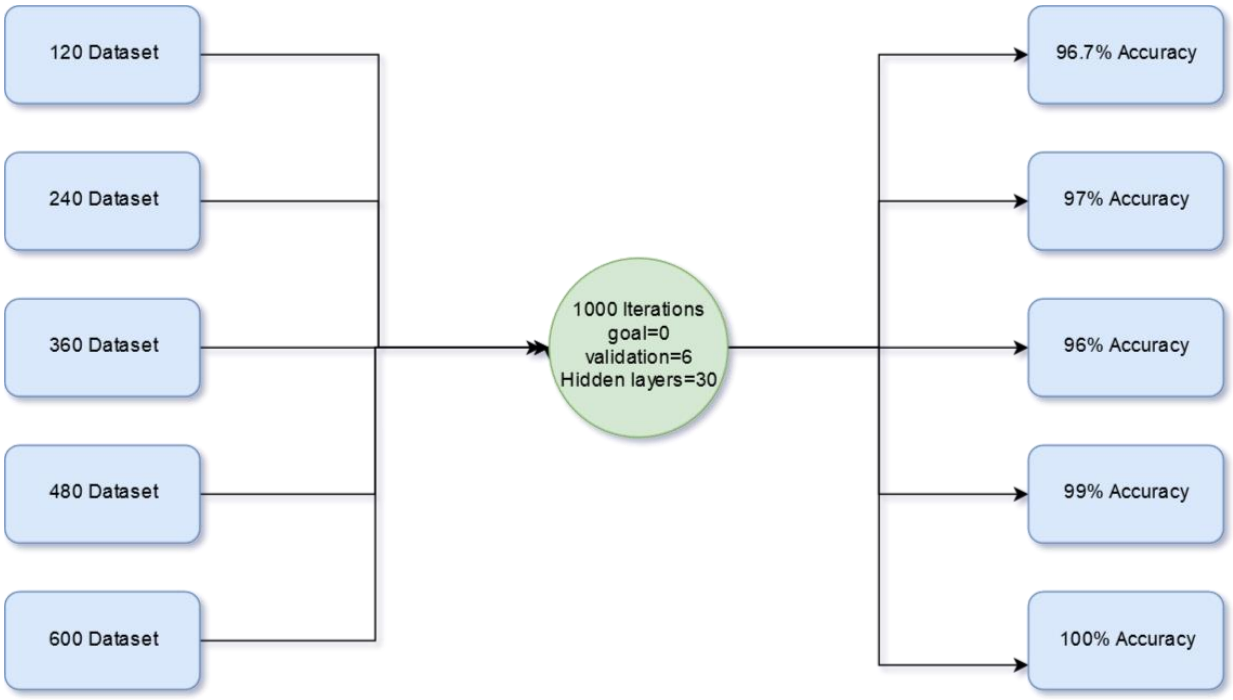


Analysis:

First set of results that was tested having default values with ten hidden layers generated results in the favor of the hypothesis, generating more accurate results as the dataset for training increases with approximately 0.7 percent gain with every increase of around hundred values.

SECOND TRAINING:

Second training was done with the same default values, but with increasing hidden values from 10 to 30, and the results are as follows:

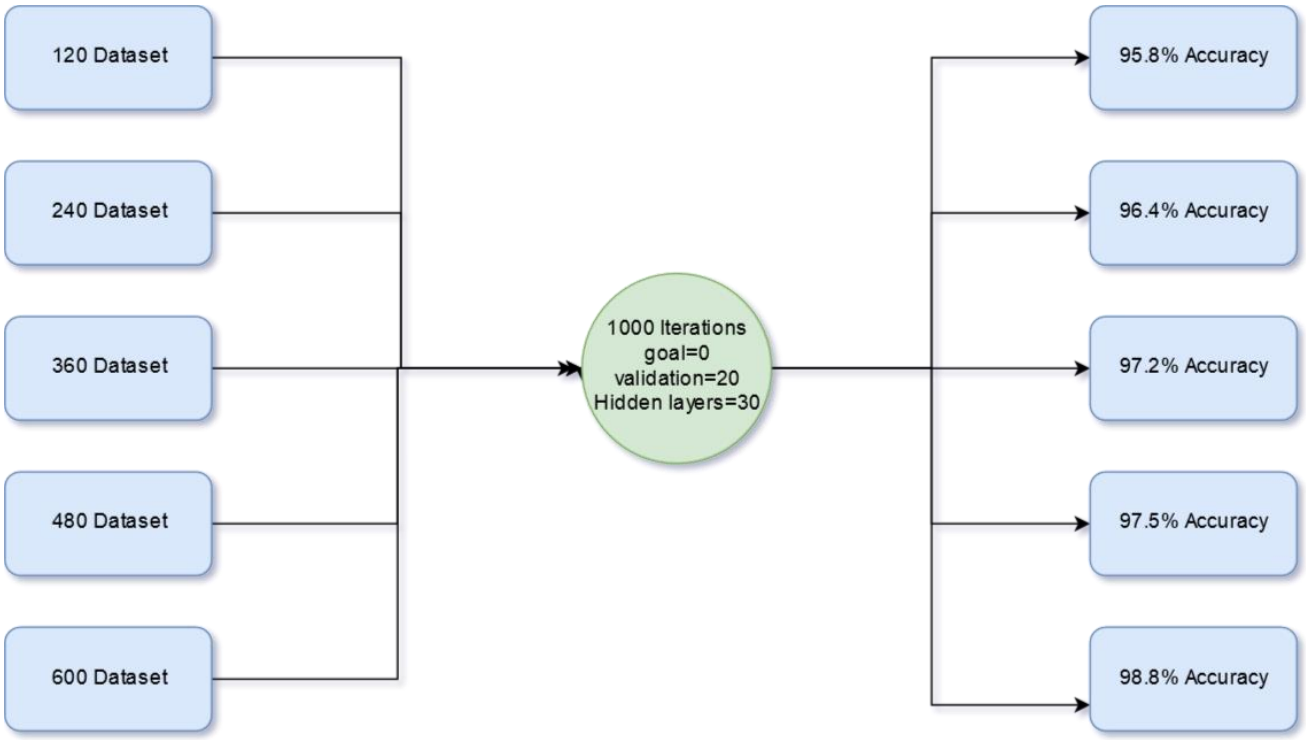


Analysis:

By increasing hidden layers with default values, a minor though significant gain seen in the accuracy of the neural network, but if we look at fourth dataset, there is an abnormal behavior in the result. Though tested thrice on these settings for simulation, result is abnormal for the fourth set of values, else the results favor hypothesis.

THIRD TRAINING:

Now, by increasing hidden layers a gain is seen in the accuracy of the neural network, but there is an abnormal activity with fourth set of dataset, that is around 75 percent of the dataset. Now, with modified validation value to 20, the neural network was trained again, and the results are as follows:

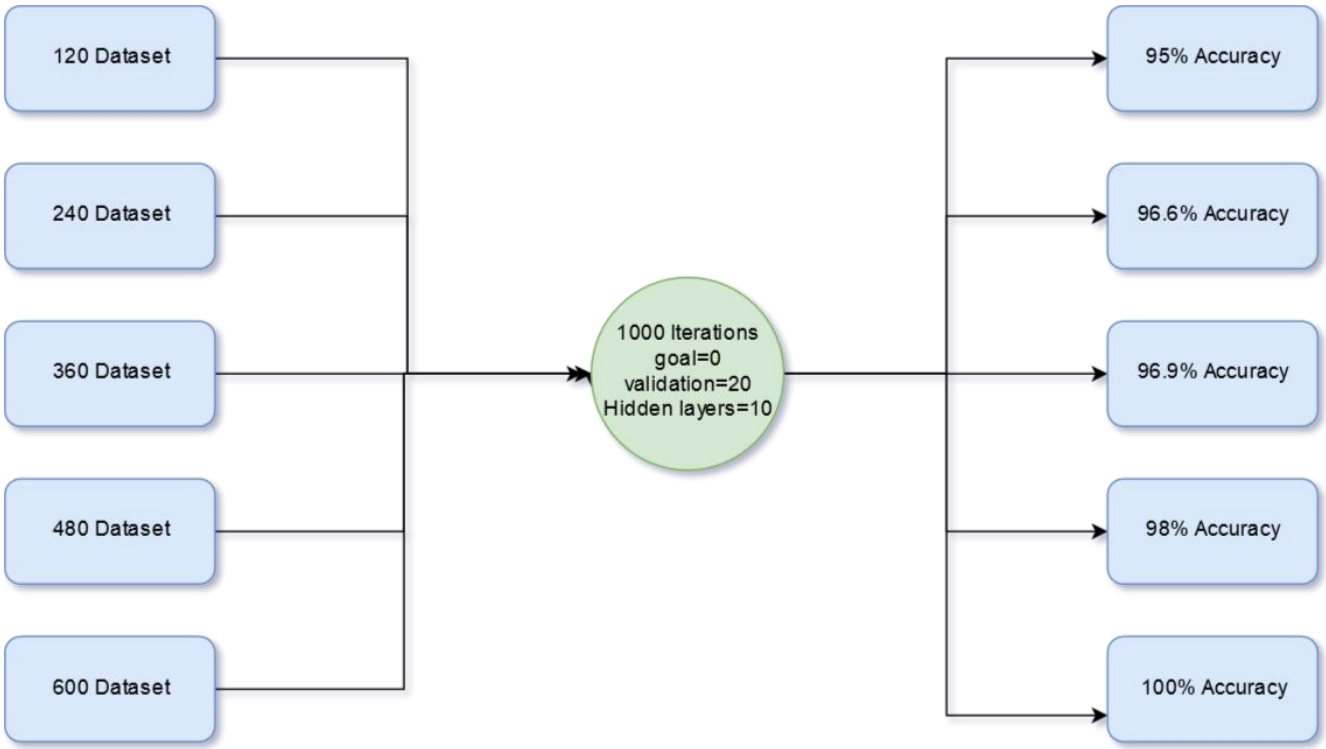


Analysis:

Results generated by this configuration came in the favor of the hypothesis, but there is a significance drop on the accuracy of the neural network as compared to the previous configuration with exception of fourth dataset.

FOURTH TESTING:

As our previous result generated the results in the favor of hypothesis, one test with the same configuration but with lesser number of hidden layers that are 10 is done, the results are as follows:

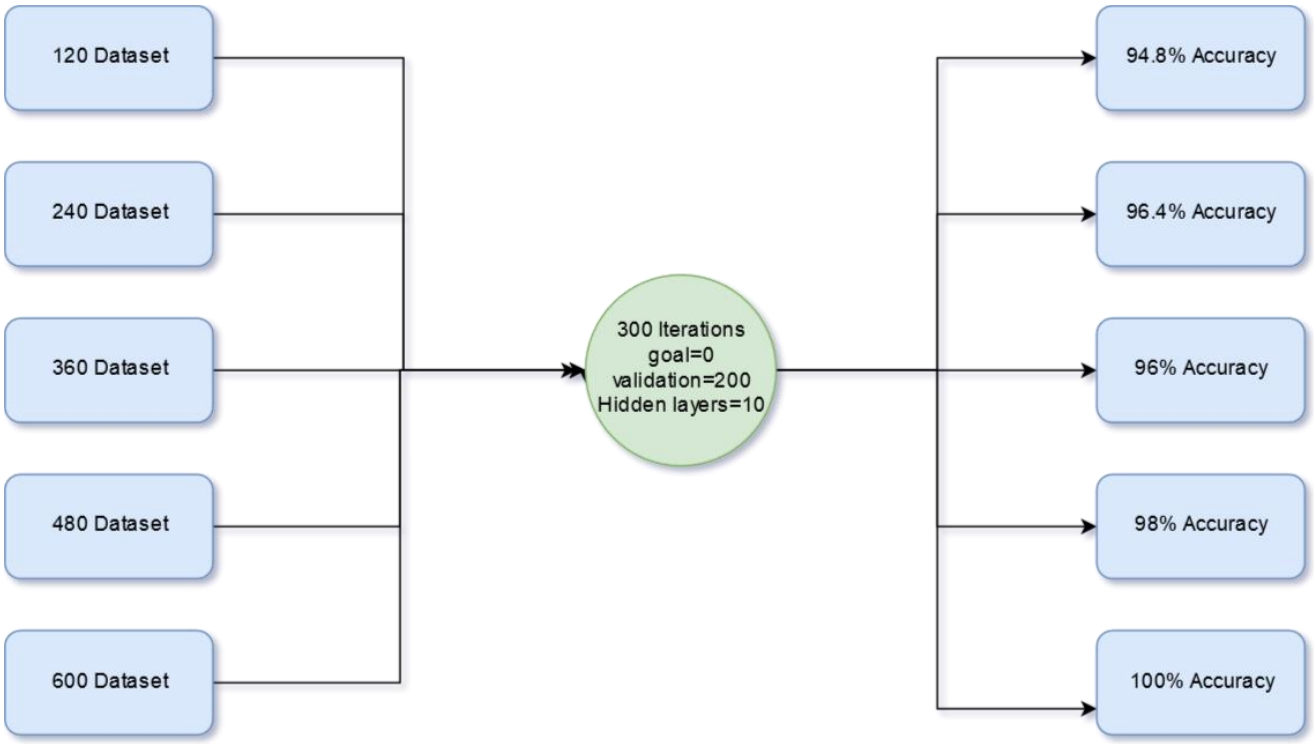


Analysis:

By decreasing hidden layers and keeping validation check to 20, the results generated have a drop in accuracy of all datasets, except the biggest one, that touched the accuracy to hundred. Which means that lesser number of hidden layers with a greater validation check works better for the greater number of training data while increasing hidden layers have shown a better result when it comes to lesser number of values to be trained upon.

FIFTH TESTING:

By keeping previous settings but changing the value of iterations to 300 so that a test on lesser number of iterations was done. Results are as follows:



Analysis:

The results are again in the favor of hypothesis, but even after five different methods and settings, the result being generated is still ambiguous when it accuracy, though a blurred hypothesis about validation and its relation with different hidden layers can be generated.

CONCLUSION:

Tests conducted are all in the favor of given hypothesis with a variation in second test where all values are kept default and the number of hidden layers is 30. By looking at other results, it can be said that hidden layers and validation check works hand to hand but with inverse reaction to each other, it cannot be said that this hypothesis is a law because with different type of dataset, there may come a different result and a different hypothesis apply, but in this case, it can be said that this hypothesis is true, though a different hypothesis can also be true as results change with different kind of settings and different kind d results can be seen. In results being generated, many attributes are being neglected, such as time efficiency which is not a main concern until it reaches infinity, because sometimes neural networks can take days to train. Though, iterations and validation check and goal, all effects time factors.

References

Breastcancer.org, November 9, 2016. *Types of Breast Cancer.* [Online]
Available at: http://www.breastcancer.org/symptoms/types

Cheprasov, A., n.d. *Benign vs. malignent: Definition, Characteristics & Differences.* [Online]
Available at: http://study.com/academy/lesson/benign-vs-malignant-definition-characteristics-differences.html

O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).