May-June 2020
ISSN: 0193-4120 Page No. 3806-3809

Decision Tree based Multi-Classifier Model for Breast Cancer Screening System

[1]K. Vijayalakshmi, [2]M. Vinayakamurthy, [3]Anuradha V

[1]Associate professor, [2]Professor, [3]Head PG Department
[1,2]School of CSA, REVA University, [3]STC College, Pollachi, TamilNadu

Article Info
Volume 83
Page Number: 3806-3809
Publication Issue:
May-June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract
Breast cancer is recognized as most commonly diagnosed life threatening cancer among women. Breast cancer is a form of cancer which develops in mammary glands in the form of invasive tumor. According to WHO(World Health Organization), due to the above, almost 6,27,000 women died in the year 2018. By implementing Machine learning algorithms and techniques, researches have been undertaken to predict the threats of breast cancer. In this paper, the proposed work is presented with the integrated probability results of two decision tree based classifiers: J48 and Random forest algorithms. Finally the model has proven with the better accuracy, precision and recall with preprocessing when compared with unprocessed data.

Keywords: *Breast cancer, Machine learning, J48 decision tree, Random forest algorithm, preprocessing*

1. Introduction

In Breast cancer, malignant tumors originate from the breast tissues. Most regularly they originate from the inner linings of milk ducts that supply milk to the ducts. There are two types of Breast cancers: Ductal carcinomas, which affect the ducts of mammary gland and Lobular carcinomas, which affect the lobules of mammary gland. The Breast cancer affects mostly the women than the men. The Breast cancer symptoms are : change in breast shape, fluid from the nipple, skin patch in red and, lumps in the breast.

The factors for Breast cancer include drinking alcohol, lack of physical exercise, obesity, not at all having children or having children late, ionizing radiation. It can also occur due to prior history of Breast cancer in family history. About 5-10% of Breast cancer cases are inherent. Early detection isvery crucial in order to improve survival and outcomes of Breast cancer. For early detection of Breast cancer there are two strategies: Early diagnosis and screening. According to the WHO (World Health Organization), every year 2.1 million women are impacted due to Breast cancer. Breast cancer causes greatest number of cancer related deaths among women. It can be diagnosed using mammogram, MRI, biopsy and ultrasound.

Machine Learning as a branch of AI provides a variety of probabilistic and statistical methods. These methods will make the system to learn through vast history and experience to determine the relevant and significant patterns from huge datasets. The steps involved in Machine Learning are: Data collection, choosing the model, training the model, analyzing the model. Machine Learning had been used for decades to classify tumors and predict gene sequences that are responsible for cancer.

Supervised Learning is a type of learning in which we train the machine using training data. Both Random Forest and J48 algorithm comes under supervised learning. Both are different types of classification algorithms. Random forest shapes multiple decision trees and joins them together to get more precise and stable output. Decision trees with majority votes are chosen for Random forest prediction that can be used as classifiers and regression models. It has the ability to manage the missing values of an attribute and maintains accuracy for missing data. It won't support over fitting of the model.J48 is extended from ID3 algorithm. It's an Open source java application of C4.5 in WEKA data mining tool. The other features of J48 are raising the missing values, derivation rules etc. In J48, for predicting the target variable rules are generated. The additional features of J48 are accounting for missing values, decision tress pruning, continuous attribute value ranges, derivation of rules etc.[9]

May-June 2020
ISSN: 0193-4120 Page No. 3806-3809

2. Literature Survey

The work compares between KNN and Naïve Bayes classifier and calculated the accuracy using cross validation[1]. CNN method to classify the images of breast cancer into multiclasses -8 classes that helps in reducing the death rate[2]. Jyothismitha Talukdar and Dr Sanjib kr.Kalitha proposed diagnosis of BC using Data Mining techniques J48 and zero R. [3] B.Padmapriya, T.Velmurgan used J48, AD Tree and CART for finding the accuracy of classification techniques which is evaluated based on selected attributes of mammogram images[4].
S. Padmapriya , M.Devika, V.Meena, S.B Dheebika and R.Vinodhini used various classification techniques and listed the performance accuracy from them to select the best algorithm for prediction using WEKA. [5]. Daniele Soria and Jonatham M Garibaldi proposed thevalidation on fuzzy quantification , an extension on three different breast cancer datasets[6]. Phonethep Douanoulack and Veera Boorjing proposed that, out of J48 ,REP &Randon Tree, J48 classifier was the best classifer with accuracy of 97.36% with 2% rules respectively[7].
Nitasha proposed that classification algorithms show better accuracy and precision in pretending the breast cancer.It is also said that the accuracy depends on the type of data mining algorithm used. [8] Gaganjot Kaur and Amit Chhabra proposed changed version of J48 which improves in the accuracy rate using WEKA[9]. T.L Octaviani and Z. Rustam proved that Random forest runs efficiently in large databases and it is good at classification but not as good for regression. [10]

3. Methodologyused In Proposed Work

WEKA tool has a wide collection of machine learning algorithms and hence it is used in this paperwork for Breast cancer data modeling. Then methodology described in the procedure stepwise of the algorithm mentioned below is applied on the dataset provided below. The Breast cancer dataset has 10 attributes as shown in Fig 3.1 with 286 instances. The type of all the attributes are of nominal including the class label. For the proposal, the dataset is splitted into 60:40 % of training and test sets. The steps are implemented and the observations are made accordingly.

SL NO	ATTRIBUTE	DATA TYPE
1	Age	Nominal
2	menopause	Nominal
3	Tumour size	Nominal
4	Inv-nodes	Nominal
5	Node-caps	Nominal
6	Deg-malig	Nominal
7	Breast	Nominal
8	Breast-quad	Nominal
9	Irradiat	Nominal
10	Class	Nominal

Figure 3.1: Dataset Description
The algorithm for the proposed model is given below:

Algorithm:

STEP 1– Select the appropriate cancer dataset for Machine Learning.
STEP 2 – Apply J48 & Random Forest algorithm along with the integrated model that works on J48 and Random Forest using Meta classifier on the dataset.
STEP 3 – Apply preprocessing techniques to select the relevant attributes and transform nominal attributes to numeric values.
STEP 4 – Normalize the data.
STEP 5 – Apply the step 2 and measure the accuracy.

The flowchart for the proposed model is given below in the Fig 3.2:

Flowchart

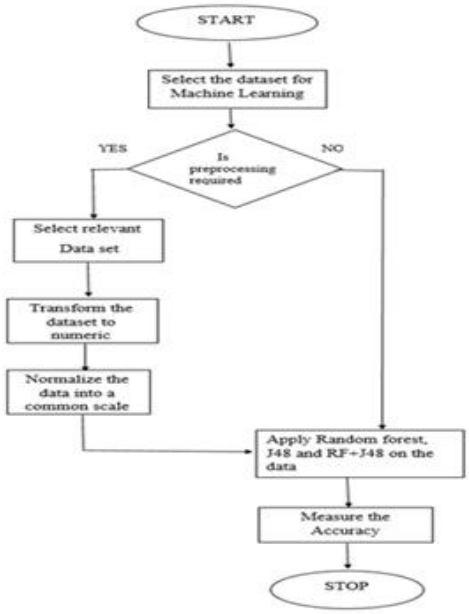

Figure 3.2: Flowchart of the proposed

J48 without preprocessing

The J48 decision tree is constructed with 4 leaf nodes and the size of the tree is 6. It has take 0.03 seconds to build this model. The J48 pruned tree is given below in fig 3.3:

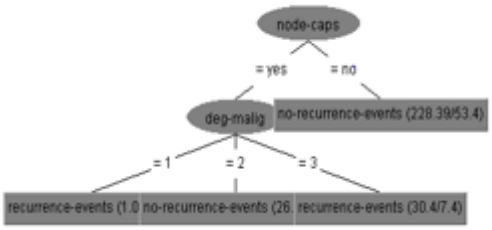

Figure 3.3: Decision Tree-J48

If node-caps == yes then
If deg-malig = =1 then
recurrence-events (1.01/0.4)
else if deg-malig == 2 then :
no-recurrence-events (26.2/8.0) else

May-June 2020
ISSN: 0193-4120 Page No. 3806-3809

recurrence-events (30.4/7.4) else
no-recurrence-events (228.39/53.4)

```
=== Confusion Matrix ===

  a  b   <-- classified as
 67  8 |  a = no-recurrence-events
 26 13 |  b = recurrence-events
```

Random forest Without preprocessing

It generates 10 trees from the subsets by considering 4 random features. The out of bag error is 0.3182. Time required to construct the model is 0.04 seconds

```
=== Confusion Matrix ===

  a  b   <-- classified as
 62 13 |  a = no-recurrence-events
 28 11 |  b = recurrence-events
```

Proposed Model with unprocessed data

The new hybrid model using the meta classifier voting algorithm combines the probability distributions of the above base learners J48 and random forest. 0.03 seconds is needed to build model.

```
=== Confusion Matrix ===

  a  b   <-- classified as
 68  7 |  a = no-recurrence-events
 26 13 |  b = recurrence-events
```

Proposed Model with processed data

The new hybrid model using the meta classifier voting algorithm combines the probability distributions of the above base learners J48 and random forest on the dataset which is preprocessed by selecting the relevant attributes. Out of 10, 6 attributes are selected and converted all the nominal to numeric values.

The numeric values of all attributes transformed into a common range of values. Finally the dataset is cleaned by replacing the missing values. The time taken to build model: 0.1 seconds

```
=== Confusion Matrix ===
| a  b   <-- classified as
 71  4 | a = no-recurrence-events
 25 14 | b = recurrence-events
```

The proposed work compares the performance of the integrated classifier model with J48 and Random Forest algorithms before and after preprocessing. The observation shows that the efficiency of the multi-classifier decision tree model has improved better in terms of prediction parameters after preprocessing.

4. Performance Metrics And Results

The model which is developed by integrating the multiple decision tree based classifiers has to be measured using various metrics. The outcome of the model is interpreted based on the measured values of those performance metrics. Those parameters highlights explicitly the behavior of the model and shows how varies from the existing in terms of relevance and precision factors. Sometimes beyond accuracy measure, few other metrics also need to be focused for the improvement in the model optimality. The following are the parameters used to measure the model :

a. Accuracy: it is an estimation metrics on how a model performs.

Accuracy = Number of correct prediction / Total number of predictions

b. Precision: it determines how near the measured values are to each other.

Precision = True positive / Actual results

c. Recall: it states to the percentage of total related results correctly classified by the algorithm

RECALL = True positive / Predicted result.

d. Error rate: The number of wrongly classified instances.

Error rate = 1- Accuracy

e. ROC AREA: Receiver Operating Characteristic curve is a graphical plot used to show the problem-solving ability of binary classifiers.

5. Interpretation of Results

From the table 5.1 , the accuracy given by J48 for the dataset is 70.1754% and for random forest is 64.0351%. The number of correctly classified instances of proposed model without preprocessing is 71.0526% and for proposed model with preprocessing is 75%.

Table 5.1: Breast cancer dataset results on various classifiers

	Accuracy	Precision	recall	ROC Curve
J48	0.7	0.686	0.702	0.572
Random Forest	0.64	0.61	0.64	0.64
Proposed model with unprocessed data	0.71	0.698	0.711	0.641
Proposed model with processed data	0.75	0.753	0.746	0.667

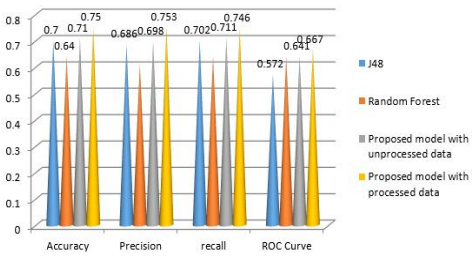

Figure 5.1: Comparison of multi-classifier hybrid classifier results

May-June 2020
ISSN: 0193-4120 Page No. 3806-3809

The accuracy of Random forest when combined with j48 is increased by 7% without preprocessing and an increase of 11% with preprocessing. Also form the Fig 5.1, the improvement in the recall and ROC values, improves the efficiency of the model proposed with data processed to be a better model than the exisiting.

6. Conclusion

In this paper, machine learning classsifers are used to construct the new prediction model for diagnosis of Breast cancer. Results show that the combination of J48 and Random forest gives more accuracy when compared with their individual performance. The measures to evaluate the performance of the model proposed are accuracy, PrecisonRecall and ROC area. Based on the overall measures, it clearly shows that the weak classifier as Random forest. When the weak classifier is combined with an another decision based classifier J48 which performs better, the performance of the integrated model can have better improvement in the results towards optimality of the model.

References

[1] Amrane, Meriem, Saliha Oukid, Ikram Gagaoua, and Tolga Ensarİ. "Breast cancer classification using machine learning." In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), pp. 1-4. IEEE, 2018.

[2] Nguyen, Phu T., Tuan T. Nguyen, Ngoc C. Nguyen, and Thuong T. Le. "Multiclass Breast Cancer Classification Using Convolutional Neural Network." In 2019 International Symposium on Electrical and Electronics Engineering (ISEE), pp. 130-134. IEEE, 2019.

[3] Talukdar, Jyotismita, and Sanjib Kr Kalita. "Detection of Breast Cancer using Data Mining Tool (weka)." International Journal of Scientific & Engineering Research 6, no. 11 (2015): 1124.

[4] Padmapriya, B., and T. Velmurugan. "Classification Algorithm Based Analysis of Breast Cancer Data." International Journal of Data Mining Techniques and Applications 6, no. 1 (2016): 43-49.

[5] Padmapriya, S., M. Devika, V. Meena, S. B. Dheebikaa, and R. Vinodhini. "Survey on Breast Cancer Detection Using Weka Tool." Imperial Journal of Interdisciplinary Research (IJIR) 2, no. 4 (2016).

[6] Soria, Daniele, and Jonathan M. Garibaldi. "Validation of a quantifier-based fuzzy classification system for breast cancer patients on external independent cohorts." In 2016 15th IEEEInternational Conference on Machine Learning and Applications (ICMLA), pp. 576-581. IEEE, 2016.

[7] Douangnoulack, Phonethep, and Veera Boonjing. "Building Minimal Classification Rules for Breast Cancer Diagnosis." In 2018 10th International Conference on Knowledge and Smart Technology (KST), pp. 278-281. IEEE, 2018.

[8] Nitasha. "Review on Breast cancer prediction using mining algorithms.", International Journal of Computer Science Trends and Technology(IJCST)-volume 7, issue 4, 2019.

[9] Kaur, Gaganjot, and Amit Chhabra. "Improved J48 classification algorithm for the prediction of diabetes." International Journal of Computer Applications 98, no. 22 (2014).

[10] T. L Octaviani and Z. Rustam, "Random Forest for Breast cancer prediction", AIP Conference proceedings 2168, 020050, 2019.