JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Breast Cancer Classification Using Artificial Neural Networks

V. Ambikavathi[1*], P. Arumugam[2], P. Jose[3]

[1,2]Dept. of Statistics, Manonmaniam Sundaranar University, Tirunelveli, India
[3]Dept. of Computer Science & Engineering, Saveetha Institute of Medical and Technical Science, Chennai, India

Corresponding Author: ambikavathi05@yahoo.com Tel.: 94435 54217

DOI: https://doi.org/10.26438/ijcse/v7i4.964968 | Available online at: www.ijcseonline.org

Accepted: 21/Apr/2019, Published: 30/Apr/2019

Abstract— Breast cancer is a fatal disease causing high mortality in women. By applying data mining techniques people can work on the extraction of hidden, historical and previously unknown large databases. The development of the technique have promised towards intelligent component in medical decision support systems. Here efficient information have been mined from the machine learning. ANN has been widely used in breast cancer diagnosis. In the proposed system the desired output were chosen and applied to ANN for preprocessing, classification and so on. The breast cancer data set from UCI data sets will be used to demonstrate different activities.

Keywords— ArtificialNeuralNetwork,ANN,DataMining,BreastCancer

I. Introduction

Brain tumor is an abnormal growth of tissue in the brain or central spine that can hamper proper brain function. The masses grow rapidly in an uncontrolled way. A primary brain tumor starts from cells in the brain. Secondary tumors are mainly generated from another part of the body. Secondary brain tumors are actually composed of cancer cells from somewhere else in the body that have been metastasized, or spread already to the brain. Cancer can be diagnosed by classifying tumors in two different types such as malignant and benign. Benign tumors represent an unnatural outgrowth but rarely lead to a patient’s death; yet, some types of benign tumors, too, can increase the possibility of developing cancer. On the other hand, malignant tumors are more serious and their timely diagnosis contributes to a successful treatment. As a result, predication and diagnosis of cancer can boost the chances of treatment, decreasing the usually high costs of medical procedures for such patients. Generally, there are two types of BC that are in situ and invasive. In situ starts in the milk duct and does not spread to other organs even if it grows. Invasive breast cancer on the contrary, is very aggressive and spreads to other nearby organs, and destroys them as well. It is very important to detect the cancerous cell before it spreads to other organs; thus, the survival rate for patient will increase to more than 97%. Htet Thazin Tike Thein[13] describes a to create an effective tool for building neural models to help us making a proper classification of various classes of breast cancer. Using this model, an automated classification of various types of breast cancer was performed by avoiding the question of the expert concerning the recognition of cancer required, improving the identification of breast cancer classification. Janghel[15] proposed the ability of the different neural networks to classify the applied input into either of the two classes configurations It saw that 95.82 had a testing accuracy of LVQ, 74.48 had a testing accuracy of CL and 51.88 had a testing accuracy of MLP.

II. Related Work

Bray [1] describes a Cancer, medically defined as a malignant neoplasm, is a board group of disease involving unregulated cell growth. Smith [6] describes an automated, fast and robust method for segmenting MR head scans into brain and non-brain parts. It is very robust and accurate and has been tested on thousands of data sets from a wide variety of scanners and taken with a wide variety of MR sequences.Somasundaram [7] focuses on two unsupervised and knowledge based methods to extract brain parts automatically using region labelling and morphological operations. It performed to obtain the fine brain on the assumption that brain is the largest connected component [LCC].Amit Tate [8] describes a major classification techniques used for prediction of classes using supervised learning dataset. It presents a comparative study on the performance of various classifiers of data mining over high dimensional data. It observe that Random Forest algorithm performed well with respect to all the factors.Beant Kaur [10] describes the huge amounts of data generated for prediction of heart disease are too complex and voluminous to be processed and analyzed by traditional methods. An different survey papers in which one or more algorithms used in prediction of heart disease. By applying different algorithms the best results found by the neural networks that gives the 100% accuracy & decision tree gives 99.62 % accuracy of results in prediction of heart disease. The use of data mining techniques to identify a suitable treatment for heart disease patients has received less attention. Different classification algorithms produce different result on base of accuracy, training time, precision, recall.Venkatesan [11] proposed a classification algorithms predict the hidden information in the medical domain. Classification is used to classify the elements permitting to the features of the elements through the predefined set of classes. The performances of classification accuracy of J48, AD Tree, BF Tree and classification and regression trees (CART) algorithms using various accuracy measures like False Positive Rate, True Positive Rate, Recall, Precision, Receiver Operating Characteristic Area and F-measure. In the implementation process, it is considered only the numerical values of some attributes in the breast cancer data.

III. METHODOLOGY

For experiment analysis the data set breast cancer is selected from UCI data repository. This data sets consists of 2 different types of breast cancer they are namely no recurrence events, recurrence events. Table 1 shows data set consists of 9 attributes and 286 instances.

Table 1: Breast file

Class	Age	MP	TS	IN	NC	DM	Breast	BQ	IR
NRC	32	premeno	33	1	no	3	Left	Lefi-low	no
NRC	45	premeno	22	2	no	2	Right	Right- up	no
NRC	47	premeno	23	1	no	2	Left	Lefi-low	no
NRC	65	ge40	16	2	no	2	Right	Lefi-up	no
NRC	46	premeno	3	0	no	2	Right	Right-low	no
NRC	62	ge40	15	1	no	2	Left	Lefi-low	no
NRC	53	premeno	26	2	no	1	Left	Lefi-low	no
NRC	64	ge40	22	1	no	2	Left	Lefi-low	no
NRC	44	premeno	51	2	no	2	Left	Lefi-low	no
NRC	46	premeno	23	1	no	1	Right	Lefi-up	no
NRC	49	premeno	3	1	no	3	Left	Central	no

NRC-No Recurrence Events MP – MenoPause
DM - Deg-malig TS - Tumor Size
BQ - Breast quad IN - Inv nodes
IR - Irradit NC - Node Caps

Preprocessing:
This exercise illustrates some of the basic data preprocessing operations that can be performed. It is required to prove method of resolving such issues. Data cleaning is process of fill in missing values, smoothing the noisy data, identify or remove outliers, and resolve inconsistencies. Integration of multiple databases, data cubes, or files.

Classification:
It is a Data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (sub populations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.Table2 shows various comparison classifier algorithm results by various breast cancer data sets. It describes the Mean Absolute Error, Root Mean Square Error.

Table 2.Comparison of Classifier Algorithm & Results

Algorithm	Time taken to build the model(Secs)	Correctly classified instance (%)	InCorrectly classified instance (%)	Total Number of instance	Mean Absolute Error	Root Mean Error
ZeroR	0	63.63	36.36 %	11	0.5	0.51
OneR	0	81.81	18.18	11	0.18	0.42
J48	0.06	63.63	36.36 %	11	0.31	0.48
Random Forest	0.05	63.63	36.36	11	0.39	0.44
Decision Table	0.02	63.63	36.36 %	11	0.52	0.53
Bayes Net	0	72.72	27.27	11	0.26	0.34
Naïve Bayes	0	45.45	54.54	11	0.51	0.63
Logistic	0.01	72.72	27.27	11	0.27	0.52
Multilayer Perceptron	0.08	81.81	18.18	11	0.24	0.42

The main aim to analyze the classification algorithms performance for breast cancer data as per table1. They are analyzed using classification ZeroR, OneR, J48, Random Forest, Decision Table, Bayes Net, Naive Bayes, Logistic, Multilayer Perceptron algorithms. A comparative study of classification accuracy in ZeroR, OneR, J48, Random Forest, Decision Table, BayesNet, NaiveBayes, Logistic, Multilayer Perceptron is carried out in this work. The TP Rate FP Rate and precision analysis is also carried out. The various formula as used for the calculation of different measures are as follows. The following formula is used to calculate the proportion of the predicted positive cases, Precision P using TP = True Positive Rate and FP = False Positive Rate as,

$$Precision\ P = \frac{TP}{TP+FP} \tag{1}$$

It has been defined that Recall or Sensitivity or True Positive Rate (TPR) means the proportion of positive cases that were correctly identified. It will be computed as

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

Where FN =False Negative Rate

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

The above formula will calculate the accuracy (the proportion of the total number of predictions that were correct) with TN = True Negative. Sensitivity is the percentage of positive records classified correctly out of all positive records.

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (4)$$

Specificity is the percentage of positive records classified correctly out of all positive records.

$$Specificity = \frac{TN}{(TN+FP)} \quad (5)$$

ROC stands for Receiver Operating Characteristic.A graphical approach for displaying the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) of a classifier are given as follows

Table 3.Classification Result Of Testing Data

Algorithm	TP Rate	FP Rate	Precision	F-Measure	ROC Area
ZeroR	1.000	1.000	0.636	0.778	0.018
OneR	1.000	0.500	0.778	0.875	0.750
J48	1.000	1.000	0.636	0.778	0.750
Random Forest	1.000	1.000	0.636	0.778	0.821
Decision Table	0.857	0.000	1.000	0.923	0.857
Bayes Net	1.000	0.750	0.700	0.824	0.893
Naïve Bayes	0.714	1.000	0.556	0.625	0.214
Logistic	0.857	0.000	1.000	0.923	0.857
Multilayer Perceptron	0.857	0.250	0.857	0.857	0.857

Breast cancer data contains the tumors that represents the severity of the disease and the tumors are correctly classified from the training data set, the error rates and accuracy are calculated using classifiers. The confusion matrix helps us to find the various evaluation measures like accuracy, recall and precision, F-Measure and Receiver Operating Characteristic Area etc. Table 3 shows classification results of breast cancer testing data by using its various classification algorithm Logistic Regression and Multilayer Perceptron produce better result. Table 4 shows the classification results of breast cancer testing data accuracy by weighted average.

Table 4. Accuracy by weighted average

Algorithm	TP Rate	FP Rate	Precision	F-Measure	ROC Area
ZeroR	0.636	0.636	-----	-----	0.018
OneR	0.818	0.318	0.859	0.799	0.750
J48	0.636	0.636			0.750
Random Forest	0.636	0.636	-----	-----	0.821
Decision Table	0.909	0.052	0.927	0.911	0.857
Bayes Net	0.727	0.477	0.809	0.670	0.893
Naïve Bayes	0.455	0.740	0.354	0.398	0.214
Logistic	0.727	0.370	0.720	0.717	0.821
Multilayer Perceptron	0.818	0.211	0.818	0.818	0.857

Performance Evaluations

Mean Absolute Error

The mean absolute error (MAE) is a quantity used to measure predictions of the eventual outcomes. The mean absolute error is an average of the absolute errors. The mean absolute error is given by

$$MAE = \frac{1}{n} \sum_{j=1}^{n} |y_{j-} \hat{y}_j|$$

MAE – Mean Absolute Error

Where

$\{ y_j \}$is the actual observations time series

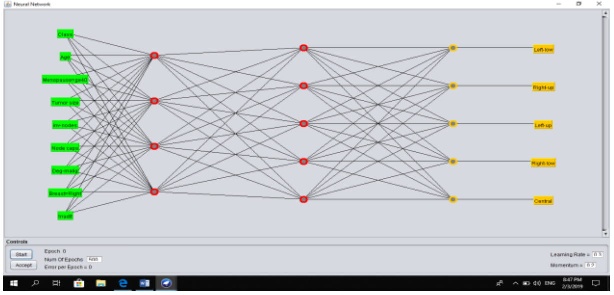
$\{ \hat{y}_j \}$ is the estimated or forecasted time series

Root Mean Squared Error

The difference between values predicted by a model and the values actually observed from the environment that is being modelled. It is the square root of the average of squared errors. The effect of each

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^{n} |y_{j-} \hat{y}_j|}$$

Figure 1 shows the multilayer perceptron network such as.input layer,hidden layer and output layer.



Epoch 0
Num Of Epochs 500
Error per Epoch = 0
Learning Rate = 0.3
Momentum = 0.2

Figure 1: Multilayer Perceptron

IV. Results and Discussion

Above section involves the study of nine algorithms and each of its technical applications introduced previously, and testing each one of them and classified on a set of breast cancer data related to medical information. The Accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. Incorrectly classified instances means the sum of FP and FN. The total number of correctly instances divided by total number of instances gives the accuracy. Correctly classified instances give the accuracy of the model. Correctly and incorrectly classified instances will be partitioned in numeric and percentage value. In the next part, mean absolute error, root mean squared error will be consider as parameters for evaluation. Table 5 shows the time taken to bulid the model.

Table 5: comparison of time taken to build the model (secs)

ZeroR	OneR	J48	Random Forest	Decision Table	Bayes Net	Naïve Bayes	Logistic	Multilayer Perceptron
0	0	0.06	0.05	0.02	0	0	0.01	0.05

V. Conclusion and Future Scope

This research work evaluate the performances in terms of classification accuracy of ZeroR, OneR, J48, Random Forest, Decision Table, BayesNet, Naive Bayes, Logistic, Multilayer Perceptron using various accuracy measures like FP rate, TP rate, Recall, Precision, ROC Area and F-measure. .The experimental results shows that the highest accuracy 81% is found in OneR and Multilayer Perceptron where as ZeroR, OneR, J48, Random Forest, Decision Table 63% and accuracy 72% is found in logistic and BayesNet algorithm. Rest of the algorithm take 45%.Based on the classification results of all the algorithms, the performance of OneR and Multilayer Perceptron algorithm is better than the other algorithms for the chosen data set

References

[1] Bray F,Jemal A,Grey N,Ferlay J,Forman D*,"Global cancer transitions according to the human development index(2008-2030)",*:A population case study.The Lancet oncology 2012;13(8):790-801

[2] Sulochana Wadhwani, A.K Wadhwani, Monika Saraswat, "*Classification of breast cancer using artificial neural network*", Current Research in Engineering, Science and Technology Journals, December 2009

[3] Dr. J. Abdul Jaleel, Sibi Salim, Aswin.R.B, *"Artificial Neural Network Based Detection of Skin cancer"*, International Journal of Advanced Research in Electrical, Electronics and Instrmentation Engineering, ISSN 2278 – 8875 Vol. 1, Issue 3, September 2012

[4] Miller KD, Siegel RL, Lin CC, Mariotto AB, Kramer JL, Rowland JH, JA. Cancer treatment and survivorship statistics. CA: A Cancer Journal for Clinicians. 2016;66(4):271-289

[5]G. Holmes; A. Donkin and I.H. Witten *"Weka: A machine learning workbench"*, Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia, 1994.

[6] Stephen M. Smith, *"Fast Robust Automated Brain Extraction,"* Human Brain Mapping, 17:143–155(2002).

[7]K. Somasundaram, T. Kalaiselvi, *"Automatic Brain Extraction Methods for T1 Magnetic Resonance Images Using Region Labeling and Morphological Operations,"* Computers in Biology and Medicine, 41 (2011), 716–725.

[8] Amit Tate, Bajrangsingh Rajpurohit *"Comparative Analysis of Classification Algorithms Used for Disease Prediction in Data Mining"*, International Journal of Engineering and Techniques, Volume 2 Issue 6, Nov –2016.

[9] S. a. E. N. Sharma, *"Brain Tumor Detection and Segmentation Using Artificial Neural Network Techniques"*, International Journal of Engineering Sciences & Research Technology, August 2014.

[10]Beant Kaur, Williamses Singh *"Review on heart disease prediction using data mining techniques," International Journal on recent and innovation trends in computer and communication" ,* Volume- 2, Issue-10,Page No(3003-3008), October2014.

[11]E. Venkatesan, T. Velmurugan *"Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification" Indian Journal of Science and Technology",* Vol 8(29), November 2015.

[12]Kariuki Paul Wahome *"Towards Effective Data Preprocessing for Classification Using WEKA",* International Journal of Science and Research (IJSR), 2016.

[13] Htet Thazin Tike Thein, Khin Mo Mo Tun,"*An Approach For Breast Cancer Diagnosis Classification Using Neural Network"*, Advanced Computing: An International Journal (ACIJ), Vol.6, No.1, January 2015.

[14] Shiv Shakti S, Sant A, Aharwal RP. *"An Overview on Data Mining Approach on Breast Cancer data",* International Journal of Advanced Computer Research. 2013; 3(13):256–62.

[15] R.R.Janghel, Anupam Shukla, Ritu Tiwari, Rahul Kala*,"Breast Cancer Diagnosis using Artificial Neural Network Model*",Research Gate,2010.

[16] Sanjay Agrawal et al. *"A Study on Fuzzy Clustering for Magnetic Resonance Brain Image Segmentation Using Soft Computing Approaches",* Applied Soft Computing, 24(2014), 522–533.

[17]D.Jude hemanthl et al. *"Effective Fuzzy Clustering Algorithm for Abnormal MR Brain Image Segmentation",* IEEE International Advance Computing Conference (IACC 2009), Patiala, India, 6-7 March 2009.

[18] Rajesh K, Anand S. *"Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm",* International Journal of Advanced Research in Computer and Communication Engineering. 2012 Apr; 1(2):72–7.

[19]Sandeep Chaplot et al. *"Classification of Magnetic Resonance Brain Images Using Wavelets as Input to Support Vector Machine and Neural Network",* Biomedical Signal Processing and Control, 1 (2006) 86–92.

[20] Dipali M. Joshi et al. *"Classification of Brain Cancer Using Artificial Neural Network",* 2nd International Conference on Electronic Computer Technology (ICECT 2010).

[21] Martin Fodslette Moller*, "Scaled Conjugate Gradient Algorithm for Fast Supervised Learning",* Neural Networks, Vol. 6, pp. 525-533, 1993.

[22] Amanpreet Singh, Narina Thakur, Aakanksha Sharma, *"A Review of Supervised Machine Learning Algorithms"*,2016 International Conference on Computing for Sustainable Global Development.

[23] Vrushali Y Kulkarni, Dr Pradeep K Sinha. *"Random Forest Classifiers: A Survey and Future Research Directions",* International Journal of Advanced Computing, ISSN: 2051-0845, Vol.36, Issue.

[24] Poomani N, Porkodi. R. *“A comparison of Data Mining classification algorithms using breast cancer microarray dataset: A study”,* International Journal for Scientific Research and Development. 2015; 2(12):543–7.

[25] Patel Pinky S, Raksha R. Patel, Ankita J. Patel, Maitri Joshi *“Review on Classification Algorithms in Data Mining”*, (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 1, January 2015).

[26] Pankaj Sapra, Rupinderpal Singh, Shivani Khurana, "*Brain Tumor Detection using Neural Network*", International Journal of Science and Modern Engineering, ISSN: 2319-6386, Volume-1, Issue-9, August 2013

[27] Ian H.Witten and Elbe Frank, (2005) "*Data mining Practical Machine Learning Tools and Techniques,"* Second Edition, San Fransisco.

[28[P.Arumugam,P.Jose,”*Efficient Tree Based Data Selection and Support Vector Machine Classification materials today*”Proceedings,Vol:5,Issue 1,2018,Pages 1679-1685.www.sciencedirect.com.

Authors Profile

V.Ambikavathi pursuing Part time Ph.D in Statistics from Manonmaniam Sundaranar University.She received M.E degree in Computer Science & Engineering from P.S.R Engineering college in 2011,India.Her research interest include data mining,Machine Learning and Artificial Neural Network.