BREAST CANCER DIAGNOSIS USING NEURAL NETWORKS

Submitted to

Sir Junaid Akthar

Submitted by

Aqsa Samreen

14031250

Contents

Introduction

20 October is the awareness day for breast cancer worldwide But this battle against this disease is far from finished. Breast cancer is a common problem in both developed and under developed countries. It is evaluated that more than 5,08,000 women died in 2011 due to breast cancer (Global Health Estimates, WHO 2013).Developed countries faces almost 50% breast cancer cases and 58% deaths occur due to this disease in under developed countries (GLOBOCAN 2008)[1].Artificial intelligence is now widely used in many real world problems including breast cancer classifiers and systems. This report explains that how neural network classify the dataset of breast cancer to identify either patient is facing cancer or not. This report covers all the steps from basic structure of the network to the analysis of the results.

Background

Classification is one of the most important and essential task in machine learning and data mining. About a lot of research has been conducted to apply data mining and machine learning on different medical datasets to classify Breast Cancer. Many of them show good classification accuracy. Vikas Chaurasia and Saurabh Pal11 compare the performance criterion of supervised learning classifiers; such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, Decision trees (J48) and simple CART; to find the best classifier in breast cancer datasets. The experimental result shows that SVM-RBF kernel is more accurate than other classifiers; it scores accuracy of 96.84% in Wisconsin Breast Cancer (original) datasets. Djebbari et al.12consider the effect of ensemble of machine learning techniques to predict the survival time in breast cancer. Their technique shows better accuracy on their breast cancer data set comparing to previous results. S. Aruna and L. V Nandakishore13, compare the performance of C4.5, Naïve Bayes, Support Vector Machine (SVM) and K- Nearest Neighbor (K-NN) to find the best classifier in WBC. SVM proves to be the most accurate classifier with accuracy of 96.99%. Angeline Christobel. Y and Dr. Sivaprakasam14, achieve accuracy of 69.23% using decision tree classifier (CART) in breast cancer datasets. A. Pradesh15 compares the accuracy of data mining algorithms SVM, IBK, BF Tree. The performance of SMO shows a higher value compared with other classifiers. T.Joachims16 reaches accuracy of 95.06% with neuron fuzzy techniques when using Wisconsin Breast Cancer (original) datasets. In this study, a hybrid method is proposed to enhance the classification accuracy of Wisconsin Breast Cancer (original) datasets (95.96) with 10 fold cross validation. Liu Ya-Qin's, W. Cheng, and Z. Lu17 experimented on breast cancer data using C5 algorithm with bagging; by generating additional data for training from the original set using combinations with repetitions to produce multisets of the same size as you're the original data;

to predict breast cancer survivability. Delen et al. Lu18 take 202,932 breast cancer patients records , which then pre-classified into two groups of “survived” (93,273) and “not survived” (109,659). The results of predicting the survivability were in the range of 93% accuracy.[5]

Iranpour, et al. discussed the application of Support Vector Machines (SVM), Radial Basis Function (RBF) networks for breast cancer detection and obtained an accuracy of 98.1% which is compared favorably to accuracies obtained in other studies like linear SVM classifier (94%), fuzzy classifiers (95.8%), and edited nearest neighbor with pure filtering (95.6%). [1].

Bat optimization algorithm is proposed for the selection of appropriate features from the WDBC dataset. This is one of the optimization algorithms that optimize the features of the breast cancer dataset to increase the accuracy of final results.[2]

Preprocessing

Wisconsin Diagnostic Breast Cancer (WBDC) dataset contains 699 breast cancer cases with their diagnosis as benign or malignant. The dataset has been divided into two datasets one for the input, which is passed to feed forward neural network and other one, is output dataset. Total 11 columns are in data set. The first column has id numbers. Input is followed from the column 2 to column 10 in the dataset. Last column is output column having two type of values i-e 2 or 4, 2 is for benign and 4 for malignant.

Column 6 of the input has 16 missing values are represented as ‘?’ in the dataset. To solve this problem ‘?’ is replaced with the median of that column which is ‘1’.

Matlab is used as platform for creating and training a neural network. Newff is used as function for creating the neural network. Preprocessing is completed.

Network Structure

Feed forward network is created using newff function as describe before. Input and output data pass to newff with 20 hidden layers in network. Tansig, trainr, learngd and mse used as activation functions are set with default values. In the next step net made by newff passed to train function. Use the activation functions of training to set the number of generations to 100 and goal to 0.01. After training use function y=net (input) which return the y as output. Set the threshold for converting the output to 2 and 4.After that find accuracy using y=minus (a, b) where a is output dataset and b is dataset obtained by function net(input).

Experimental Results and Analysis

For the purpose of experiment, separate the data into different number of proportions like 80 by 20, 60 by 40, 50 by 50, 40 by 60 and 20 by 80. These proportions of input and output data used for the different experiments and for the results to the above described network.

Hypothesis 1

Hypothesis:

Reducing training data will reduce the accuracy of the network, increase the error, and vice versa.

Effects of data distribution to accuracy and error:

Training Data (%)	Testing Data (%)	Accuracy (%)	Error
80	20	98.5714	1.4286
60	40	97.4910	2.509
40	60	96.8974	3.1026
20	80	95.5277	4.4723

Analysis:

Above table shows that as we are reducing the data for training, accuracy is also reducing and error is increasing which means that the network is trained on large dataset than mean square error will be low. Network will be able to detect more data precisely.

Result:

After the analysis, Hence proved that first hypothesis is right.

Hypothesis 2

Hypothesis:

If the percentage of input data for training and output data for testing is same, then its accuracy will be less than if the percentage of input data for training is greater than output data for testing.

Effects of data distribution to accuracy:

Training input data (%)	Testing output Data (%)	Accuracy (%)
80	80	96.2433
80	20	98.5214
60	60	95.2267
60	40	97.4910

Analysis

Above table shows the result of different experiments to prove the hypothesis. If the training data is 80% and testing data is also 80% then its accuracy is less than when the training data is 80% and testing data is 20%. Because the network is trained on large data and it is easy for the network to detect the right values for a small number of testing data as compare to 80% dataset for testing which is equal to the number of trained data set. Therefore, accuracy of detecting the right values for less testing data is greater than more testing data.

Result

Results of the accuracy show that the hypothesis is correct for this system.

Hypothesis 3

Hypothesis

Increase in learning rate will lead to the decrease in accuracy.

Effects of Learning rate on accuracy:

Learning Rate	Accuracy (%)	Error
0.09	97.2818	2.7182
0.08	96.8526	3.1474
0.07	96.4235	3.5765
0.06	96.4235	3.5765
0.05	95.8512	4.1488
0.04	94.9928	5.0072
0.03	96.8526	3.1474
0.02	95.9943	4.0057

Analysis

From the above table as we are increasing the learning rate accuracy is showing an increasing and decreasing pattern. Due to change in accuracy error rate is also fluctuating. Increasing the learning rate accuracy is not changing by following a sequence. It means that learning rate is not affecting the accuracy directly.

Result

Above analysis shows that our hypothesis is not correct.

Refined Hypothesis

Increasing the learning rate accuracy does not follow a decreasing sequence but it fluctuates.

References:

[1]http://www.who.int/cancer/detection/breastcancer/en/index1.html

[2]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5555532/

[3] Iranpour M, Almassi S and Analoui M, "Breast Cancer Detection from fna using SVM and RBF Classifier", In 1st Joint Congress on Fuzzy and Intelligent Systems, 2007.

[4] K. Menaka and S. Karpagavalli, "Breast Cancer Classification using Support Vector Machine and Genetic Programming," Int. J. Innov. Res. Comput. Commun. Eng., pp. 1410–1417, 2013.

[5] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," Procedia Comput. Sci., vol. 83, no. Fams, pp. 1064–1069, 2016.