Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 9, Issue. 3, March 2020, pg.230 – 235

Trend Analysis of Breast Cancer Stages using Supervised Machine Learning Algorithm

Abirami J; Balaji S; Aashish M; Anubharathi.B.U
Dept. of computer science and Engineering, Rajalakshmi Engineering College, Chennai
abirami.j.2016.cse@rajalakshmi.edu.in; balaji.s2.2016.cse@rajalakshmi.edu.in;
aashish.m.2016.cse@rajalakshmi.edu.in; anubharathi.bu@rajalakshmi.edu.in

Abstract: Breast cancer is one of the common forms of cancer among women. Breast cancer mostly affects women's only. Breast cancer is the second leading cancer that cause to death. In very few rare case men can be affected by breast cancer representing the majority of new breast cancer cases and cancer-related deaths according to global statistics, making it a most important human health problem in today's society. Women's are affected by breast cancer due to the growth in cells inside the beast. Either it can be left breast or right breast. In this paper we analyze the past breast cancer data to know how many peoples are affected by this breast cancer in the range of year 1990 to 2017. Here we will visualize graph analysis to show how breast cancer was affecting people. First one is based on time, second one is based on age group and the third one is based on tumor size. Here we analyze the breast cancer deaths based on the information in the data set. Data cleaning, data validation, data preprocessing will be done on the entire data set. Here we are using supervised machine learning algorithm for analyzing the historical report to predict the breast cancer.
Keywords: Data set, prediction, Graph Analysis, supervised machine learning algorithm, historical report.

Introduction

Carcinoma is a term that is used to identify and describe a cancer that begins in the epithelial of organs like breast. Most of the breast cancer are carcinomas. Carcinomas are starts in granular tissue, which is called as adenocarcinomas. The normal breast contains tiny tubes, which is also called as ducts and that ends with group of lobules.

Cancer starts in the cells lining, when a normal cell becomes a carcinoma cell. As long as the carcinoma cells are still compact to the breast ducts or lobules, without breaking out and growing into neighbouring tissue, it is considered in-situ carcinoma (or carcinoma in situ). Once the carcinoma cells have grown and broken out of the ducts or lobules, then it is known as invasive or infiltrating carcinoma. In an invasive carcinoma, the tumor cells can spread (metastasize) to other parts of human body. Based on microscopic view Breast cancer are often divided into 2 main types. One is invasive ductal carcinoma and another one is invasive lobular carcinoma. In some cases, the tumor can have symptoms of both carcinomas and is called a mixed ductal and lobular carcinoma. Another name for invasive ductal carcinoma is invasive mammary carcinoma of no special type, because it is the most common type of breast cancer carcinoma. In general, invasive lobular and invasive ductal carcinomas of the breast aren't treated separately. Most important medical issue found in middle aged women is, breast cancer. Breast cancer incidence rate can be reduced if it can be detected at an early stage of breast cancer. With the help of latest, efficient and advanced screening methods, the majority of such cancers are getting reduced. The utility of machine learning techniques in human healthcare analysis is growing good. Certainly analysis of patient's clinical data and doctor's suggestions are the most considerable terms in diagnosis. Most of the possible medical flaws can be recovered by the using classification systems, and it is also offer healthcare data to be analyze in lesser time and in more exhaustive manner. Accurate and timely prediction of breast cancer is used by the doctors to make most accurate decision about the patient treatment. In vision of the problem statement described in this section, a classification model is proposed with more accuracy to predict the breast cancer patient. Classifications of the entire data sets are done on the basis of specific properties possess by the sample variable that this capable to classify the data set and each sample variable is assigned as a malignant or benign class. Classification is mostly done by making predictions based on known sample data that has been learned from training data. Designed algorithm is first trained on the known data labels and further uses this learning to predict the class labels for the new unknown data sample.

Machine learning is used to predict the future from past data. Machine learning (ML) is comes under artificial intelligence (AI). ML provides computers with the capability to learn without being explicitly programmed. Data scientists may use different kinds of machine learning algorithms to discover patterns from past dataset in python that lead to actionable awareness. At a high level, these different algorithms can be classified into two category based on the way they "learn" about data to make accurate predictions: supervised learning and unsupervised learning. Here we are using supervised machine learning algorithm. Majority of machine learning uses supervised machine learning algorithms to predict the results. Supervised learning method having 2 variables. One is input variable (X) and another one is output variable (y) and it uses an appropriate algorithm to learn the mapping function from the input variable (X) to the output variable (Y) is $y = f(X)$. Our goal is to find the mapping function .it is useful to find the output variable from the input variable. The techniques of supervised machine learning include logistic regression, decision tree, support vector machine and naive Bayesian.

Knowledge Extraction:

Data source:

We have collected the hospital data set from kaggle website. Our dataset having the following variables age, menopause, tumor size, Inv-nodes. Node caps, Deg-malig, Breast, Breast-quad, Irrad, time and Class. After collecting the data set, data validation and preprocessing will be done on the data set.

Method of processing:

We split our project into 7 modules. Here we use 6 library packages. Numpy is for numerical calculation. Pandas for data handling. Matplot is for graphical representation. Seaborn is for advanced graphical representation. Sklearn is a machine learning package and it is used to perform 80% of implementation. Tkinter is used for creation of GUI. First one is for data validation and data preprocessing. Data validation is the process of eliminating duplicate values and null values in the data set. Second module for data visualization and here we perform graphical analysis based on age group, tumor size and time. In third module we use logistic regression and decision tree algorithm for prediction. Both are giving 100% accuracy. In fourth model we use random forest and support vector machine algorithm for prediction. Random forest gives 100% accuracy and SVM gives 96.42% accuracy. In fifth module we use naïve bays and k-nearest neighbor algorithm. Naive bays give 100% accuracy and KNN gives 85.71% accuracy. In sixth module we use ensemble voting classifier to give final result based on naïve bays, logistic regression, decision tree and random forest. In seventh module we are going to create one GUI. Our final result will be grade and TNM. TNM for tumor size and node and metastasis.

Architecture Design:

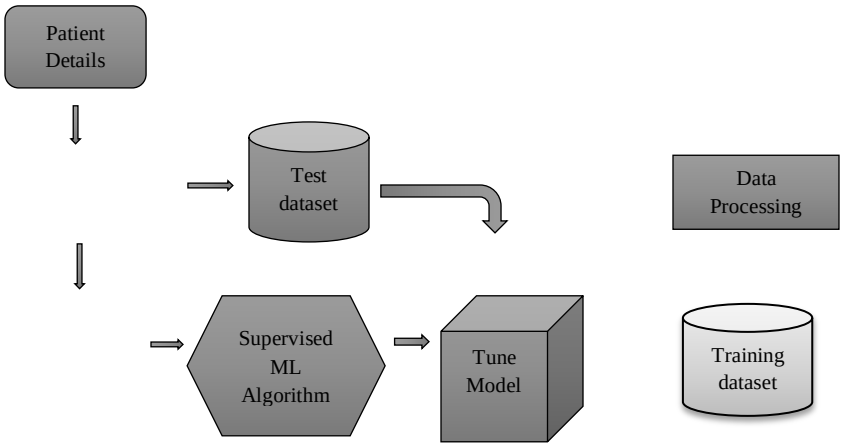

Fig: Data processing

Existing system:

The patterns frequently appearing within the tumors with the identical label is considered a possible diagnostic rule. Subsequently, the diagnostic rules are very useful to construct component classifiers of the Adaboost algorithm by using unique rules combination strategy which sort out the problem of classification in numerous feature spaces (PC-DFS). Finally, the AdaBoost learning is performed to find effective combinations and integrate them into a robust classifier. The proposed approach has been validated employing a large ultrasonic dataset of 1062 breast tumor instances (including 418 benign cases and 644 malignant cases) and its performance was compared with several conventional approaches. The experimental results show that the proposed method yielded the most effective prediction performance, indicating a decent potential in clinical applications.

Proposed system:

EXPLORATORY DATA ANALYSIS

Here we are using *Jupyter notebook* to work on this given dataset and will first go with importing the necessary libraries packages and import our dataset to Jupyter notebook platform.

Splitting the dataset

The data set we are using is usually divide into training data and test data. The training set contains already familiar output and the model learns on this data in order to be generalized to other data later on. It has the test dataset (or subset) in order to test our model's prediction on this subset and it will do this using SciKit-Learn library package in Python using the train_test_split method.

Data collection

The data set collected for predicting patient is categorized into Training set and Test set. Generally, 7:3 ratios are used to split the Training set and Test set. The Data Model which was created using naive baysien algorithm are applied on the Training set and based on the test result accuracy, Test set prediction will be done.

Preprocessing

The data which was collected might contain missing values that ends up to inaccuracy. To get better results data need to be preprocessed so as to provide 100% accuracy of the algorithm. The outliers must be off from the data set and also variable conversion need to be done. Based on the correlation among attributes it absolutely was observed that attributes that are significant individually include tnm, stages, grade, age, which is that the strongest among all. Some variables like an applicant income and co- applicant income aren't significant alone, which is strange since by intuition it's considered as important.

Building the classification model

The predicting the breast carcinoma problem, decision tree algorithm prediction model is robust thanks to below mentioned reasons: It provides the simplest leads to classification problem.

- It is efficient in preprocessing outliers, irrelevant variables and a combination of continuous, categorical and discrete variables.
- It produces out of bag estimate error which has proven to be unbiased in various tests and it's relatively easy to tune with.

Trend Analysis:

Trend analysis is nothing but analyze the past data in order to predict the future results. Here, we analyze the breast cancer patient past details to show how people were affected by breast cancer in past 20 years. We create 3 graphs to show how people were affected by breast carcinoma. One is based on age group and another one is based on time and last one is based on tumor size. Example given below

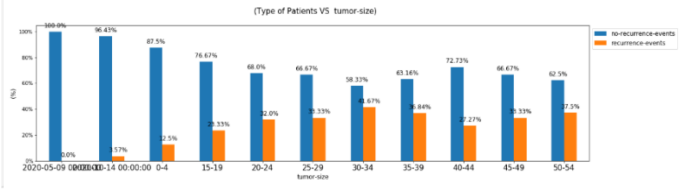

Fig: Trend analysis based on tumor size

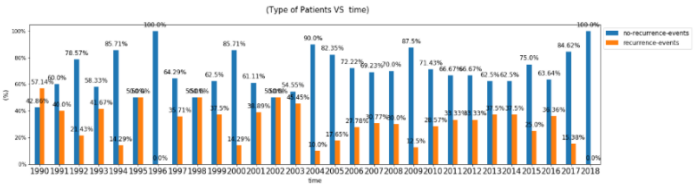

Fig: Trend analysis based on time

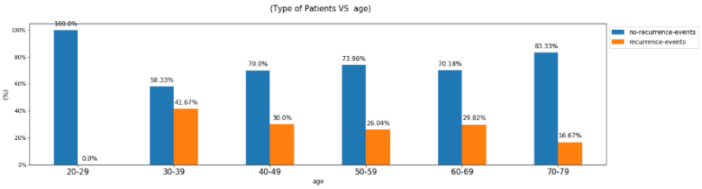

Fig: Trend analysis based on age group

GUI creation:

Finally, we create one Graphical User Interface. In that we will show 2 outcomes. We will show in which grade you will be healthy. Another one is tumor size, node and metastasis. Using this we can identify our healthy stages. Example given below

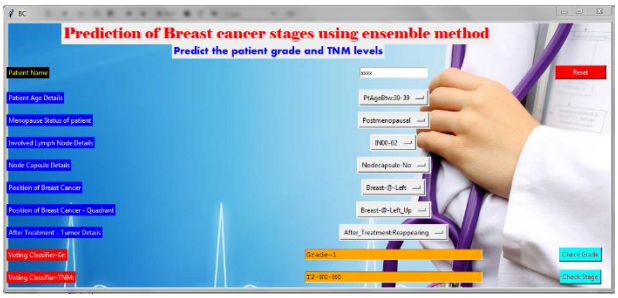

Fig: Breast cancer healthy stage prediction

Conclusion:

The analytical process started from data cleaning and processing, missing value, exploratory analysis and eventually model building and evaluation. Finding the patient stages and grade with parameter like accuracy, classification report and confusion matrix on public test set of given attributes by supervised machine learning method. So finally we've built our classification model and to know that machine learning classification algorithm gives the foremost effective results for our dataset. Well it's not always applicable to each dataset. To make decision our model, we always must analyze our dataset then apply our machine learning model.

References:

[1]. A sharp decrease in breast cancer incidence in the united states in 2003 proceedings from the 2006 annual San Antonio Breast Cancer Symposium (SABCS) San Antonio, TX, USA December 14, 2006.
[2]. Howe HL,Wu X, Ries LA,Collinidies V, Ahmed F, Jemal A, Miller B, Williams M, Ward E, Wingo PA et al: Annual report to the nation on the tatus of cancer, 1975-2003, featuring cancer among U.S. Hispanic/Latino populations, Cancer 2006,107:1711-1742.
[3]. Miller BA, Feuer Ej, Hankey BF: The increasing incidence of breast cancer since 1982: relevance of early detection. Cancer Causes Control 1991,2:67-74.
[4]. White E, Lee CY, Kristal AR: Evaluation of the increase in breast cancer incidence in relation to mammography use. J Natl Cancer Inst 1990, 82:15446-1552.
[5]. Garfinkel L, Boring CC, Health CW Jr: Changing trends. An over view of breast cancer incidence and mortality. Cancer 1994, 74(1 Suppl) 222=227.
[6]. Tarone RE, Chu KC: Implications of birth cohort patterns in interpreting trends in breast cancer rate. J Natl Cancer Inst 1992, 884:1402-1412.

