ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

A Novel Preprocessing Scheme For Predicting Breast Cancer Decease.

R Jeberson Retna Raj*, and Senduru Srinivasulu.

Department of IT, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India.

ABSTRACT

Classifying the data precisely is an important process for grouping the similar items. Analyzing and predicting the data is an important technique for any real time applications. Breast cancer is a dangerous decease which commonly available among woman. This type of cancer is known for its existence only after the later stage. The doctors may use default methods such as oral observation to identify the decease manual. However, these kind of procedure may resulted into wrong. The automatic classification techniques used with the machine learning algorithms will provide better results. The data is loaded and preprocessed using data preprocessing techniques. The machine learning techniques such as Logistic Regression, Decision Tree, KNN, Random Forest and Support Vector Machine algorithms are applied with the data set and predicted. The algorithm is implemented and tested with the UC data set and the performance of the algorithm is compared.

Keywords: component, formatting, style, styling, insert (key words)

https://doi.org/10.33887/rjpbcs/2019.10.4.18

**Corresponding author*

INTRODUCTION

Breast cancer is one of the dangerous decease which affects the woman seriously. WHO report shows that more than 508000 woman in worldwide dies due to breast cancer. There are four stages of breast cancer in the medical terminology. Stage 0 and 1 is very serious type of cancer which requires immediate treatment. This type of cancer may not know to the individual at the initial stage of the decease. If the decease is identified at the earlier stage it could have treated and recovered. The existing systems like expert systems, decision support system, and artificial intelligence may provide the better recommendations. However, they have their own advantage and disadvantages. Recently, machine learning algorithms emerged and they provide a better solution for predicting and analyzing the data. In this paper, the five machine learning algorithms such as Logistic Regression, Random Forest, KNN, SVM and Decision tree are used for experiment with the UCI data for breast cancer. The machine learning algorithms are rightly classified with the true positive and negative cancer deceases. The data set contains the fields such as catalog number, cancer position, sex, age, pathology, grade, stage, tumor node metastasis, and type of cancer.

Factors of breast cancer

Various factors are associated with breast cancer. Importantly, the chest tumor leads the way for breast cancer for the woman in their late ages. If The mother have the history of breast cancer, then the chances of affecting the daughter is more. Furthermore, the relative of a woman have chest infection, then the chances affecting the breast cancer is more for a woman in her life. More than 70% of woman in the age group of 40-50 have the risk of breast cancer. In the child age group of 0-20, more than 5% having the risk of break cancer. It is not only affecting women, Some men also rarely getting the decease. The breast cancer is categorized into three stages. The malignant stage is most dangerous, it requires immediate medical attention. The second stage is a starting stage and it is called normal. It also requires immediate medical attention, but not dangerous. The third stage is NAT which meant for not having the breast cancer. In this work, we have using the classification algorithm for classifying the cancer data using machine learning algorithms. In this work, we have using the supervised classification algorithm used. The main objective of this work is to find the best algorithm for predicting the breast cancer accurately. The python sklearn tool kit is used for prediction purpose.

Literature survey

Numerous papers have been presented in literature for predicting the breast cancer. They are based on preprocessing the input data and make use of classification algorithms to classify the data. The major challenges of classification or prediction is accuracy. The over classification or under prediction is an important threat to accuracy. There is a chances of over classification or under prediction which may hinder the accuracy. This work is based on preprocessing the cancer data using a novel algorithm and make use of the machine learning algorithms the cancer disease is predicted. Many papers have been cited in literature. In[1], Principal Component Analysis (PCA) is used to predict the breast cancer decease. Hooda et al proposed a second order derivative method is applied for the cancer images and the time delay estimation method is used to predict the cancer deceases. In[3], CNN method is used to identify the cancer cells. In [4], breast cancer is detected using deep convolutional neural networks method. In[5], a deep sea localization method is used to detect the cancer disease. Authors of [6] presented a multimodal deep neural network for predicting the survival life of human cancer is discussed. Existing system based on adaboost learning techniques used for classification. This algorithm limited in classifying top features and some of the important features may be ignored. This ignored feature is mainly required for classification. Furthermore, the accuracy of this classification is not up to the expected level.

METHODOLOGY

In supervised classification algorithm, we are using clever algorithm. The detailed flow is shown in figure1.

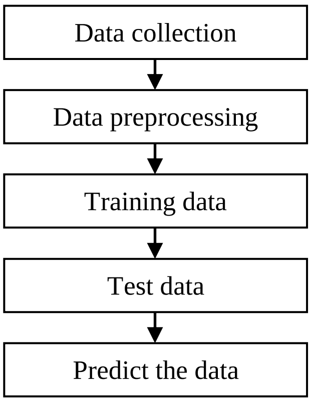
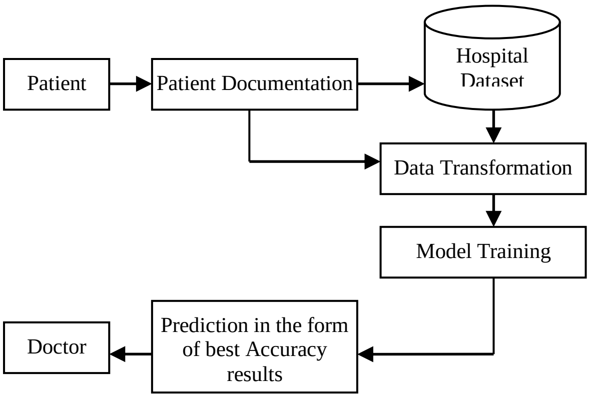


Figure 1 Flow of prediction process

Figure 1 show the flow of the prediction process. The cancer data is collected and it will be preprocessed for cleaning the data. The training and testing data is defined and finally the data is predicted.



The following steps are required to implement the proposed system :

Step 1: Exploration data analysis of variable identification
Step 2: Loading the given dataset
Step 3: Import required libraries packages
Step 4: Analyze the general properties
Step 5: Find duplicate and missing values
Step 6: Checking unique and count values Uni-variate data analysis
Step 7: Rename, add data and drop the data
Step 8: Pre-processing the given dataset
Step 9: Splitting the test and training dataset
Step 10: Comparing algorithm to predict the result
Step 11: Based on the best accuracy

Data collection

The data set for breast cancer is taken from UCI dataset. The description of the data set contains 9 fields. The followings are the fields:

Catelognum – patient id

Position – cancer cell position

Sex - Female or male

Age - 0-80

Pathology - grade(0-No cencer, 1-Normal treatment , 2- Dangerous)

Stage – four stages of cancer (I and II A –early stage, IIB – Most dangerous, III –dangerous)

Tumer Node Metasis(tnm) – tumor three types(0-normal, 1-medium, 2-dangerous)

Node (0-normal, 1-medium, 2-dangerous)

Metasis (0-normal, 1-medium, 2-dangerous)

Type – NAT (Not having breast cancer), Normal (beginning stage), Malignant(dangerous)

Data validation

Data validation is the process to validate the considered data is the right data or not. The CSV file format is validated and any missing and wrong data will be notified. The missing values are replaced with the predicted data with the collaborative filtering methods.

Data visualization

In data visualization, the data is represented in a graph. The user can visualize the age, type of cancer affected by in numbers etc. furthermore, which type of cancer dominating among the data.

Preprocessing

In preprocessing step, the raw data is converted into integers and it shows the classified view of the data. Here the training and testing data is defined. The training is defined as 70% and testing data is given as 30%. The label encoder method is used in python for complete the process.

Outlier detection process

In this outlier detection, we are using ski learn kit tool is used to calculate the mean, median and matrix. These methods show the stages of cancer decease. This show the data of Tumor Node Metasis(TNM) in a tabular column format. Furthermore, the grade of cancer is also showing in tabular column.

Machine learning algorithm for prediction

The machine leaning algorithm such as logistic regression, K-nearest Node, SVM, Random Forest and decision Tree algorithms are used and applied with the data set. The precision, recall, and F1-score is calculated. The values 0 to 0.5 defined as normal, 0.5 to 1 defined as dangerous. Next we are calculating the best accuracy of these five algorithms.

ALGORITHM AND TECHNIQUES

Breast cancer decease is a dangerous decease among woman. The early detection of this decease can get immediate medical attention which cures the decease. This paper employs the five machine learning algorithms and the result is compared.

Binary regression

In binary regression, binary value 1 defined as desired outcome variable. An independent variable should be included so that it independent to each other. The independent variables are linearly related to log adds and it requires large sample sizes.

Support Vector Machines (SVM)

The support vector machine works based on selecting the right hyperplane and number of classes can be defined.

K-Nearest Neighbor (KNN)

The algorithm based on distance between the items with the class. The k items closest with the class are defined and classified. The new item is placed in the respected class based on the closest distance with the group of items.

Decision tree (DT)

The decision tree is worked based on the following procedure:

a) Define the root by assign all the training set data.
b) Assume the attributes that should be continuous for information gain.
c) Based on the value of attributes, distribute it recursively.
d) Order the attributes based on statistical method for defining the root or internal node.

Random Forest (RF)

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks,

Step 1 : select k features randomly from m features, where k<m.
Step 2 : using best split method, calculate d as a best split point in k features.
Step 3 : daughter nodes to be identified by using the split point d.
Step 4 : Repeat the step 1 to 3 until l number of nodes reached.
Step-5 : Repeat the step 1 to 4 by n number of times for building forest to create n number of trees.

EXPERIMENTAL RESULTS

Necessary library packages

The system is implemented using python and the necessary library packages are configured. Sklearn is a packages is used for applying unsupervised classification algorithms and numpy used for calculations. Pandas used for making explicit report. Matplotlib is used for plotting and visualizing the data.

Accuracy Calculation

The accuracy of the prediction is quantified by the measurements precision and recall values. Higher the precision implies higher the accuracy of prediction. Furthermore, higher the recall value implies higher the accuracy. The values true positives, true negatives, false positive and false negatives are used to calculate precision and recall values.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Where, TP-True Positive, TN-True Negative, FP-False Positive, FN-False Negative. Furthermore the F1 score can be calculated using the following:

F1 Score = (2 * (Recall *Precision)) / (Recall + Precision)

F1 Score = 2TP / (2TP + FP + FN)

Table 1 shows the comparison of precision, recall f1-score and accuracy of five machine learning algorithm. According to this cancer data set, Decision Tree algorithm provides better precision and recall values. The decision tree algorithm provides better accuracy comparing with other algorithms. Next, KNN algorithm provides second higher accuracy with other algorithms.

Table: 2 Performance of the Algorithms

Algorithm	Precision	Recall	F1-Score	Accuracy (%)
LR	0.88	0.92	0.90	92.06
DT	0.97	0.98	0.98	98.41
SVM	0.85	0.92	0.88	92.06
RF	0.93	0.95	0.94	95.23
KNN	0.96	0.97	0.96	96.82

CONCLUSION

Prediction of breast cancer using machine learning algorithm is successfully implemented and tested. Five machine learning algorithms are applied with the data set and classified. The algorithm includes logistic regression, Decision Tree, Support Vector Machine, Random Forest and K-Nearest Neighborhood is applied with the dataset. The decision tree algorithm with a highest accuracy of 98.4%.

REFERENCES

[1] Hongchao Song, Aidong Men and Zhuqing Jiang, “Breast tumor detection using empirical mode decomposition features”,

[2] Hoda S. Hashemi, Stefanie Fallone and Mathieu Boily, “Assessment of Mechanical Properties of Tissue in Breast Cancer-Related Lymphedema Using Ultrasound Elastography”,

[3] Zhiqiong Wang, Mo Li, Huaxia Wang, Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion with CNN Deep Features, 2018

[4] Ravi K. Samala, Heang-Ping Chan and Lubomir Hadjiiski, Breast Cancer Diagnosis in Digital Breast Tomosynthesis: Effects of Training Sample Size on Multi-Stage Transfer Learning using Deep Neural Nets, 2018

[5] Dong Wei, Susan Weinstein, Meng-Kang Hsieh, “Three-Dimensional Whole Breast Segmentation in Sagittal and Axial Breast MRI with Dense Depth Field Modeling and Localized Self-Adaptation for Chest-Wall Line Detection

[6] Dongdong Sun, Minghui Wang, “A multimodal deep neural network for human breast cancer prognosis prediction by Integrating multi-dimensional data”, 2018