VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology and Engineering

(SITE)

M. Tech (Software Engineering)

Software Development Project

REVIEW - 2

"A MACHINE LEARNING FRAMEWORK FOR PREDICTION OF BREAST CANCER"

Submitted for the course

SWE3004 Software Development Project Report

Fall Semester 2019-2020

Guided By

(Prof. BRINDHA K)

Submitted By

G.BHARGAV -- (15MIS0115)

OCTOBER 2019

Guide Signature

1. Module Description:

There are 4 modules, which are:

- Pre-processing
- Future Selection
- Classification

Pre-processing:

The pre-processing is the most important step in the mammogram analysis due to poor captured mammogram image quality. Pre-processing is very important to correct and adjust the mammogram image for further study and processing. There are Different types of filtering techniques are available for pre-processing. This filters used to improve image quality, remove the noise, preserves the edges within an image, enhance and smoothen the image. In this paper, we have performed various filters namely, average filter, adaptive median filter, average or mean filter, and wiener filter.

The steps to be taken are:

- Mean Filter or Average Filter
- Median Filtering
- Adaptive median filter

Future Selection:

It has been an attempt over last many years that how to detect and cure cancer. Cancer which can be of various types like breast cancer, lung cancer, throat cancer, blood cancer etc. is known to be the deadliest disease which still now have not got any cure

There are several levels of cancer from 1 to 6. However, on the good side if cancer is detected when it is in level 1 or 2 or at the very initial stage, there is a significant probability that it will get cured within a period of time. With the advent of new technologies in the field of medicine, we get new ideas of curing the disease with methods like machine learning. The problem for the cancer can be broadly classified into three types.

1. **Firstly,** the problem is to predict whether a person in a particular stage of cancer has the chance to survive or not.

2. **Secondly,** the problem is to predict whether a person who has already encountered the disease in the past and got cured, has the probability of having the same disease in future.

3. **Thirdly,** the domain in which we are working includes the detection of cancer at the earliest stage.

Classification:

Classification of benign tumors can help the patients avoid undertaking needless treatments. Most of them show that classification techniques give a good accuracy in prediction of the type of tumor. Our methodology involves use of classification techniques like Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Logistic Regression, with Dimensionality Reduction technique i.e. Principal Component Analysis (PCA) .

A classification problem is when the result is a category like filtering emails “spam” or “not spam”. Unsupervised Learning: Unsupervised learning is giving away information to the machine that is neither classified nor labeled and allowing the algorithm to analyze the given information without providing any directions. In unsupervised learning algorithm the machine is trained from the data which is not labeled or classified making the algorithm to work without proper instructions. In our dataset we have the outcome variable or Dependent variable i.e. Y having only two set of values, either M (Malign) or B (Benign). So Classification algorithm of supervised learning is applied on it. We have chosen three different types of classification algorithms in Machine Learning.

Classification Techniques:

- **Logistic Regression:**
  - Logistic Regression is a supervised machine learning technique, employed in classification jobs (for predictions based on training data).
  - Logistic Regression uses an equation similar to Linear Regression but the outcome of logistic regression is a categorical variable whereas it is a value for other regression models.
  - Binary outcomes can be predicted from the independent variables. The outcome of dependent variable is discrete. Logistic Regression uses a simple equation which shows the linear relation between the independent variables.
- **K-NN (K-Nearest Neighbor):**
  - K-Nearest Neighbor is a supervised machine learning algorithm as the data given to it is labeled. It is a non-parametric method as the classification of test data point relies upon the nearest training data points rather than considering the dimensions (parameters) of the dataset.
  - In Classification technique, it classifies the objects based on the k closest training examples in the feature space.
  - It catches the idea of proximity based on mathematical formula called as Euclidean distance, calculation of distance between two points in a plane.
- **SVM (Support Vector Machine):**
  - Support Vector Machine is a supervised machine learning algorithm which is doing well in pattern recognition problems and it is used as a training algorithm for studying classification and regression rules from data.
  - SVM is most precisely used when the number of features and number of instances are high. A binary classifier is built by the SVM algorithm.

- This binary classifier is constructed using a hyper plane where it is a line in more than 3-dimensions.

Classification states:

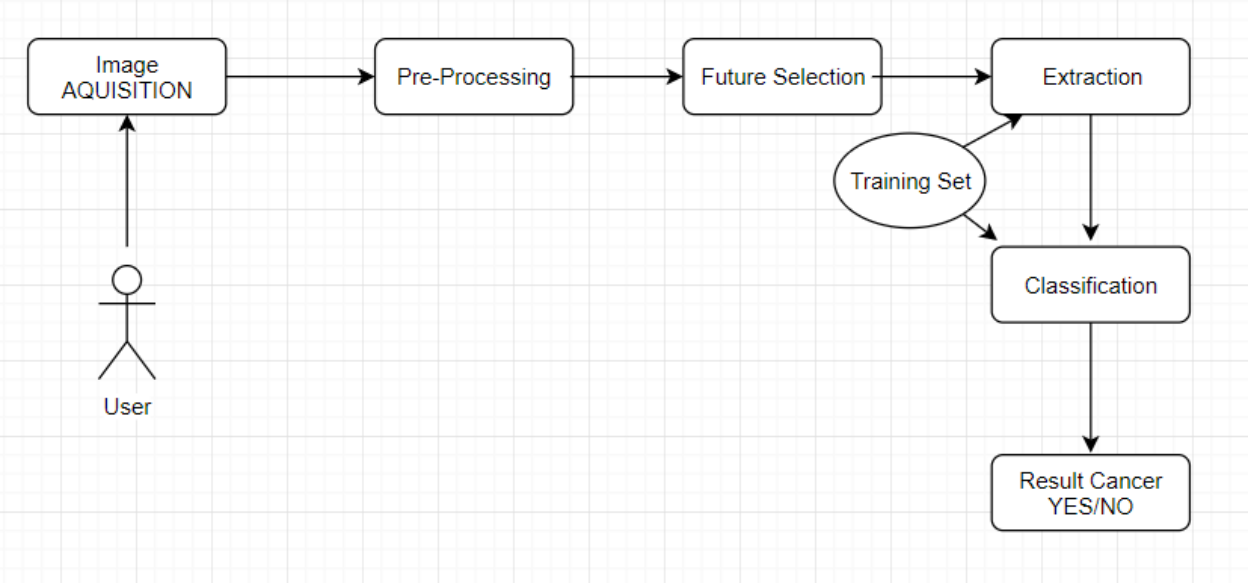
1. **Benign**

Benign is a state of tumor or lump which does not spread across body. In simple terms these are non-cancerous type of tumor.

2. **Malignant**

Malignant is a state of tumor or lump which spreads across the body. In simple terms these are cancerous type of tumor.

2. Complete Design:



3. Algorithms Used:

1. Pre-processing:

The main goal of the pre-processing is to improve the image quality to make it ready to further processing by removing or reducing the unrelated and surplus parts in the background of the mammogram images Mammograms are medical images that complicated to interpret. Hence pre-processing is essential to improve the quality. It will prepare the mammogram for the next two-process segmentation and feature extraction. The noise and high frequency components removed by filters.

a. Mean Filter or Average Filter:

The goal of the mean filters used to improve the image quality for human viewers. In this, filter replaced each pixel with the average value of the intensities in the neighborhood. It locally reduced the variance, and easy to carry out. Limitations of average filter

- Averaging operations lead to the blurring of an image, blurring affect features localization.
- If the averaging operations applied to an image corrupted by impulse noise, the impulse noise attenuated and diffused but not removed.
- A single pixel with a very unrepresentative value affected the mean value of all the pixels in neighborhood significantly.

b. Median Filtering:

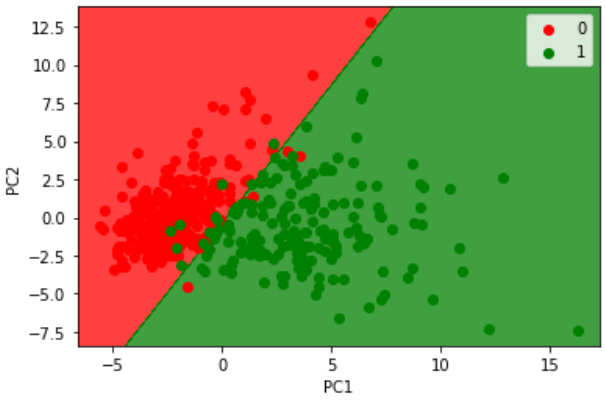
A median filter is a nonlinear filter is efficient in removing salt and pepper noise median tends to keep the sharpness of image edges while removing noise. The several of median filter is

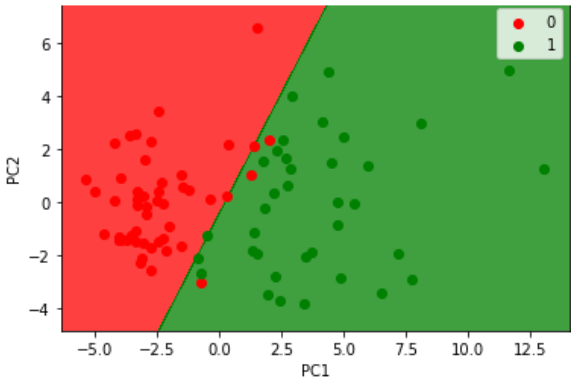
- Centre-weighted median filter
- Weighted median filter
- Max-median filter, the effect of the size of the window increases in median filtering noise removed effectively.

c. Adaptive median filter:

Adaptive Median filtering used to smooth the non-repulsive noise from two-dimensional signals without blurring edges and preserved images. This makes, it particularly suitable for enhancing mammogram images. The preprocessing techniques used in mammogram, orientation, label, artifact removal, enhancement and segmentations. The preprocessing involved in creating masks for pixels with highest intensity, to reduce resolutions and to segment the breast

2. Classification – SVM:





4. Novelty:

Existing System:

- Rule based approach is used for classification which gives static range value for different classes. Therefore we will not able dynamic images or outlier behaviour images.
- Multi classification problem is not optimized and ignore the class imbalance problem. Therefore in learning all classes is not input in learning phase so it became a biased learning.
- Features set is not normalize. Therefore different features show different outputs and show different representation during training phase of classifiers.

Proposed System:

In our proposed system, mammogram image can be enhancement using Gaussian filter. Second the segmentation is done using Fuzzy C means for partitioning the mammogram image into multiple segments to identify the mass easily and features are extracted using HWT (Haar Wavelet Features). Further tumor has been analyzed and classified using Multi –SVM (Support Vector Machine) classifier.

SVMs deliver a unique solution, since the optimality problem is convex. This is an advantage compared to Neural Networks, which have multiple solutions associated with local minima and for this reason may not be robust over different samples.

Advantages:

- More Accuracy
- Reduced time consumption

5. **Implementation**:

```
import os
import numpy as np
import pandas as pd
import seaborn as sns
import datetime as dt
import matplotlib.pyplot as plt
from IPython import get_ipython
get_ipython().run_line_magic('matplotlib', 'inline')
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
data = pd.read_csv('C:/Users/Gurramkonda Bhargav/Desktop/breastcancer.csv')
data.info()
data.head()
data.diagnosis.unique()
data.describe()
data.drop('id',axis=1,inplace=True)
data.drop('Unnamed: 32',axis=1,inplace=True)
data['diagnosis'] = data['diagnosis'].map({'M':1,'B':0})
datas = pd.DataFrame(preprocessing.scale(data.iloc[:,1:32]))
datas.columns = list(data.iloc[:,1:32].columns)
datas['diagnosis'] = data['diagnosis']
datas.diagnosis.value_counts().plot(kind='bar', alpha = 0.5, facecolor = 'b', figsize=(12,6))
plt.title("Diagnosis (M=1 , B=0)", fontsize = '18')
plt.ylabel("Total Number of Patients")
plt.grid(b=True)
data.columns
data_mean =
data[['diagnosis','radius_mean','texture_mean','perimeter_mean','area_mean','smoothness_mean',
```

```
'compactness_mean', 'concavity_mean','concave points_mean', 'symmetry_mean', 'fractal_dimension_mean']]
plt.figure(figsize=(14,14))
foo = sns.heatmap(data_mean.corr(), vmax=1, square=True, annot=True)
_ = sns.swarmplot(y='perimeter_mean',x='diagnosis', data=data_mean)
plt.show()
from sklearn.model_selection import train_test_split, cross_val_score, cross_val_predict
from sklearn import metrics

predictors = data_mean.columns[2:11]
target = "diagnosis"

X = data_mean.loc[:,predictors]
y = np.ravel(data.loc[:,[target]])

# Split the dataset in train and test:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
print ('Shape of training set : %i || Shape of test set : %i' % (X_train.shape[0],X_test.shape[0]) )
print ('The dataset is very small so simple cross-validation approach should work here')
print ('There are very few data points so 10-fold cross validation should give us a better estimate')
from sklearn import svm

# Initiating the model:
svm = svm.SVC()

scores = cross_val_score(svm, X_train, y_train, scoring='accuracy' ,cv=10).mean()

print("The mean accuracy with 10 fold cross validation for Random Forest Model is %s" % round(scores*100,2))
from sklearn.neighbors import KNeighborsClassifier
```

```
# Initiating the model:
knn = KNeighborsClassifier()

scores = cross_val_score(knn, X_train, y_train, scoring='accuracy' ,cv=10).mean()

print("The mean accuracy with 10 fold cross validation for Tree is %s" % round(scores*100,2))
from sklearn.ensemble import RandomForestClassifier
```

6. References:

- Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classificationwith deep convolutional neural networks," in InternationalConference on Neural Information Processing Systems, Lake Tahoe,Nevada, USA, December 2012, pp. 1097–1105.

- Chi C.L., Street W.H. and Wolberg W.H., "Application of Artifical Neural Network-based Survival Analysis on Two Breast Cancer Datasets", Annual Symposium Proceedings / AMIA Symposium, 2007.

- William H Wolberg, Olvi Mangasarian, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA

- Chao-Ying ,Joanne, PengKukLida Lee, Gary M. Ingersoll –"An Introduction to Logistic Regression Analysis and Reporting ", September/October 2002 [Vol. 96(No. 1)]

- AlirezaOsarech, BitaShadgar,"A Computer Aided Diagnosis System for Breast Cancer",International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.