Breast cancer classification using Neural Network Approach

By

Ameer Abdullah

Namal College Mianwali



Abstract:

Breast cancer is a common human disease after skin cancer. It happens both in man and woman but it is far more common in women. The report contains an implementation of Neural Networks to diagnose the breast cancer. The data is provided in which we can identify positive and negative cases of Breast Cancer. This data is the base of the report as it will provide means to diagnose the disease by system developed at the end of this report.

Date: 3rd December 2017

1. **Introduction:**

Breast cancer is a very common disease occurring in women. But it also occurs in men. It happens when the cells of breasts grow abnormally or rapidly. This problem is a very serious problem as the disease can lead to death if not handled properly [1]. According to a survey by WHO the number of cases are increasing in a very rapid fashion and could touch 12 million by (2030 George *et al.*, 2012). Being such a fatal disease there is a need to establish a system to identify the breast cancer properly.

Now a day many health organizations are keeping records of this disease from different cases. This is to support and provide enough means for future research and diagnosing techniques in inter-disciplinary sciences. Because in these days it is very common to use data mining techniques for classification and many other problems (Raad, Kalkech & Ayache, no date).

In this report a method is prosed to diagnose the disease with the help of symptoms using Wisconsin Breast Cancer Database (UCI Machine Learning Repository: Breast Cancer Original Dataset, no date). Neural Networks are being used to diagnose the problem given that all symptoms are scaled from 1 to 10. This model can be deployed in any hospital to identify the breast cancer.

1. **Background:**

**Breast Cancer Classification**: This classification is done between two main classes of breast cancer: Benign and Malignant. The former one is the non-cancerous type of tumor in which the tumor doesn’t spread to the whole body. Instead it is contained by our immune system in the region it is produced and thus it can easily be cure. The latter one is the cancerous type of tumor in which tumor spread to other parts of body through blood. This can also be cured if diagnosed at early stage (Benign and Malignant Tumors: What is the Difference, no date). Classification is the technique in which we place anything in one of the many categories. There may exist two or more classes in which we classify the data. Here we have two classes, benign and malignant. We need to classify all the data from Wisconsin’s Breast Cancer’s database.

**Neural Networks:** In neural networks the machines go for pattern matching which is very intuitive for humans. For example, if we go to doctor for any sort of diagnosis, the doctor starts asking for basic questions. Then he tries to place our case to a certain case which is already diagnosed. Machines also go for similar kind of pattern. Here we are using neural networks and back propagation as a training algorithm. The algorithm calculates the error from targeted output and then propagates back to change the biases and weights (Raad, Kalakech and Ayache, no date). This is a very efficient and fast learning algorithm for pattern matching problems used commonly.

1. **Main Part**

The data has 699 instances in which there are 11 attributes. Following table contains the overview of the data and values it contains in each column:

|  |  |  |
| --- | --- | --- |
| # | 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 |

**Pre-Processing:** The data is needed to be refined as it contains some issues which may lead us to erroneous results.

1. There are 16 instances in the data in which 7th attribute contain missing values. These missing values must be dealt. One approach is to remove all the instances which contain missing values. In my case, I have placed the mean value of column 7 to all missing values in it. This mean value is calculated and it is 3.
2. In the class table 2 stands for benign and 4 stands for malignant. The machine may mislead to give the value 3. To avoid this, I have place 0 for benign and 1 for malignant to be clearly classified in either class.

**System Architecture:** In this process, MATLAB R2015a is used for developing our network. MATLAB contains a fully equipped toolbox of Neural Networks and multiple variations can be tried. But in our case we have used feedforward backpropagation neural network. The data is divided in training and testing data.

Training: The data is divided and tested multiple times. I have gone through multiple ratios of training and testing data to achieve the best results. For training ‘newff’ from MATLAB is used to create a network. There are multiple parameters in the ‘newff’ function to be set. Then it is trained using the ‘train’ function. There are also some other parameters for the network which can be changed and get the best of it. These are learning rate (newff.trainParam.lr), number of iterations (newff.trainParam.epochs), minimum error (newff.trainParam.goal) and max attempts of failing (newff.trainParam.max\_fail).

Testing: In this step, the trained network is tested on the testing data. For this, the network is given the testing data and it resulted in the output predicted by the network. The whole output is compared to the target output and number of correctly classified instances are counted. The accuracy is calculated through following equation:

Accuracy =

1. **Experimental Results and Analysis**

In this section, multiple variations of the networks are tried to get the best network to be used anywhere. The only thing which will be constant is the ratio of training and testing data. It will be 70 percent and 30 percent respectively.

1. **Effect of Neurons**: The number of neurons in the hidden layer of the network matters a lot. So deciding the number of neurons is important. The increasing number of neurons should increase the accuracy but to a certain limit. Because increasing too much neuron will cause a problem.

**Experiments**

|  |  |
| --- | --- |
| **Number of Neurons** | **Accuracy** |
| 1 | 99.4975 |
| 4 | 98.4925 |
| 9 | 99.4975 |
| 18 | 98.9950 |
| 36 | 99.4975 |
| 50 | 98.9950 |
| 100 | 99.4975 |
| 200 | 99.4975 |
| 1000 | 99.4975 |

In our case the number of neurons are not affecting as much as expected because our accuracy is changing very little bit even if I have increased the neurons to 1000.

1. **Effect of Multiple Layers of Neurons:** The number of layers of neurons matter a lot if the data is not consistent. As it helps a lot in creating the boundaries between multiple classes. It matters a lot when the boundaries are not clear. We will keep the number of neurons in each layer to 9 because the size of input vector is 9.

**Experiments**

|  |  |
| --- | --- |
| **Number of Layers** | **Accuracy** |
| 1 | 99.4975 |
| 2 | 100 |
| 3 | 97.4874 |
| 4 | 98.9950 |
| 5 | 99.4975 |

We can see that increasing the number of layers is not affecting the layers which mean that our data is very much accurate according to the classes. It doesn’t contain any noise cases or boundary ambiguity to mislead.

1. **Effect of Learning Rate:** Learning rate is an important factor towards making a classifier. Learning rate should not fall a certain limit but it should also not exceed too much. We will see the results in following experiments:

**Experiments**

|  |  |
| --- | --- |
| Learning Rate | Accuracy |
| 0.01 | 98.2249 |
| 0.001 | 98.8166 |
| 0.0001 | 99.4083 |
| 0.1 | 98.8166 |
| 0.5 | 98.9950 |
| 1 | 99.4083 |
| 5 | 100 |
| 10 | 99.4975 |

The learning rate is not affecting the accuracy too much. It means that initial biases and weights assigned in the training are much affective. So even if we increase or decrease the learning rate highly the accuracy doesn’t get below a certain limit.

1. **Effect of ratios of division:** In this experiment multiple variations of splitting are tried and results are achieved. If we decrease the number of training data the accuracy should decrease highly.

**Experiments**

|  |  |  |
| --- | --- | --- |
| Training Percentage | Testing Percentage | Accuracy |
| 10 | 90 | 92.3688 |
| 20 | 80 | 96.9589 |
| 30 | 70 | 96.3190 |
| 50 | 50 | 98.2808 |
| 70 | 30 | 92.5238 |
| 90 | 10 | 100 |

It is seen clearly that if we decrease the number in training data our accuracy decreases. Hence the training data must not be less than 50 percent to achieve satisfactory results but we can see that accuracy doesn’t fall below 90 percent, which is a clear proof that our data doesn’t contain noisy cases and is very accurate for using anywhere.

1. **Conclusion**

According to my experiments the important factor in neural network is performance. Because what so ever other variables are increased or decreased, the initial weight assigning and biases are good enough to predict the result accurately.

This system can be used fairly to get the predictions of Breast Cancer as it is trained and tested enough. Many variations are tried and every combination is giving more than 90 percent result which is far enough accuracy to be utilized anywhere.

Bibliography:

1. What is Breast Cancer? (no date). Available at:<https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html>
2. George, Y. M. *et al.* (2012) ‘Breast Fine Needle Tumor Classification using Neural Networks’, 9(5), pp. 247–256
3. Raad, A., Kalakech, A. and Ayache, M. (no date) ‘Breast Cancer Classification Using Neural Network Approach: Mlp and Rbf’, *Networks*, 7(8), p. 9.
4. Benign and Malignant Tumors: What is the Difference? (no date). Available at: <https://thetruthaboutcancer.com/benign-malignant-tumors-difference/>
5. UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set (no date). Available at: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29