Neoteric Breast Cancer through Machine Learning Algorithms

Atiya Khan
M.Tech (CSE)
Department of Computer Science and Engineering
Jamia Hamdard, New Delhi

Tabrez Nafis
Assistant Professor
Department of Computer Science and Engineering
Jamia Hamdard, New Delhi

ABSTRACT

Various machine learning algorithms are being used in healthcare today for better analysis and prediction of disease. Machine learning has enabled us to extract knowledge from the huge and complex healthcare data. This paper discusses about the breast cancer recurrence data. Breast cancer is a very common type of cancer in women. It usually occurs at the age of 50 or above. But these days it can be seen in younger women. Some women are the higher risk as compared to other but that totally depends on the personal medical history, hereditary or changes in gene. It is very important to diagnose the breast cancer in early stage. This will not only help in treating the cancer but also prevent it from reoccurring. This calls for the need of real time streaming of data and analysis of disease in real time. In this paper we have compared three algorithms J48 decision tree algorithm, naïve bayes algorithm and multilayer perceptron (MLP) algorithm. We have compared them on the basis of accuracy, sensitivity, specificity and Area under ROC curve on two type of dataset, training set and testing set.

General Terms

Algorithms, Healthcare, sensitivity, Specificity, Accuracy, ROC.

Keywords

Machine learning, J48 decision, Naïve Bays, Multilayer Perceptron (MLP), Breast Cancer.

1. INTRODUCTION

Now a day’s computers are being used in healthcare industries in a much more sophisticated way. Huge amount of data are analyzed by computers in research laboratory which makes the lab scientists free from the task of analyzing the collected data [9]. Many hospitals have moved from paper based systems to electronic systems. As a result of which large amount of data are gathered by these electronic systems on a day to day basis. These collected data contain a lot of hidden valuable information that can directly increase the efficiency and even cut down the expenditures of any organization [3]. This information can help doctors to provide better healthcare services to patients. Even the patients can avail the advantage of low healthcare costs. Healthcare industries are generating large amount of data in the form of electronic medical record (EMR), patient records, hospital resources, billing systems, medical devices etc [6]. Since the volume of data is increasing, various new and efficient ways to extract knowledge and interact with data are emerging. Machine learning is a method of data analysis and developing models [4].

Machine learning is a branch of computer science that provides machine the ability to learn from previous experiences without human intervention and improve their performance. Machine learning algorithm can be supervised, unsupervised, semi supervised and reinforcement machine learning algorithm. Machine learning plays a very important role in various applications such as data mining, NLP, expert system, image recognition, prediction. Machine learning has a very wide spread impact in healthcare [1]. Machine learning can help in developing tools that can be used by physicians for detecting the disease in early stage and hence increasing the survival rate of the patients [5]. Machine learning can help in reducing the cost of health care and can also improve patient doctor relationship. It can provide plenty of solutions to healthcare related issues such as, enabling doctors to perform personalized treatment for patients; also patients can determine when they can schedule their appointments [4]. Machine learning has helped us in predicting various deadly diseases such as heart diseases, stroke, various types of cancer, diabetes, genetically inherited diseases and many more.

There are various types of machine learning algorithm that can be efficiently used in healthcare industries. The various types of machine learning algorithms are decision tree, neural network, support vector machine (SVM), naïve bayes, genetic algorithm, fuzzy sets etc.

Here we will be discussing about the breast cancer and analyzing breast cancer data through various machine learning algorithms. We know that breast cancers are very common these days. After skin cancer, breast cancer is the most common type of cancer in women. It is the second leading cause of death among women. Even men can also suffer from breast cancer but that is not very frequent [7]. Less than 1% of breast cancer occurs in men. Some women are at the higher risk of breast cancer than others because of their personal medical history, hereditary or some changes in their gene.

Breast cancer symptoms includes

- Formation of a lump in breast or armpit.
- Swelling in some part of breast
- Irritation in breast skin.
- Redness in nipple.
- Pain in nipple or any area of breast.
- Change in shape and size of breast.

Women who are 50 or more are at the higher risk of breast cancer. But now a day’s breast cancer is diagnosed in women at the age of 45 only. In 2017, it is expected to have around 252, 710 new diagnoses of breast cancer in women, out of which around 40,610 women are most likely to die from the disease. Breast cancer screening is also very important. We all know that screening will not prevent breast cancer but it can help to diagnose breast cancer at early stage so that it can be treated at early stage [7].

There are several ways of treating breast cancer and that totally depend on the type of breast cancer or the stage of breast cancer. Some common types of treatments include surgery, chemotherapy, hormonal therapy, biological therapy and radiation therapy. Some times in some cases even the combination of these therapies can be applied [7].

In this paper we will be comparing decision tree, naïve bayes and multilayer perceptron (MLP) algorithm in terms of sensitivity specificity and accuracy and also analyzing ROC curve and area under ROC curve (AUC) and we will try to find out which algorithm will be better for analysis. We will be using weka tool for this purpose. WEKA is data mining software that uses a collection of machine learning algorithms. These algorithms can be applied directly to the data or called from the Java code.

2. METHODOLOGY

J48 Decision Tree: J48 is an implementation of C4.5 algorithm in WEKA. This algorithm works by building some rules and then uses those rules for predicting class label of the data being tested. J48 algorithm is reliable and robust algorithm. It also has the ability to handle the missing data. This algorithm is also very easy to understand that is the reason why it is so popular.

Naïve Bayes: It is also a type of classification technique. It is good to use naïve bayes in case we have large set of data. Naive bayes works on the basis of baye's theorem which defines the relationship between conditional and unconditional probability. Below given equation describes the baye's theorem and also the relationship between probabilities.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Artificial Neural Network: Artificial Neural network is inspired from human nervous system. An ANN consists of large number of interconnected neurons. Neurons in human body consist of dendrites, axon, cell body, synapse etc. Neurons of ANN are composed in the same way. An Artificial Neural Network consists of three layers: Input layer, output layer and hidden layer. All these three layers are composed of neurons that are interconnected. Neurons in ANN are connected with links and these links are provided with some weight. Learning in neuron is achieved by adjusting these weights. Neural Network plays a very important role in the field of healthcare. It can help in diagnosing various diseases by assigning weight to each of the changes that have occurred in the body. It can even help in suggesting medication for that particular disease. It can also help in processing medical images [2]. We have used multilayer perceptron of Neural Network in this paper. In multilayer perceptron each node except for input node uses activation function. It uses back propagation for training purpose. The activation function of MLP is non linear which distinguishes it from linear perceptron [10].

Procedure

1. Collect datasets of breast cancer from any authentic repository. In this case we have taken data from UCI repository which is having two class label reoccurring and non-reoccurring dataset.

2. Apply data preprocessing in which any missing value is identified and removed from the dataset. Also perform type conversion of data in required.

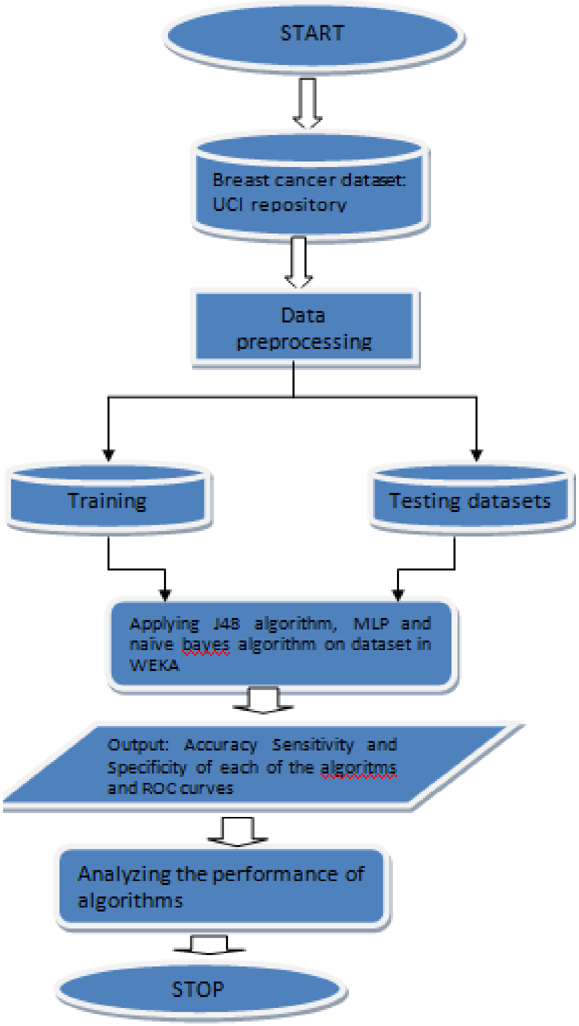
3. Divide the data set into two categories training set and testing set. This division must be done in such a way that that training set is two-third and testing set is one-third of the dataset.

4. Upload the preprocessed data into WEKA tool for analysis.

5. Apply J48, MLP and naïve bayes algorithm on dataset. First by doing evaluation on training set and then by evaluating on testing set.

6. Output of these algorithms tells us about the accuracy sensitivity and specificity of these algorithms.

7. This can be used in determining which algorithm will be better in diagnosing a patient.


START
Breast cancer dataset:
UCI repository
Data preprocessing
Training
Testing datasets
Applying J48 algorithm, MLP and naïve bayes algorithm on dataset in WEKA
Output: Accuracy Sensitivity and Specificity of each of the algoritms and ROC curves
Analyzing the performance of algorithms
STOP

3. RESULT AND DISCUSSION

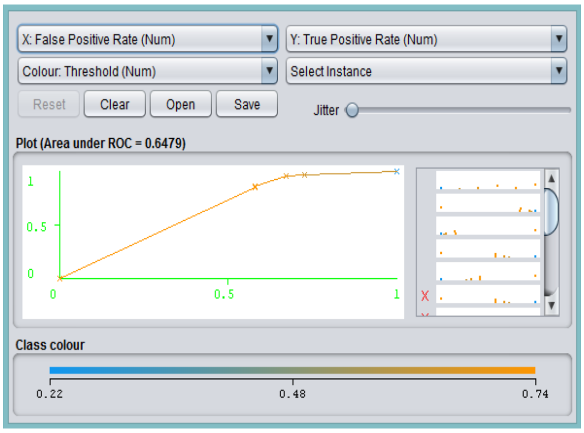
We have already mentioned in the previous section that we are using J48 decision tree, MLP and naïve bayes algorithm for comparison. These two algorithms are tested on the dataset of breast cancer that we acquired from UCI repository. The data set consisted of 286 data instances and 10 different attribute. These attributes include age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad, irradiat and class. We first divided our dataset into two parts training set and testing test. Training set comprises two-third of the dataset whereas testing set contained one-third of the whole dataset. Then we executed both algorithms first evaluating on the training set and then evaluating on supplied test set using 10 fold cross validation method. After executing both the algorithm we have compared them on the basis of accuracy, specificity, sensitivity and Area under ROC curve (AUC).

Table 1. The formula for accuracy sensitivity and specificity

Accuracy	**(TP+TN)/ (TP + TN + FP + FN)**
Specificity	TP/ (TP + FN)
Sensitivity	TN/(TN + FP)

Output for J48 algorithm: when evaluating on training set:

It took 0.03 second to execute. The correctly classified instances are 127 and incorrectly accuracy is 74%. Sensitivity and specificity are also calculated which are mention in table-II. We have also generated ROC curve for this which is shown in *figure-I*. We have also calculated area under ROC curve (AUC) and are shown in table-III

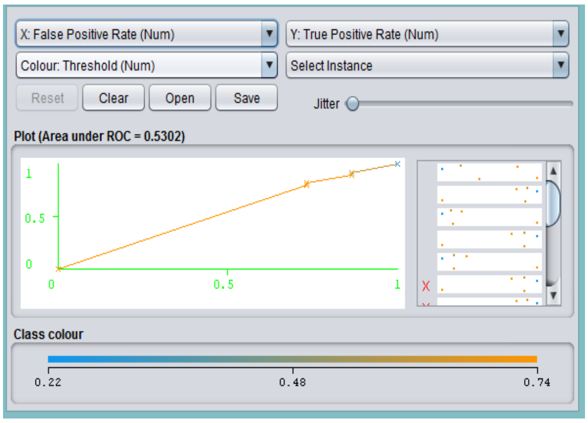


X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Plot (Area under ROC = 0.6479)

Fig 1

Output for J48 algorithm: when evaluating on supplied test set:

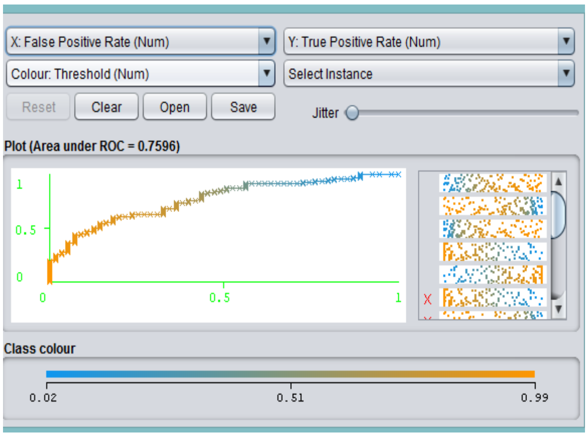
It took 0.01 second to execute. The correctly classified instances are 41 and incorrectly classified instances are 17. Which means the accuracy is 70%. Sensitivity and specificity are also calculated which are mention in table-II. We have also generated ROC curve for this which is shown in *figure-II*. We have also calculated area under ROC curve (AUC) are shown in table-III



X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Plot (Area under ROC = 0.5302)

Fig 2

Output for Naïve bayes algorithm: when evaluating on training set: It took 0 second to execute. The correctly classified instances are 125 and incorrectly classified instances are 46. Which means the accuracy is 73%. Sensitivity and specificity are also calculated which are mention in table-II. We have also generated ROC curve for this which is shown in *figure-III*. We have also calculated area under ROC curve (AUC) and are shown in table-III

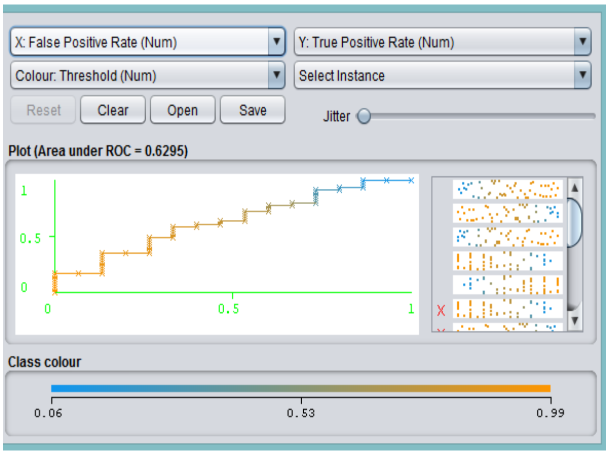


X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Plot (Area under ROC = 0.7596)

Fig 3

Output for Naïve bayes algorithm: when evaluating on supplied test set:

It took 0 second to execute. The correctly classified instances are 39 and incorrectly classified instances are 19. Which means the accuracy is 67%. Sensitivity and specificity are also calculated which are mention in table-II. We have also generated ROC curve for this which is shown in *figure-IV*. We have also calculated area under ROC curve (AUC) and are shown in table-III.

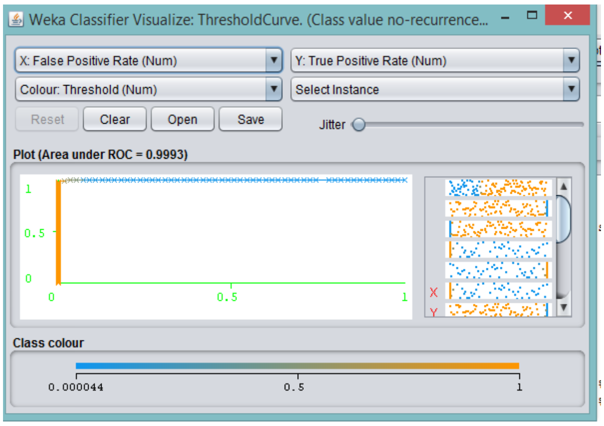


X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Plot (Area under ROC = 0.6295)

Fig 4

Output for MLP algorithm: when evaluating on training set:

It took 0.03 second to execute. The correctly classified instances are 168 and incorrectly classified instances are 3. Which means the accuracy is 98%. Sensitivity and specificity are also calculated which are mention in table-II. We have also generated ROC curve for this which is shown in *figure-V*. We have also calculated area under ROC curve (AUC) and are shown in table-III.



Weka Classifier Visualize: ThresholdCurve. (Class value no-recurrence...

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

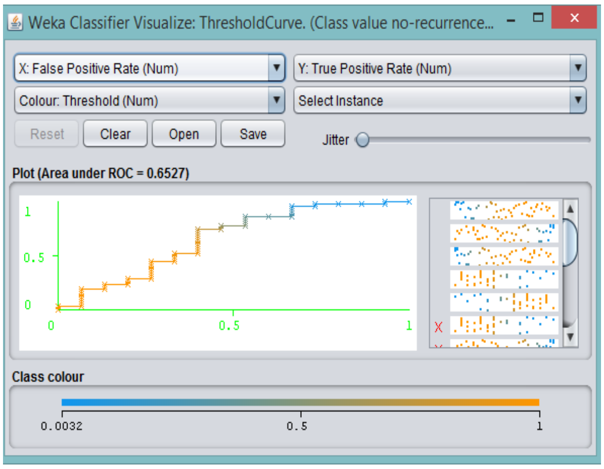
Select Instance

Plot (Area under ROC = 0.9993)

Fig 5

Output for MLP algorithm: when evaluating on supplied test set:

It took 0.02 seconds to execute. The correctly classified instances are 41 and incorrectly classified instances are 17. Which means the accuracy is 70%. Sensitivity and specificity are also calculated which are mention in table-II. We have also generated ROC curve for this which is shown in *figure-VI.* We have also calculated area under ROC curve (AUC) and are shown in table-III



Weka Classifier Visualize: ThresholdCurve. (Class value no-recurrence...

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.6527)

Fig 6

The given below table compares the accuracy, sensitivity and specificity analysis of both the algorithm when they are executed on training set and supplied test set.

Table 2. Accuracy Analysis of Algorithm

Algorithm	Confusion Matrix Using Training Set	Confusion Matrix Using Supplied Test Set	Accuracy (when evaluating on training set)	Accuracy (when evaluating on supplied test set)	Sensitivity (when evaluating on training set)	Sensitivity (when evaluating on supplied test set)	Specificity (when evaluating on training set)	Specificity (when evaluating on supplied test set)
J48	N R N 112 4 R 40 15	N R N 39 4 R 13 2	74%	70%	96.6%	90.7%	27.3%	13.3%
NAÏVE BAYES	N R N 98 18 R 18 27	N R N 34 9 R 10 5	73%	67%	84.5%	79.1%	49.1%	33.3%
MLP	N R N 115 1 R 2 53	N R N 33 10 R 7 8	98%	70.68%	99.1%	76.7%	96.4%	53.3%

The given below table compares the Area under ROC curve (AUC) of both the algorithm when they are executed on training set and supplied test set.

Table 3. AUC Analysis of Algorithms

Algorithm	Area under ROC curve (AUC) when evaluating on training set	Area under ROC curve (AUC) when evaluating on supplied test set
J48	64%	53%
NAÏVE BAYES	75%	62%
MLP	99%	65.2%

4. CONCLUSION

In this paper we have discussed the breast cancer disease and importance of detection of breast cancer in early stage. After implementing the algorithms J48 decision tree, naïve bayes and Multilayer perceptron (MLP) algorithm and evaluating on training set and testing set we have compared the accuracy sensitivity and specificity of both the algorithm. We found that accuracy of J48 is 74% on training data and 70% on supplied test set, the accuracy of naïve bayes is 73% on training set and 67% on supplied test set and accuracy of MLP is 98% on training set and 70% on supplied test set. Therefore the accuracy of MLP is very much better than J48 and naïve bayes whereas J48 and naïve bayes showed almost similar accuracy. We have also taken into account the Area under ROC curve. ROC is a visualization tool that will tell you that whether your classifier is appropriate or not. It is a cost sensitive method analysis. We also found that area under ROC curve for J48 is 64% on training set and 53% on supplied test set, Area under ROC curve for naïve bayes is 75% on training test set and 62% on supplied test set and area under ROC curve for MLP is 99% on training set and 65% in supplied test set. Still the MLP seems to be better. So we find that MLP shows a very good percentage of accuracy so it should be used for analyzing breast cancer for better evaluation result. Further the more recent and real time data can be used to predict the recurrence of breast cancer disease in future.

5. REFERENCES

[1] Expert System: http://www.expertsystem.com/machine-learning-definition/, 2018.

[2] Akshay Raul, Atharva Patil, Prem Raheja, Rupali Sawant "knowledge discovery, analysis and prediction in healthcare using data mining and analysis" 10, 2106.

[3] Mohammad Hossein Tekieh, Bijan Raahemi "Importance of Data Mining in Healthcare: A Survey" 2015.

[4] Rohan Bhardwaj, Ankita R. Nambiar, Debojyoti Dutta "A Study of Machine Learning in Healthcare" IEEE 41st Annual Computer Software and Applications Conference, 2017.

[5] Dana Bazazeh, Raed Shubair "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis" 978-1-5090-5306-3/16/$31.00 c2016 IEEE.

[6] Tania Cerquitelli, Elena Baralis, Lia Morra and Silvia Chiusano "Data mining for better healthcare: A path towards automated data analysis?" 2016.

[7] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2018.

[8] D. Tomar and S. Agarwal, 'A survey on Data Mining approaches for Healthcare', *Int. J. Bio-Sci. Bio-Technol.*, vol. 5, no. 5, pp. 241–266, 2013.

[9] Computer Science Degree Hub https://www.computersciencedegreehub.com/faq/role-healthcare-industry-computer-programmers/ 2018.

[10] Wikipedia: https://en.wikipedia.org/wiki/Multilayer_perceptron, 2018.