Breast Cancer Prediction

Saurabh Biswas & Gourav Verma

DSC-630 Predictive Analytics

Bellevue University, NE

Fall-2020

# Abstract

Breast cancer or breast carcinoma is uncontrolled growth of epithelial cells in the breast. The uncontrollable division of one cell results in visible mass named tumor. Tumor can be benign or malignant. By Johns Hopkins Pathology, benign tumors are non-malignant/non-cancerous tumor and malignant tumors are cancerous growths. A cancer is another word for a malignant tumor. Most benign tumors respond well to treatment. But, malignant tumors are often resistant to treatment, may spread to other parts of the body and they sometimes recur after they were removed. Under a study by University of Wisconsin, 569 patients (212 with cancer and 357 with benign masses) provided the data for diagnostic algorithm. Diagnostic 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 in the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34]. Breast cancer is the most common malignancy in women, with over 200,000 being diagnosed in the US every year. 40,000 women will die from it each year. The objective of this research is to provide a comparative study on the utilized potential classification tools (Linear regression, random forest) on the problem by a benchmark dataset which consist of numeric cellular shape features extracted from preprocessed Fine Needle Aspiration biopsy image of cell slides.

*Keywords:* K-Mean, AFINN, elbow method, clustering, scikit-surprise, recommendation model, BaselineOnly.

Table of Contents

[Abstract 2](#_Toc52833366)

[Problem Statement 4](#_Toc52833367)

[Data 4](#_Toc52833368)

[Methods 10](#_Toc52833369)

[Results 11](#_Toc52833370)

[Acknowledgments 11](#_Toc52833371)

[References 11](#_Toc52833372)

# Problem Statement

It is 2nd most common cancer in women. But, on rare scenario can also happen to men. It is 2nd leading cause of deaths of women after lung cancer. A fine needle biopsy is an effective tool in evaluating and diagnosing suspect lumps or masses. With early diagnosis of breast cancer, patients can be isolated for early treatment for a better chance of survival. Common biopsies for breast cancer diagnosis includes fine-needle aspiration (FNA), core needle biopsy, and MRI-guided biopsy. In this analysis, we will be using ten features of tumour cell nuclei extracted from the digital image processing of an FNA of a breast mass to predict breast cancer. The data is collected from UCI Machine learning repository.

# Data

The benchmark dataset in this research will obtained from the UCI Irvine machine learning repository http://archive.ics.uci.edu/ml/index.html. This dataset was originally created by Dr. Wolberg, Street and Mangasarian all from University of Wisconsin. Data items in the dataset are composed of ID number, the diagnosis which will either be classified as malignant (M) or benign(B) and numeric shape features of extract cellular nuclei such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, and symmetry and fractal dimension. Each of the dataset in the observation is composed of 30 variables and 10 of the featured variables are related to the aforementioned characteristics. As described the website following 10 features were computed for each cell nucleus and will be used as input for our analysis. All feature values are recoded with four significant digits.

1. radius (mean of distances from center to points on the perimeter)
2. Texture (standard deviation of gray-scale values)
3. Perimeter
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness (perimeter^2 / area - 1.0)
7. Concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)
9. Symmetry
10. Fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, and field 23 is Worst Radius. There are no missing values in the dataset. ‘diagnosis’ is the target variable. The total numbers of data samples with different labels are not balanced. In fact, the total number (357) of data samples labelled as B (benign) is almost doubled the total number (212) of data samples labelled as M (malignant).



For the purpose of running machine learning algorithm, the target variable values have been translated from B (benign) to 0 and M (malignant) to 1. Target variable ‘0’ will be treated as no cancer and ‘1’ will be treated as breast cancer. Snapshot of the correlation table:



Based on the above heatmap and correlation table value - radius\_mean, perimeter\_mean & area\_mean have higher correlation, radius\_worst, perimeter\_worst & area\_worst have higher correlation (>0.98). So, we keep one of each column and drop other correlated columns for further analysis. We also can see that diagnosis variable has higher correlation value with perimeter\_mean, concave points\_mean, perimeter\_worst, concave points\_worst (>0.70). Before performing any data cleaning activity, we will split our dataset into train and test dataset in 75/25 ratio.





Any outliers in perimeter\_mean, concave points\_mean, perimeter\_worst, concave points\_worst will have impact in final outcome of the model. So, it is important to identify the outliers and remove them from training dataset. perimeter\_mean, concave points\_mean, perimeter\_worst contained outliers. So, we removed them from training dataset and used the cleaned dataset for model training.

# Methodology

Our aim is to build a classification model that will identify breast cancer using diagnosis label. For this purpose, we are planning to use logistic regression classifier, Random Forest classifier. Logistic Regression classifier will establish the baseline training results and we will compare that with results from Random Forest.

Logistic regression (LR) is a statistical method similar to linear regression since LR finds an equation that predicts an outcome for a binary variable, Y, from one or more response variables, X. However, unlike linear regression the response variables can be categorical or continuous, as the model does not strictly require continuous data. To predict group membership, LR uses the log odds ratio rather than probabilities and an iterative maximum likelihood method rather than a least squares to fit the final model. This means the researcher has more freedom when using LR and the method may be more appropriate for nonnormally distributed data or when the samples have unequal covariance matrices. Logistic regression assumes independence among variables, which is not always met in morphoscopic datasets. However, as is often the case, the applicability of the method (and how well it works, e.g., the classification error) often trumps statistical assumptions.

Random forest is a classifier that evolves from decision trees. It actually consists of many decision trees. To classify a new instance, each decision tree provides a classification for input data; random forest collects the classifications and chooses the most voted prediction as the result. The input of each tree is sampled data from the original dataset. In addition, a subset of features is randomly selected from the optional features to grow the tree at each node. Each tree is grown without pruning. Essentially, random forest enables a large number of weak or weakly-correlated classifiers to form a strong classifier. The Random Forests algorithm had a substantial impact on medical image computing over the last decade. This chapter presents basic algorithmic details, some variations proposed in the recent years and applications in medical image computing. Arguably, Random Forests' main impact was on the analysis tasks that required understanding spatial context within the images. We take a specific angle and view Random Forests as a machine learning tool that can integrate contextual information. We position the algorithm and its contributions within the larger field from this respect. Lastly, we briefly discuss how Random Forests and deep learning methods relate to each other and how they differ.

We used sklearn’s GridSearchCV to fine tune the parameters for above model and then compared the results. Our plan is to use Pipeline and pass different parameters for different models to GridSearchCV using Pipeline for module tuning. Random Forest classifier are not best suited for skewed class distribution. Our dataset is imbalanced dataset. So, in order to use Random Forest classifier, we need to calibrate the model.

# Results

After tuning our models, we will select the parameters with best results in terms of f1-score & precision recall curve. Precision Recall curve is a good way to measure the performance of an imbalanced dataset. We will also look at the complexity of the model. e.g. if we have close f1-score for random forest and SVM classifier, then we will select Random Forest for our final model as this is easier to explain to business stakeholders.

# Discussion

Early detection of breast cancer cells can be predicted accurately by the use of machine learning techniques. This may result in the decrease of health cost and may enhance time required for a patient to receive treatment. In this project the linear regression and the Random Forest have been discussed in providing diagnostic and prognosis assessment for breast cancer. The LR has been determined to be more superior to Random Forest since it provides higher prediction accuracy.

# Acknowledgments

I want to acknowledge Dr. William H. Wolberg, W. Nick Street & Olvi L. Mangasarian for putting together this breast cancer dataset. I want to also acknowledge UCI Machine Learning Repository team for hosting this dataset in their repository.

# References

<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>

<https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian. Computerized breast cancer diagnosis and prognosis from fine needle aspirates. Archives of Surgery 1995;130:511-516.

You, H., & Rumbe, G. (2010). Comparative study of classification techniques on breast cancer FNA biopsy data.

W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian. Computer-derived nuclear features distinguish malignant from benign breast cytology. Human Pathology, 26:792--796, 1995.

Koelliker, S. L., Chung, M. A., Mainiero, M. B., Steinhoff, M. M., & Cady, B. (2008). Axillary lymph nodes: US-guided fine-needle aspiration for initial staging of breast cancer—correlation with primary tumor size. Radiology, 246(1), 81-89.

Mu, T., & Nandi, A. K. (2007). Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM–RBF classifier. Journal of the Franklin Institute, 344(3-4), 285-311.

McManus, D. T., & Anderson, N. H. (2001). Fine needle aspiration cytology of the breast. Current Diagnostic Pathology, 7(4), 262-271.

Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), 570-577.

Azmi, M. S. B. M., & Cob, Z. C. (2010, December). Breast cancer prediction based on backpropagation algorithm. In 2010 IEEE Student Conference on Research and Development (SCOReD) (pp. 164-168). IEEE.

Xiong, X., Kim, Y., Baek, Y., Rhee, D. W., & Kim, S. H. (2005, May). Analysis of breast cancer using data mining & statistical techniques. In Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Network (pp. 82-87). IEEE.