Breast Cancer Prediction

Anmol Jain
MT19005
anmol19005@iiitd.ac.in

***Abstract*— Breast Cancer is the most often identified cancer among women and major reason for increasing mortality rate among women. As the diagnosis of this disease manually takes long hours and the lesser availability of systems, there is a need to develop the automatic diagnosis system for early detection of cancer. Data mining techniques contribute a lot in the development of such system. For the classification of benign and malignant tumor I have used classification techniques such as decision tree , random forest of machine learning in which the machine is learned from the training data and can predict test data.**

I. Problem Statement

Predicting Breast Cancer from features extracted from digitized image of a fine needle aspirate (FNA) of a breast mass and doing detailed analysis of the type and version of the feature using various machine learning algorithms is the goal of the project. This will help in better and fast cancer prediction , help women to take better actions in near future.

II. Literature Review

Chaurasia and Pal[1] compare the performance criterion of supervised learning classifiers, such as Nave Bayes, SVM-RBF kernel, RBF neural networks, Decision Tree (Dt) (J48), and simple classification and regression tree (CART), to find the best classifier in breast cancer datasets. The experimental result shows that SVM-RBF kernel is more accurate than other classifiers; it scores at the accuracy level of 96.84% in the Wisconsin Breast Cancer (original) datasets.
Chaurasia and Pal[2] conducted an experiment to identify the most common data mining algorithms, implemented in modern Medical Diagnosis, and evaluate their performance on several medical datasets. Five algorithms were chosen: Nave Bayes, RBF Network, Simple Logistic, J48 and Decision Tree. For the evaluation two Irvine Machine Learning Repository (UCI-UC) databases were used: heart disease and breast cancer datasets. Several performance metrics were utilized: percent of correct classifications, True/False Positive rates, area under the curve (AUC), precision, recall, F-measure, and a set of errors. Delen et al.10[3] had taken 202,932 breast cancer patients records, which then pre-classified into two groups of survived (93,273) and not survived (109,659). The results of predicting the survivability were in the range of 93% accuracy.

III. Database Details

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.Dataset characteristics is Multivariate with 32 features and 569 instances.Every instance is classified into two category Every datapoint was classified into two classes i.e. malignant and benign.

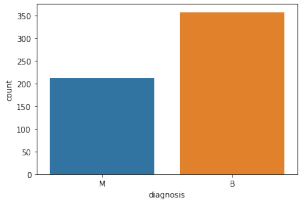

Fig. 1. Malignant and Benign Count

IV. Task Performed

A. Data Pre-processing

Diagnosis is our class label. There is an ID column that cannot be used for classification and a feature named Unnamed:32 which includes NaN so it is removed leaving 30 features. From the table it is observed that max value of area_mean = 2501 and compactness_mean is 0.16 there is so much

difference in these value area value will overpower value of smoothness so we need to standardize or normalize our data. As values have different units we will standardize our data.

area_mean , concavity_mean in swarm plot looks like malignant and benign are seprated not totaly but mostly. Hovewer, smoothness_se , fractal_dimension_worst in swarm plot looks like malignant and benign are mixed so it is hard to classify while using this feature.

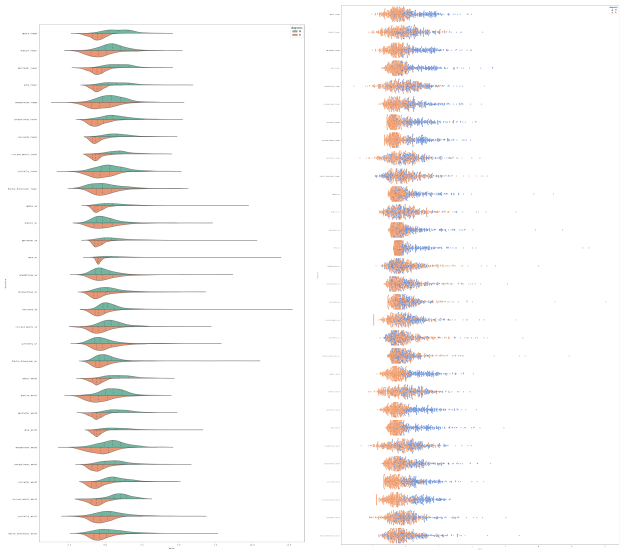

Fig. 2. Data Variation

B. Data Analysis

1) Univariate Analysis: Best 5 feature to classify is that area_mean, area_se, texture_mean, concavity_worst and concavity_mean.

After selecting these 5 features from Correlation Matrix following Accuracy were observed on some known models

- Decision Tree - 90.78 %
- Naive Bayes - 91.66%
- Logistic Regression -90.78%
- Random Forest-93.105%

2) Multivariate Analysis: From Correlation map we can drop features which are highly correlated , as one can derive the other and those features can be neglected.It can be seen in map heat figure radius_mean, perimeter_mean and area_mean are correlated with each other so we will use only area_mean.

After selecting feature from Correlation Matrix following Accuracy were observed on some known models:

- Decision Tree - 92.98%
- Naive Bayes - 93.42%
- Logistic Regression -92.98%
- Random Forest-93.90%

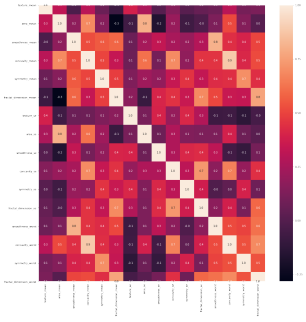

Fig. 3. Correlation Matrix

C. Outlier Analysis

To remove the outliers in the dataset, we used the boxplot method. In the Remaining features of the dataset, all those datapoints that lie outside the Inter Quartile range are considered as outliers. All those data points which have outlier even in one feature are removed from the dataset.

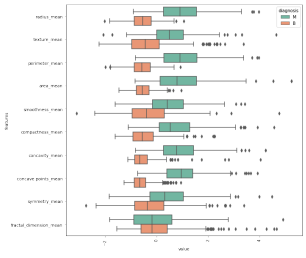

Fig. 4. Outlier Analysis

D. Feature selection

Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. Features are ranked by the feature_importances attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model. Recursive feature elimination technique is applied on

different model to analyse cross validation score with amount of feature to be selected . Optimum features on applying recursive elimination:

- Decision Tree - 10

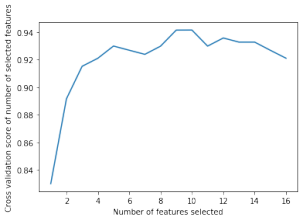

Fig. 5. CVscore vs Feature

- Logistic Regression - 16

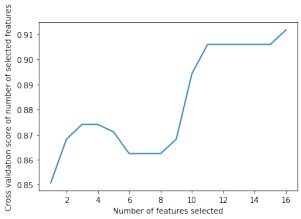

Fig. 6. CVscore vs Feature

- Random Forest - 13

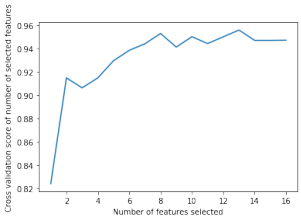

Fig. 7. CVscore vs Feature

E. Applying Model

During analysis, after dropping high correlated features , we observed that most of the features had no correlation so we applied Naive Bayes as it is best suited for datasets that have independent features. We also tried Decision Tree , Logistic Regression and Random Forest to observe the predictions as data set is small and has binary classification.

V. CONCLUSIONS

Accuracy of data set increased from 89% to 93% when data set is processed and analysed. On analyzing data on different models accuracy of random forest classifier turns out to 94% when test data splits into 60-40 ratio of train and test data respectively, 5 features 'area_mean', 'concavity_mean', 'area_se', 'concavity_se', 'concavity_worst' contributes most to the predictions.

REFERENCES

[1] Chaurasia, V, Pal, S. Data mining techniques: to predict and resolve breast cancer survivability. Int J Comput Sci Mobile Comput 2014; 3: 1022.

[2] Chaurasia, V, Pal, S. A novel approach for breast cancer detection using data mining techniques. Int J Innovative Res Comput Commun Eng 2014; 2: 24562465.

[3] Delen, D, Walker, G, Kadam, A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005; 34: 113127. .