IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-2 E-ISSN: 2347-2693

Prediction of Breast Cancer using Decision tree and Random Forest Algorithm

N.Sridevi[1*], S.Anitha [2]

[1*]Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India
[2]Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India

[*]*Corresponding Author: sridevi.n78@gmail.com*

Available online at: www.ijcseonline.org

Received: 22/Jan/2018, Revised: 31/Jan/2018, Accepted: 13/Feb/2018, Published: 28/Feb/2018

Abstract— Breast cancer is one of the most leading causes of death among women. The early detection of anomalies in breast enables the doctor's in diagnosing the breast cancer easily which can save numerous of lives. In this work, Wisconsin Diagnosis Breast Cancer database is used for experiments in order to predict the breast cancer either benign or malignant. Supervised Machine Learning algorithms namely Decision tree and Random Forests are used to classify the breast cancer. R programming language is used to classify the breast cancer. The performances of the algorithms are measured in terms of accuracy, specificity and sensitivity. The functionality of the algorithms are analysed and the results were discussed.

Keywords— *Breast Cancer, Classification, Decision tree, Random Forests, R programming*

I. INTRODUCTION

Breast cancer is one of the most common cancer among women. The malignant tumour develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division. Although breast cancer is the leading cause of death in women, the survival rate is high when the cancer is identified early. Early diagnosis requires an accurate and reliable procedure that allows doctors to distinguish benign breast tumours form malignant tumours without undergoing surgical biopsy.

The primary objective of this research paper is to classify the patients to either a "benign" class or a "malignant" class which is cancerous. Predicting the disease is one of the most interesting and challenging tasks for which data mining applications can be developed. Use of computers with automated tools, large volumes of medical data are collected and made available to the medical researchers. As a result of this data mining techniques, has become most popular tool for researchers to identify and analyse the patterns and relationship among large number of features which made them to predict the type of disease using these data. According to Jain et al. [1], data mining can be used for classification, estimation, prediction, association rules, clustering, and visualization activities. Among these activities, prediction, classification, and estimation come under supervised learning. The main objective of this research paper is to use the supervised machine learning techniques to classify the breast tumours as benign or malignant.

The organization of the paper is as follows, section II discuss about the few related works available in prediction or analysis of breast cancer. Section III, describes about the dataset used for this research work. In section IV the methodology used to predict the breast cancer is explained and Section V describes briefly about the experimental result.

II. RELATED WORK

Numerous methods and algorithms have been adopted on classification of breast cancer. Among which few related works have been discussed here.

1. "Analysis of Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection" by Borges and Lucas Rodrigues [2]. In this paper, two machine learning algorithms namely Bayesian Networks algorithm and J48 are investigated. Several experiments were conducted using these algorithms and they found out that Bayesian network has a much better performance than the J48 algorithm.

2. "Analysis of k-means clustering approach on the breast cancer Wisconsin dataset" by Ashutosh Kumar Dubey et.al [3]. This study was aimed to find the effects of k-means clustering algorithm with different computation measures like centroid, distance, split method, epoch, attribute, and iteration and to carefully consider and identify the combination of measures that has potential of highly accurate clustering accuracy.
3. "Classification of Cancer Dataset in Data Mining Algorithms Using R Tool" by P.Dhivyapriya and Dr.S.Sivakumar [4]. Here Naïve Bayes and Support Vector Machine classifiers are used. Accuracy is compared over classifying two different cancer datasets. The best results are achieved using Naive Bayes classifier and Support Vector Machine after data preprocessing and adjustment of the classifiers.
4. "Classification of Breast cancer by comparing Back propagation training algorithms" by F.Paulin et al [5]. Here Feed Forward Artificial Neural Networks is used to classify the breast cancer. Levenberg marquardt algorithm gives the highest accuracy.

III. DATASET

The dataset used for this research work is Breast Cancer Wisconsin (Diagnostic) Dataset. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Ten real-valued features are computed for each cell nucleus. The information about the features are shown in Table 1. The data set consists of 32 attributes, out of which attribute 1 is id number and attribute 2 is diagnosis, the mean value, extreme value and standard error of each feature for the image is calculated, returning a 30 real-valuated attributes. All feature values are recoded with four significant digits. Each feature are evaluated on the scale of 1 to10, with 1 being the closest to benign and 10 closest to malignant. The dataset consists of 569 instances, among which 357 is benign, 212 is malignant. The dataset is divided into training and testing dataset in the ratio of 70% and 30% respectively.

Table 1: Features of Cell Nucleus

1	radius	6	compactness
2	texture	7	concavity
3	perimeter	8	concave points
4	Area	9	symmetry
5	Smoothness	10	fractal dimension

IV. METHODOLOGY

Classification is a technique that will assign category to a collection of data which can be used to perform analysis and prediction. Thus the goal of classification is to predict the target class for the given instance of data using acquired knowledge. For this research work two supervised machine learning classification algorithms namely Decision tree and Random Forest are considered to predict the patients to either a "benign" or a "malignant".

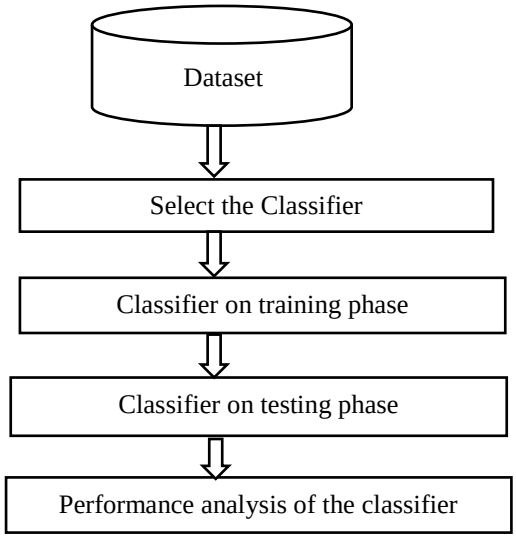

Fig 1: Classification Steps

A. Decision Tree Algorithm

Decision tree algorithm is a well-known supervised algorithm suitable to solve regression and classification problems. The objective of using decision tree is to create a training model that can be used to predict class of target variables by learning decision rules inferred from training data. Initially the dataset is divided into smaller subsets and incrementally builds an associated decision tree consisting of decision nodes and leaf nodes. The decision node has two or more sub trees where leaf node represents a classification. The root node is the topmost decision node that performs the prediction. Decision tree is suitable for both categorical and numerical data.

Pseudo code of Decision Tree Algorithm

1. Place the best attribute of the dataset at the root of the tree.
2. Divide the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until leaf nodes are found in all the branches of the tree.

For prediction using decision tree a class label for a record is compared with the **root** of the tree. Root attribute value is compared with the record attribute. Based on the comparison, select the branch corresponding to that value and follow the

next node. The process of comparing record's attribute values with other internal nodes of the tree is continued until a leaf node is reached with predicted class value.

B. Random Forest Algorithm

Random forest algorithm is a supervised classification algorithm. Random forest algorithm is similar to that of decision tree and creates the forest with a number of trees. In the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.

Pseudo code for Random Forest Creation

1. Randomly select "k" features from total "m" features, where k << m
2. Among the "k" features, calculate the node "d" using the best split point.
3. Split the node into child nodes using information gain as the best split method
4. Repeat 1 to 3 steps until "l" number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for "n" number times to create "n" number of trees.

Pseudo code for Random Forest Prediction

1. Test features are taken and the rules are used for each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the votes for each predicted target.
3. Consider the high voted predicted target as the **f**inal prediction from the random forest algorithm.

To perform the prediction using the trained random forest algorithm it is necessary to pass the test features through the rules of each randomly created tree.

V. RESULT AND DISCUSSION

In this paper, R Programming is used for implementing the classification algorithms. R is a programming language and environment which supports statistical computing and graphics. It provides software for data manipulation, calculation and graphical display. R provides packages that support classification through machine learning. For the experimental purpose the data set is divided into training and testing data in the ratio of 70% and 30% respectively. The effectiveness of the classification algorithms are measured using the evaluation metrics namely Accuracy, Specificity and Sensitivity. These metrics are calculated using the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) cases. These values are obtained from confusion matrix. Confusion matrix is a table that is often used to describe the performance of a classification algorithm on a collection of test data for which the true values are known. In our testing dataset there are 67 malignant data out of which 63 have been predicted correctly and only 4 are predicted wrongly. Also, there are 111 benign data of which 4 are predicted wrongly and 107 are predicted correctly. This analysis is shown in Table 2 and the evaluation metrics for classification algorithm are shown in Table 3

Table 2: Confusion Matrix obtained using Random Forest Classifier

Actual/Predicted	Predicted	
	Benign	**Malignant**
Benign	107	4
Malignant	4	63

Table 3: Evaluation Metrics for classification algorithm

	Classification Accuracy	**Specificity**	**Sensitivity**
Decision Tree	91.18%	82.54%	96.26%
Random Forest	95.72%	95.89	95.61

Figure 2, shows the graphical representation of the evaluation metrics for classification algorithm. From the figure it is apparent that the Random Forest performs better in terms of accuracy, sensitivity and specificity.

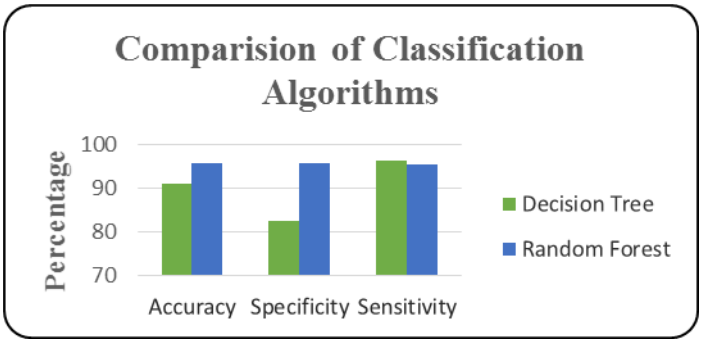

Fig 2: Graphical Representation of the quality performance of Classification Algorithms

VI. CONCLUSION

Breast cancer is the most common cancer in India. Early detection of breast cancer will increase the survival rate hence this research work is intended to predict the cancer

as benign or malignant which will assist the diagnosis process. Two supervised machine learning algorithms namely Decision tree and Random Forest algorithms are compared with the performance metrics such as accuracy, sensitivity and specificity using the Breast Cancer Wisconsin (Diagnostic) Dataset. From the experimental results it's evident that Random Forest algorithm predicts the breast cancer with the accuracy of 95.72%. Further enhancement of this work can include feature selection which may increase the accuracy of the prediction. In future, the performance of Random Forest algorithm can be compared with various classification algorithms like Naive bayes and SVM.

REFERENCES

[1] Jain R, "Introduction to data mining techniques", hhtp:// www.iasri.res.in/ebook/expertsystem/datamining.pdf

[2] Borges and Lucas Rodrigues, "Analysis of Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection", Proceedings of XI Workshop de Visão Computational, October 05th-07th, 2015.

[3] Dubey, A.K., Gupta, U. & Jain, S, "Analysis of k-means clustering approach on the breast cancer Wisconsin dataset", International Journal of Computer Assisted Radiology and Surgery, Vol.**11**, Issue **11**, pp. 2033–2047, November 2016 .

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

[5] F.Paulin et al., "Classification of Breast cancer by comparing Back propagation training algorithms", International Journal of Computer Sciences and Engineering (IJCSE), Vol **3**,No **1**, pp 327 – 332,Jan 2011.

Authors Profile

N. Sridevi, working as Assistant Professor in the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India. She has 4 years of Industrial experience and 6 years of teaching and research experience. Her area of research interest are Image processing, Pattern Recognition and Data mining.

S.Anitha, working as Assistant Professor in the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India. She has 12 years of teaching experience and 5 years of research experience. Her area of research interest are Image processing, Pattern Recognition and Data mining.

