Diagnosing Breast Cancer using Julia

Julia Poster
Julia in Intelligent Computing (JICom), IJCCI 2019

Introduction

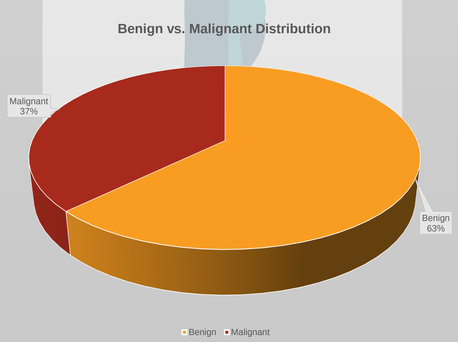
Breast cancer is one of the major causes of the high number of women's death. Therefore, early diagnosing breast cancer can promote timely clinical treatment to patients. Moreover correct classification of the benign tumor can prevent patients from undergoing unnecessary treatments. Thus, diagnosing breast cancer and accurate prediction of malignant and benign tumors is an important research issue.

In this work, we have diagnosed breast cancer on the Breast Cancer Wisconsin(Diagnostic) Dataset which is also available in the UCI Machine Learning Repository.

Dataset

This dataset has 569 transaction instance with thirty one features and one target class. Also, It has only 212 instances of benign class and other 357 instances belong to malignant class.

Parameter	Explanations and Values
Dataset Name	Breast Cancer Wisconsin(Diagnostic) Dataset
Source	Kaggle
Number of Instances	569
Features	31
Feature Type	Numeric and Categorical
Missing Values	No
Class Value	2
Learning Type	Classification
Class Instances	Benign => 357 | Malignant => 212



Methodology

In this work, a Random Forest classifier algorithm is applied. The algorithm can be described into data preparation phase and training phase.

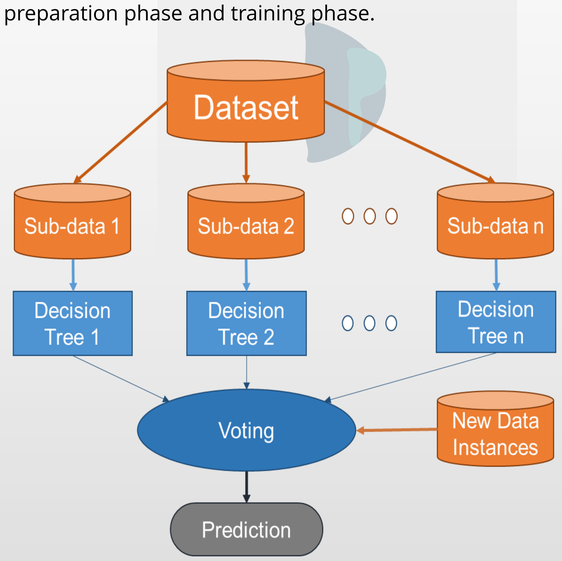


Figure: Workflow of Random Forest Classifier

In data preparation phase, the training dataset is divided into several individual sub-dataset.

In the training phase, all these balanced sub dataset, created in the data preparation phase is employed with multiple decision tree to train a ensemble classifier.

Finally, new instances is predicted by using voting mechanism.

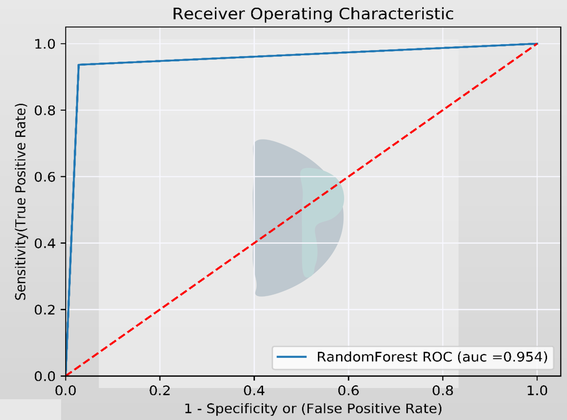
Results

Confusion Matrix		Actual		
		Malignant	Benign	Total
Predicted	Malignant	59	4	63
	Benign	3	105	108
	Total	62	111	173

Based on the confusion matrix's the performance table is calculated. The following table shows average result of 3 iteration.

Performance Parameter	Score
Accuracy	96.10%
True Positive Rate or Sensitivity or Recall	92.96%
Specificity or True Negative Rate	98.10%
AUC	95.54%
Gmean	95.49%
Precision	96.81%
F1_score	94.83%
False Positive Rate	01.90%

The ROC curve with 95.54% AUC score shows that the trained model is capable enough to distinguish between the benign and malignant classes of breast cancer.



Impacts

Economic

- Accurate classification of the benign tumor can prevents unnecessary treatment cost.
- Faster breast cancer detection compared to real time detection

Health and Safety

- Early detection of malignant cancer can increase the chance of survival.

Conclusion

The ensemble classifier "Random-Forest" can leverage cases from all the segment of the majority class instances. Thus, it shows a significant result with 0.190% False Positive rate in detecting malignant class from the imbalanced dataset.

Future Work

We shall continue the investigation on the performance of our approach in an excellent manner with other established techniques like feature selection or sampling technique. Moreover, modifying the core algorithm can open a new door in classifying malignant tumorous cancer as an malignant one with minimal error.

Team

I. Dipta Paul
ID: 011 143 008
Contact:
idpaulyd@gmail.com

Acknowledgements

Mentor:
Dr. Dewan Md. Farid
Associate Professor
United International University

References

1. Hiba Asri et al, *2016, Elsevier,* "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis".
2. Dewan Md. Farid et al**,** 2016, *Benelearn, "*A New Data Balancing Method for Classifying Multi-Class Imbalanced Genomic Data"
3. Md Yasir Arafat et al, 2017, *SKIMA,* "Cluster-based Under-sampling with Random Forest for Multi-Class Imbalanced Classification,"
4. C Nguyen et al, 2013, *Scientific Research Publishing,* "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic".
5. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data