IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-10, Oct 2018 E-ISSN: 2347-2693

Emperical Evaluation of Machine Learning algorithms for Breast Cancer Data Classification

S. Kumaravel[1]*, S. Ophilia Domanica Vithya[2]

[1,2]Department of Computer Science, A.V.V.M. Sri Pushpam College, Poondi, Thanjavur, India

Corresponding Author: skeyvel14@gmail.com

Available online at: www.ijcseonline.org

Accepted: 19/Oct/2018, Published: 31/Oct/2018

Abstract- Breast cancer (BC) is a deathly cancer disease which occurs mainly in women and the greater number of breast cancer patients leads to death according to global statistics. The early examination of Breast Cancer can augment the durability of patients, and it helps to improve the prompt medication to the patients. Machine learning plays an important role in health care and they are more powerful in classification and prediction process. There are various classification algorithms used based upon then data set. This work is the implementation of few classification algorithms such as Random Forest, K Nearest Neighbor, Navie Bayes, Support Vector Machine, and Artificial Neural Network for breast cancer data set. This paper is the comparative study of these algorithms using R tool. The goal of this paper is to analyze the accuracy of these algorithms. The implementation procedure reveal that the performance of any algorithm varies based on the data set attributes and characteristics.

Keywords- Machine Learning, Classification, Random Forest, K Nearest Neighbor, Navie Bayes, Support Vector Machine, Artificial Neural Network, R Tool**.**

I. INTRODUCTION

Machine learning is a part of (Artificial Intelligence). The main theme ofthe machine learning is to learn the design of data and fix the data into models. That make the user to utilize and understand once the data is fit as model. The next step is to test the models accuracy using the data and finally check the performance by using various data sets. Each data sets have various characteristics based on this model accuracy can be predicted and improved [1]. The machine learning technique have been widely implemented in the field of health care domain this helps to detect the deadly disease Breast Cancer by reading the past data of the patients and match them with their symptoms in the defined algorithms.

Breast cancer is cancer that develops in breast cells. Typically, the cancer forms in either the lobules or the ducts of the breast. Cancer can also occur in the fatty tissue or the fibrous connective tissue within your breast.

There are several risk factors that increase your chances of getting breast cancer. However, having any of these doesn't mean you will definitely develop the disease. Some risk factors can't be avoided, such as family history. Other risk factors, such as smoking, you can change. Risk factors for breast cancer include: Age, Drinking alcohol, Having dense breast tissue, Gender, Genes and Early menstruation.

The purpose of this work is predict the possibility of Breast cancer from the data set and initiate the necessary steps to recover the patient as earlier.

The rest of the paper is organized as follows, section II discuss about various techniques of classification process, section III deal with different machine learning approaches, section IV contain methodology of proposed work, section V explain the implementation in R environment and section VI deals with the results and discussion .

II. RELATED WORK

The author of this work Mengjie Yu, B.S [2] 2017 carried out a comparison of different classification techniques using R tool on large dataset. The data utilized in their research is the breast cancer data. The dataset is collected from Wisconsin Diagnostic Breast Cancer dataset was obtained from the UCI machine learning repository. The data set contains 357 cases of benign breast cancer and 212 cases of malignant breast cancer. In this the author implemented three different algorithms for classification. In that the highest average accuracy across the 10 fold cross-validation was achieved in KNN (accuracy = 0.97564) and SVM with linear kernel achieved the second highest accuracy (accuracy = 0.9754) when compared to others.

B.Nithya [3] et al. 2017 carried out a classification method using two different data set they are Iris data and Breast

Cancer Wisconsin (Diagnostic) data and they are predicted with three different algorithms such as Decision Tree - c5.0() , K Nearest Neighbor – (kNN), Naïve Bayes - (NB).Iris data set has 150 instances with four attributes. Breast Cancer Wisconsin (Diagnostic) data set has 569 instances with 32 attributes. In both the datasets 70% of instances have been taken as training data and the remaining 30% has been considered as test data. For Iris data Decision Tree - c5.0 () and Naïve Bayes gives the highest accuracy 95.56% when compared with K Nearest Neighbor 93.33%. In Breast Cancer data the K Nearest Neighbor 95.32 % compared to Decision Tree - c5.0 92.4% and Naïve Bayes 92.98%.

P.Dhivyapriya [4] 2017 done a prediction on breast cancer and Leukemia Cancer using two algorithms such as Navie Bayes and SVM. The breast cancer data set is downloaded from UCI Machine Learning Repository and the Leukemia dataset is downloaded from Bioinformatics Research group. Breast cancer data set contains 699 instances and 10 attributes. The leukemia data set contains 38 instances and 7129 attributes. In breast cancer prediction SVM gives the highest accuracy of 97% and Navie Bayes gives 96%. For leukemia cancer prediction both Navie Bayes and SVM produce 100% accuracy.

VikasChaurasia and Saurabh Pal [5] 2014 have done a prediction on Breast Cancer using WEKA tool the data set contains 683 rows and 10 columns and it is implemented in three algorithms they are Sequential Minimal Optimization, K Nearest Neighbours classifier,Best First tree. In this prediction SMO gives more accuracy of 96.19% and the IBK gives 95.90% accuracy, BFTree gives 95.46 % of accuracy.

Jahanvi Joshi and RinalDoshi [6] 2014 carried out a research work in WEKA tool using breast cancer data set it is downloaded from the UCI Machine Learning Repository. The main aim is to classify the Healthy and Sick Patients from the data set using different classifier rules. In that Filtered Classifier, Multiclass Classifier, J48, LMT gives more accurate result of 76% of healthy patients and 24% of sick patients.

J.S.Saleema [7] et al. 2014 have done a survey on Sampling Techniques such as Random Sampling, Stratified Sampling, Balanced Sampling and which are implemented in three classification algorithms Decision Tree, Navie Bayes, KNN and Breast cancer data set has been used. According to the classifications carried out in Stratified Sampling the Navie Bayes produce more accuracy of 95.52%. In Random Sampling also Navie Bayes gives more accuracy of 95.84%.At last in Balanced Sampling the Decision Tree gives more accuracy 98.40%.

TübaKiyanand Tülay Yildirim [8] et.al. 2004 have developed a Breast Cancer diagnosis using Statistical Neural Networks, In this work the author implemented the data set in three different Neural Network algorithms they are Probabilistic Neural Networks (PNN), Radial Basis Functions (RBF), Generalized Regression Neural Networks (GRNN) which are implemented in MATLAB 6.0. The data set is classified in two categories they are test data of 50% and training data 50%. According to the results produced GRNN gives highest accuracy of 98.8 %, second highest accuracy of 97.0 for PNN, RBF produce 96.18 %, MLP gives 95.74%.

Subrata Kumar Mandal [9] 2017 has discussed an analysis of Breast Cancer using different Data Mining algorithms the data set is collected from the UCI Machine Learning Repository the data set contains 569 instances and 32 attributes. The WEKA tool is used for classification the Logistic Regression algorithm produce highest accuracy of 97.90%, Decision Tree gives 96.50% and Navie Bayes gives 94.40% of accuracy.

Adnan Alam khan and Shariq Ahmed [10] et.al. 2015 done a research work based on performance comparative analysis between R and WEKA tool in this lung cancer data set has been used which contains 350 real data's. The data is loaded in the both tool and prediction carried out. In the result R tool produce more accuracy than the WEKA tool.

E. SathiyaPriya [11] et.al.2017 demonstrate a study on classification algorithms in which two different lung cancer data sets have been used. The first data set contains 2000 observations and 11 attributes. The second data set contains 309 observations and 16 attributes, these data sets are used in WEKA tool. The author implemented 8 different classification algorithms they are J48, KNN, REP Tree, Bayes Network, Neural Network, ID3, Navie Bayes, and SVM. In this research the ID3 algorithm produce the highest accuracy of 100%, the second accuracy is given by Neural Network of 96% when comparing other the algorithms.

ShamreenFathimaSaddique[12] et.al.2018 have done a survey on prediction of Lung Cancer using Classifier Models. The data's are implemented in R tool using Naïve Bayes Random Forest, KNN, Logistic Regression, SVM-Linear, and SVM-Radial. The highest accuracy is given by Navie Bayes 95.24%, Random Forest 93%, KNN 89.90% Logistic Regression 88.53%, SVM-Linear 88.04%, and SVM-Radial 85%.

Hlaudi Daniel Masethe, Mosima Anna Masethe[13] 2014 have done a discussion on Prediction of Heart Disease using WEKA tool. The data set is collected from medical practitioners in South Africa. The classification algorithms implemented are J48, REPTREE, SIMPLE CART , Bayes Net, Naïve Bayes. The highest accuracy is produced by three algorithms they are J48, REPTREE, and SIMPLE CART of 99.0741%.

T. Marikani[14] et.al. 2017 have done a prediction of Heart Disease using supervised Learning algorithms the data set is collected from the UCI Machine Learning Repository. Data set contains 303 instances and 14 attributes. The algorithms implemented are Classification tree, Naïve Bayes, KNN, Random Forest Classification, and SVM. The highest accuracy for the prediction is given by SVM 100%, the other algorithms are Classification tree 95.4%, Naïve Bayes 81.7%, KNN 75.7%, Random Forest Classification 96.3%.

Jothikumar and R.V. Sivabalan [15] et.al. 2016 have done a discussion about the analysis of classification algorithms for heart disease prediction the data set contains 294 instances and 14 attributes the algorithms used are Random Tree 75.14% 2 Naïve Bayes 79.25% 3 Decision Tree 78.24% 4 Random Forest 74.16%, The highest accuracy is Navie Bayes 79.25%.

III. DIFFERENT TYPES OF MACHINE LEARNING APPROACHES

Classification is the techniques mainly used in machine learning. The classification applications like sentiment analysis, ad targeting, spam detection, risk assessment, medical diagnosis and image classification[16]. Among the various classification algorithms, five major methods are depicted here.

A. Random Forest (RF)

Random forest algorithm cam also be called as decision forest and it can be used to build predictive model for both classification and regression problems. This algorithm will be used for multiple learning models to gain better predictive result. This decision tree algorithm is very simple to understand can be utilize like a top- down approach. In this approach origin node will create a dual splitting of nodes and it will providean estimated value based on internalnodes then it will leads to endmost node. In Random Forest the output will predicted to target class for each terminal node.
This algorithm can also be called as accurate learning algorithm for many data sets and it will produce more accuracy rate, it also works dynamically in large data sets. Random forest will be considered as output for some data sets with noisy classification and regression task.

B. K Nearest Neighbor (KNN)

KNN is part of the classification algorithm in Machine Learning. It deals with the Supervised Learning process and it helps to find the Pattern Recognition, Data Mining, and Intrusion Detection. The KNN algorithm is based on the unlabeled observation and declaring them to the class with the most similar labeled example. Characteristics and observations are collected from both training and test data set. The functionality of kNN classifier is shown in fig1. This approach is easy to use and helps to reduce noisy data but the distance can create irrelevant attributes and more expensive.

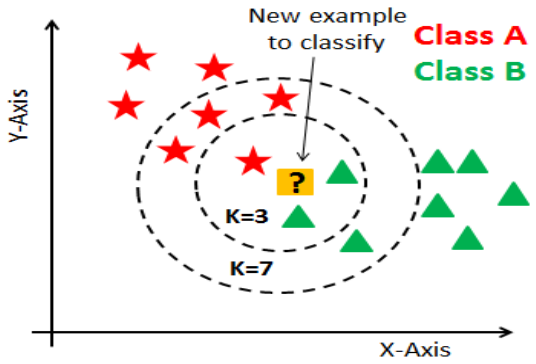
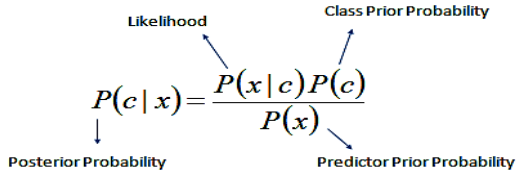


Fig1: kNN Classifier

C. Navie Bayes (NB)

The Navie Bayes classification is technique based on Bayesian Theorem and it is commonly used in Machine Learning. It is considered to be a probabilistic classifier that help to make classification using the maximum posteriori decision rule in a Bayesian setting as shown in the equ 1. This classification rule is well suited for Text Classification and Spam Detection.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$ Equ --- (1)

This classification algorithm is easy to implement and easy to calculate probabilistic predictions. Even though it is difficult to do regression.

D. Support Vector Machine (SVM)

Support Vector Machine is a powerful Machine Learning method deals with the Supervised Learning technique. The SVM creates a linearly separable Hyper lane through a data set in order to classify the data into two groups and it could be (2D),(3D),(4D+). The Hyper lane it's a line whose distance to nearest element of each tag is the largest is shown in fig 2.

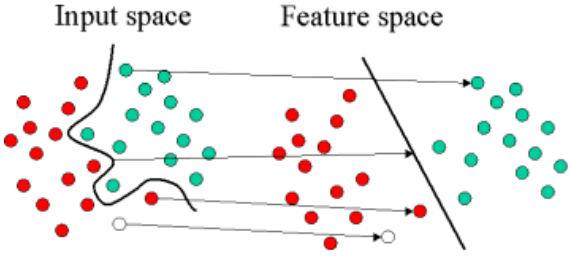


Fig 2: Support Vector Machine Classifier

SVM maintain memory efficiently and produce high dimensionality. But it is considered as Non probabilistic.

E. Artificial Neural Networks (ANN)

Neural Network is one of the Machine learning algorithm it works based on Human Neuron. ANN algorithm is an information processing technique it includes a large number of connected processing units that works together to process information. Neural Network works not only for classification it can also apply for regression. Neural Network consist of three layers: (i) Input Layer, (ii) Hidden layer and (iii) Output Layer as shown in fig 3.

The input layer used to feed the fresh information into network, in hidden layer the actual activity is done through system of weighted connection and there may be one or more hidden layer and the output layer receives connections from hidden layer and it returns an output value that corresponds to the prediction value.

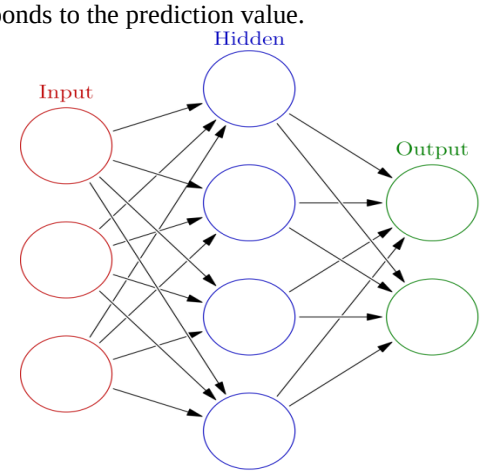


Fig3: Architecture of ANN

The advantages of ANN is storing the information on the entire network and it has the facility to function with incomplete knowledge.

IV. METHODOLOGY

For the predictive analysis using the classification algorithms such as RF, kNN, NB, SVM, ANN are implemented using the following Tool and Data Set.

A. Tool Used

R is language mainly used for data analysis and graphics. This tool mainly runs the Data Preprocessing, Cleaning, Web scraping and Visualization. The source code of classification algorithm in R tool can be modified and improved based on the user requirements. In this work, we use several packages such as, (i) caret- for classification and regression training, (ii) mlbench- for machine learning benchmark problems[17].

B. Data Set

The data sets for this work is obtained from UCI Machine Learning Repository. The breast cancer data sets are used for prediction and classification using five approaches being implemented. The cancer data set has 569 instances and 32 variables. Here the outcome to be predicted is 'M' (Malignant) or 'B' (Benign). In the datasets 70% of instances have been taken as training data and the remaining 30% has been considered as test data.

V. IMPLEMENTATION OF SUPERVISED LEARNING METHODS IN R ENVIRONMENT

A. Random Forest Algorithm

The random forest algorithm have been implemented in the Breast cancer data set and the prediction accuracy is given below.

Accuracy = Number of correct Predictions / Total number of instances for prediction.

		Predicted Class	
		B	M
Actual Class	B	106	10
	M	1	53

Prediction Accuracy = 159/170
= 93.5%

B. K Nearest Neighbor

The accuracy is calculated using the breast cancer data set in kNN algorithm.

The predicted accuracy is given below.

		Predicted Class	
		B	M
Actual Class	B	107	5
	M	0	58

Prediction Accuracy = 165/170 = 97.0%

C. Navie Bayes

The Navie Bayes algorithm have been implemented in the Breast cancer data set and the prediction accuracy is given below.

		Predicted Class	
		B	M
Actual Class	B	99	7
	M	8	56

Prediction Accuracy = 155/170 = 91.1%

D. Support Vector Machine

The Support Vector Machine algorithm have been implemented in the Breast cancer data set and the prediction accuracy is given below.

		Predicted Class	
		B	M
Actual Class	B	103	2
	M	4	61

Prediction Accuracy = 164/170 = 96.4%

E. Artificial Neural Networks

The Artificial Neural Network algorithm have been implemented in the Breast cancer data set and the prediction accuracy is given below.

		Predicted Class	
		B	M
Actual Class	B	107	3
	M	0	60

Prediction Accuracy = 167/170 = 98.2%

VI. RESULTS AND DISCUSSION

In this work the complete study on five classification algorithms were performed using R tool and observed the accuracy, which is shown in table 1.

Table 1. Performance Accuracy of various Machine Learning Algorithms

S.No	Types Of Classification Algorithm	Prediction Accuracy for Breast Cancer Data Set
1	Random Forest	93.5%
2	K Nearest Neighbor	97.0%
3	Navie Bayes	91.1%
4	Support Vector Machine	96.4%
5	Artificial Neural Network	98.2%

In classification task on Breast Cancer dataset, ANN and SVM algorithm shows highest accuracy. The accuracy will be differ based on the number of instances and attributes (features) to be considered for classification task. So that the comparative study on classification algorithms will be more efficient and effective is shown in table 2 and 3.

Table 2. Sensitivity and Specificity of ML algorithms for BCC data

ML Algorithm	Sensitivity	Specificity
RF	0.98	0.91
KNN	1.0	0.95
NB	0.87	0.93
SVM	0.93	0.98
ANN	1.0	0.97

Table 3: Precision, Recall and F-measure of ML algorithms for BCC data

ML Algorithm	Precision	Recall	F Measure
RF	0.84	0.98	0.90
KNN	0.92	1.0	0.95
NB	0.88	0.87	0.87
SVM	0.96	0.93	0.94
ANN	0.95	1.0	0.94

VII. CONCLUSION

The implementation of five classification algorithms such as Random Forest, K nearest neighbor, Naïve Bayes, Support Vector Machine and Artificial Neural Network on Breast Cancer Data Set using R environment produces the prediction accuracy. In which the classification algorithm results varies depends upon the data set charcetstics and attributes. The comparative analysis based on the performance metrics can be further enhanced by applying on various real-time datasets in future.

REFERENCES

[1]. Https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning.

[2]. Mengjie Yu, B.S. "Breast Cancer Prediction Using Machine Learning Algorithm", May 2017 - he University of Texas at Austin.

[3]. B Nithya "Comparative Analysis of Classification Methods in R Environment with two Different Data Sets", December 2017 - International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN: 2456-3307.

[4]. P.Dhivyapriya "Classification of Cancer Dataset in Data Mining Algorithms Using R Tool", February 2017 - International Journal of Computer Science Trends and Technology (IJCST) – Volume 5 Issue 1, Jan – Feb 2017.

[5]. Vikas Chaurasia1, Saurabh Pal "A Novel Approach for Breast Cancer Detection using Data Mining Techniques", January 2014 - International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 1, January 2014

[6]. Jahanvi Joshi, RinalDoshi "Diagnosis And Prognosis Breast Cancer Using Classification Rules" November 2014 - International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014 ISSN 2091-2730.

[7]. J.S.Saleema , N.Bhagawathi , S.Monica, P.Deepa Shenoy , K.R.Venugopal and L.M.Patnaik "Cancer Prognosis prediction using balanced Stratified sampling" February 2014 - International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.3, No. 1.

[8]. Tuba Kiyan, Tulay Yildirim "Breast Cancer Diagnosis Using Statistical Neural Networks" 2004 - Istanbul University - Journal of Electrical & Electronics Engineering Volume Number: 4.

[9]. Subrata Kumar Mandal "Performance Analysis of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree" 2017 - International Journal Of Engineering And Computer Science

ISSN: 2319-7242 Volume 6 Issue 2 Feb. 2017, Page No. 20388-20391.

[10]. Adnan Alam khan and Shariq Ahmed "Comparative analysis of data mining tools for lungs cancer patients" 2015 Journal of Information & Communication Technology Vol. 9, No. 1, (Spring2015) 33-40.

[11]. E. SathiyaPriya "A Study on Classification Algorithms and Performance Analysis of Data Mining using Cancer Data to Predict Lung Cancer Disease" 2017-International Journal of New Technology and Research (IJNTR) ISSN:2454-4116, Volume-3, Issue-11, November 2017 Pages 88-93.

[12]. ShamreenFathimaSaddique, Sharmithra P, Justin Xavier D "Prediction of Lung Cancer Using Classifier Models" 2016 - International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2016): 79.57 | Impact Factor (2015): 6.391.

[13]. Hlaudi Daniel Masethe and Mosima Anna Masethe "Prediction of Heart Disease using Classification Algorithms" 2014 - Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA.

[14]. T. Marikani "Prediction of Heart Disease using Supervised Learning Algorithms" 2017 - International Journal of Computer Applications (0975 – 8887) Volume 165 – No.5, May 2017.

[15]. R. Jothikumar and R.V. Sivabalan "Analysis of Classification Algorithms for Heart Disease Prediction and its Accuracies" 2016 - Middle-East Journal of Scientific Research 24 (Recent Innovations in Engineering, Technology, and Management & Applications): 200-206, 2016 ISSN 1990-9233.

[16]. Akil Bansal, Manish kumar Ahirwar, Piyush kumar sukla, "A Survey on Classification Algorithms used in Healthcare Environment of the Internet of Things". International journal of Computer Sciences and Engineering, Vol 6, Issue 7, Pp 883-887, July 2018.

[17]. J.Seetha and T.Chakravarthy, "Diabetes classification using machine learning techniques with the help of cloud computing". International journal of Computer Science and Engineering, Pp. 278-283, ISSN. 2347-2693, Vol. 6, Issue. 8, Aug-2018.

Authors Profile

Mr. S. Kumaravel, Associate Professor Department of Computer Science, A.V.V.M Sri Pushpam College, Poondi, Thanjavur. He has published 15 papers in reputed journal and one next. He has completed a UGC research project in the field of Embedded system

Miss. S.Ophilia Domanica Vithya, Research Scholar, A.V.V.M Sri Pushpam College, Poondi, Thanjavur. Her main research work focuses on machine learning.

.

.

