Breast Cancer Classification

Vineeth Kumar, Renukuntla

Spring 2021

***PROBLEM DESCRIPTION***

Breast cancer is growing problem among women, knowing the stage of it is a must and treated accordingly. Here, by using the Classification algorithms, we can classify the tumor as Benign or Malignant. The tumor is considered to be benign if the cells that are developing out of control are normal cells (noncancerous). The tumor is classified as malignant if the out-of-control cells are abnormal and do not behave like the body's normal cells (cancerous). In 2020, an estimated 3,676,262 women in the United States were diagnosed with female breast cancer. Here, by using the Classification algorithms, we can classify the tumor as Benign or Malignant. It’s a binary classification problem.

***DATESET DESCRIPTION***

Here we got the data from UCI repository, it has got 32 attributes with 569 records.

Feature Information:

1)ID number (unique, Just patient id which has no meaning)

2) Diagnosis (‘M’ = malignant, ‘B’ = benign , our target variable)

The ten real valued features here were calculated for each cell nuclei: (All were continuous variables)  
  
1) Radius (the average of the distances between the center and the perimeter points)  
2) texture   
3) area  
4) perimeter  
5) smoothness   
6) compactness   
7) concavity   
8) concave points   
9) fractal dimension   
10) symmetry

The standard error, mean and worst of above mentioned attributes were enumerated for each picture, that will result in 30 attributes . Here column 4 represents Area mean, column 14 represents Area se, and column 24 represents worst Area.

***DATA PREPROCESSING:***

As our dataset got entire null values in 33rd column, we dropped the column using pandas.

Here our target variable is ‘Diagnosis’ which has two values (‘M’ for Malignant, ‘B’ for Benign)

Chart, bar chart

Description automatically generated

As our target variable is categorical variable, we need to encode the data values (0,1 here)

A picture containing background pattern

Description automatically generated

***VISUAL ANALYSIS:***

It’s more important these days in order to find how two attributes are related, whether it’s adds weight to the target variable or not, if two attributes are adding same weight to the target variable and are correlated, we can use only one attribute from them. Here we are using correlation heat map using seaborn packages.

Some of the observations I found from the below correlation heatmap are: perimeter mean, radius mean, and area mean are correlated, here we consider radius mean feature. compactness mean, concave points mean and concavity mean are correlated. Here we choose concave points mean.

A picture containing text

Description automatically generated

***SCALING THE DATA:***

The first step in our data analysis is to scale the data. If the data isn't scaled, the algorithms can only consider magnitude rather than units. The predictors with the highest magnitudes are assigned the most weight. A standard scaler is best for symmetrical data, but a robust scaler is best for distorted data with outliers. If we don’t scale the data before applying the algorithms, higher values get the greater weights and lower values will get smaller weights. Here we used standard scaler to scale the data.

***ALGORITHMS :***

1. *Support Vector Machine*: Here we used ‘rbf’ kernel since our data is not linearly separable, used scikit learn package to apply SVM algo on our scaled data. To find the hyperplane, which can separate objects of different classes in a such way that it leaves the maximum margin.
2. *Decision Trees*: we used criterion as ‘entropy’ and used sckit learn package to apply decision tree algo on our scaled data. Decision tree is constructed from top down, starting with root node, and involves splitting the data into subsets containing instances with similar values . ID3 algorithm calculates a sample's homogeneity using entropy. An evenly split sample has an entropy of one, while a completely homogeneous sample has an entropy of zero.
3. *Logistic Regression*: In order to forecast an output value, input values (x) are linearly combined with coefficient values(y). The major difference from linear regression is that the output value being modeled is a binary value (0 or 1) rather than a numeric.Used sckit learn to implement the algo.
4. *Random Forest Classifier*: Here we used n\_estimators = 10 (i.e., 10 decision tree base models) which is randomly selected subset of training set. It then collects votes from all the base models to predict the final class of test objects. Used sckit learn package to implement the algo on our scaled data.

The accuracies resulted are listed below:

|  |  |
| --- | --- |
| Algorithm | Scaled Normal |
| SVM | 0.95 |
| Decision Tree | 0.932 |
| Logistic Regression | 0.94 |
| Random Forest | 0.959 |

***EXPERIMENTS:***

*1) Using SelectKBest:*

Choosing the 5(i.e., k=5 here) best features given by Sckit’s ‘SelectKbest’ .With the five features we will train our models .SelectKbest best inbuilt package in Sckitlearn .Here we pass chi2 as parameter.

The function Selectkbest calculate chi 2 stat measure inbetween each attribute of 𝑋(x1,x2….xn) and y(target variable). Low value must indicate that attribute (xi) is independent of 𝑦. High value must indicate that attribute is more correlated to 𝑦, and more likely to give significant information. Depends on number of k, till that attributes are retained , rest dropped.

*2) Selecting Features based on correlation:*

we here find the correlation between the features and remove one of two features that have a correlation threshold greater than 0.85. That brings down to the 13 features from 32.

The selected features are texture mean ,area mean , smoothness worst , compactness se ,concavity mean ,area se , concavity worst ,concave points se ,fractal dimension se , symmetry worst , fractal dimension worst , concavity se, fractal dimension mean.(based on correlation heatmap)

***CONCLUSION:***

Random forest and SVM performed very well even with features down to 5. XGBoost classifier generally gives highest accuracy for classication problems, since we scaled the data here ,we didn’t use it. Overall random forest is consistent in all the experiments we did.