Using Predictive Models for Tumor Classification and Recurrence

Anand Bhave
Courant Institute of Mathematical Sciences
New York University
New York, NY 10003
asb761@nyu.edu

Seong Woo Han
Courant Institute of Mathematical Sciences
New York University
New York, NY 10003
swh324@nyu.edu

Hongji Li
Tandon School of Engineering
New York University
New York, NY 11201
hl2491@nyu.edu

Abstract

The treatment of breast cancer depends on the stage of cancer, and the use of machine learning in medical area can analyze the data in an efficient way. This paper aims to develop a model built from features computed from a digitized image of a fine needle aspirate of a breast mass that predicts whether a given tumor sample is benign or malignant based on Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC) dataset.

1 Introduction

Approximately, two hundred thousand Americans suffer from breast cancer, and its treatment depends on the stage of cancer [1]. As the use of machine learning in medical domain takes root and becomes more prevalent, data can be analyzed and modeled in an efficient way. This allows us to gain insight into our data set to make more informed decisions. Breast cancer is a cancer that develops from breast tissue, which usually occurs to women. The cause of breast cancer is unclear, but its risk increases for a woman who has certain types of benign breast lumps and increases significantly for a woman who has previously had cancer of the breast [2]. Even though there is a great deal of public education and scientific research, breast cancer is considered the most common cancer in women, and its early diagnosis is crucial [3]. We aim to develop a model computed from a digitized image of a fine needle aspirate (FNA) of a breast mass, which predicts whether a given tumor sample is benign or malignant based on three different data sets having some mutually inclusive features. Another aim is to predict whether the tumor is recurrent or not based on relevant features like time elapsed and other cell nuclei properties.

2 Experiment

We use Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC) dataset, acquired from the public dataset of UCI Machine Learning.

2.1 Dataset

The Wisconsin Breast Cancer (WBC) consists of visually assessed nuclear features of fine needle aspirates (FNA) taken from patients. Tumor malignancy or benign diagnosis (Label attribute) is determined by performing a biopsy. WBC has 699 samples, and the sample distribution is 458 benign and 241 malignant.

The Wisconsin Diagnosis Breast Cancer (WDBC) datasets contains information about features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The image contains cell nuclei and has a label field that enlists whether a tumor is benign or malignant and a distinct set of features from the WBC dataset. WDBC has 569 samples, and the sample distribution is 357 benign and 212 malignant.

The Wisconsin Prognosis Breast Cancer (WPBC) dataset contains information whether a tumor is recurrent or not and the corresponding period of time elapsed. Also, the data contains tumor size and lymph nodes. WPBC has 198 samples, and the sample distribution is 151 non-recurrences and 47 recurrences.

The three data sources are relatively clean but have some missing values. For handling the missing values, the average values for the particular features were computed per class label and then the missing values were substituted with that specific feature average value with respect to the particular class label.

2.2 Preprocess

For the Wisconsin Breast Cancer(WBC) dataset, we map the class label (Malignant and Benign) into numerical label (1/0). Specifically, "malignant" tumor corresponds to 1, "benign" tumor corresponds to 0. For an initial analysis, the mean and standard deviation for every feature with respect to the output label is computed. We use the dataset with plotting histograms for the features to see the value ranges for each feature categorized by the class label (Malignant or Benign). The features, clump_thickness, size of tumor, shape of tumor, marginal adhesion, cell size, bare nuclei, band chromatin show a clear distinction in the values taken categorized by the output label (i.e, type of tumor). Mitoses and normal nuclei do not convey as much information as compared to the rest of the features with regard to categorization.

For the Wisconsin Diagnosis Breast Cancer (WDBC) dataset, the output label is also Malignant / Benign, but the features based on the label are different. The features, clump_thickness, size of tumor, shape of tumor, marginal adhesion, cell size, bare nuclei, band chromatin show a clear distinction in the values taken categorized by the output label (i.e, type of tumor). Mitoses and normal nuclei do not convey as much information as compared to the rest of the features.

For the Wisconsin Prognosis Breast Cancer (WPBC) dataset, the output label shows whether the tumor is recurrent or nonrecurrent. The dataset gives the corresponding time which is recurrence time if the tumor is recurrent or the disease free time if the tumor is non recurrent. As we can see from Figure 2, no conclusive features apart from tumor size, lymph status to some extent can be seen, which indicates the dataset containing 198 observation makes the task complex.

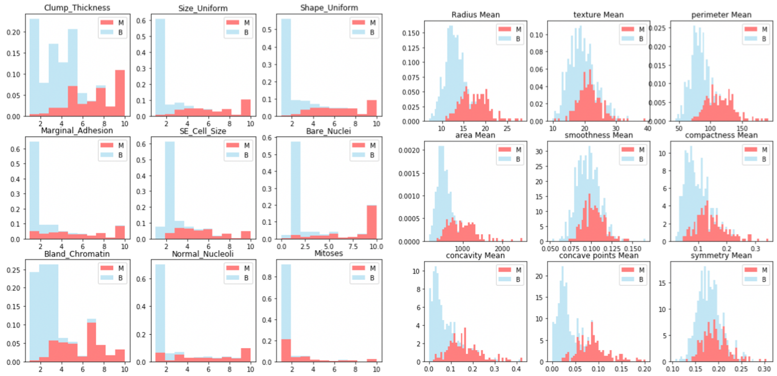

Figure 1: The left figure is WBC dataset and the right figure is WDBC dataset

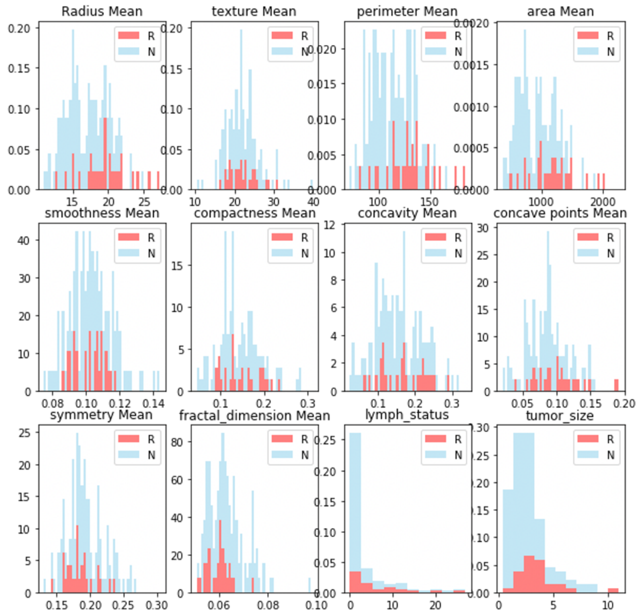

Figure 2: WPBC dataset

2.3 Training

Based on the existing feature, we compute new features using the combination of tumor size and lymph nodes. Both features are ranked from 0-10 with the severity of case. From feature analysis, we observe most of the recurrent cases have values greater than 1 for each of these features. Then, we compute new features in which we make the feature 0 (nonrecurrent) if the corresponding values in tumor size and lymph status is less than 1 and 1 if the values of tumor size and lymph status are greater than 1. We also create different features based on different criteria for value ranges of tumor size and lymph node status.

Another method we employ for visualizing whether the available features do a good job of clustering data based on specific label is PCA. PCA is used to emphasize variation and bring out distinguishable patterns in dataset that can be viewed in a 2-D plot. For each datasets, we reduce the features to two new variables namely principal components. Using PCA, we identify the 2-D plane that optimally describes the highest variance of the data. Then, the rotation of the 2-D plane gives the 2-D space and such 2-D visualization of the samples allow us to see patterns and draw qualitative conclusions about the separability based on specific conditions.

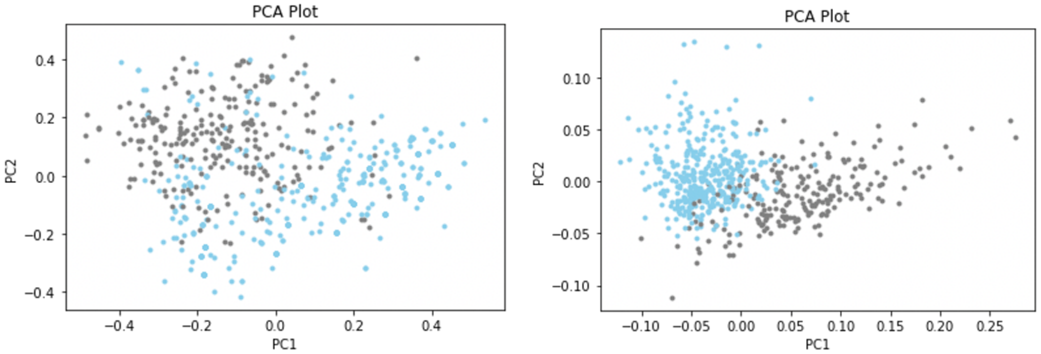

Figure 3: The left figure is PCA plot for WBC and the right figure is PCA for WDBC

From these PCA plots, for WBC and WDBC datasets, the existing features do a good job of segregating into malignant and benign cases (the blue and gray cases), but for dataset 3, the existing

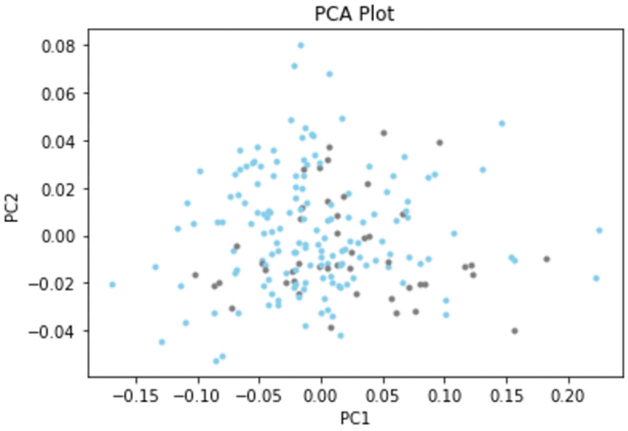

Figure 4: PCA plot for WPBC

features do not segregate the data into recurrent and nonrecurrent cases. Thus, we implement feature extraction from the existing features described above and this brings increase of accuracy.

3 Model Evaluation

Our goal is to predict whether an input instance is benign/malignant or recurrence/non- recurrence, so it makes sense to build a classification model. First, we use Logistic regression since it is useful predicting discrete levels, especially on binary class prediction.

First attribute of logistic regression is that its hypothesis class contains discrete values. In binary classification problem, there are two classes we use of "0" and "1" to represent. Second attribute of logistic regression is that the model predicts the probability of target class for one input sample and output class with highest probability. In binary classification, following expression is used:

$$P(y=1|x) = h_\theta(x) = \frac{1}{1+exp(-\theta^T x)} = \sigma(\theta^T x) \tag{1}$$

$$P(y=0|x) = 1 - P(y=1|x) = 1 - h_\theta(x) \tag{2}$$

To evaluate performance, we search for a wide range of parameters in order to find best parameters. Considerable parameters are regularization parameter, penalty type and tolerance value.

3.1 Logistic Regression

The logistic regression feature for WBC dataset are 'Clump_Thickness', 'Size_Uniform', 'Bare_Nuclei', 'Bland_Chromatin', 'Norma_Nucleoli'. The features for WDBC and WPBC datasets are mean, standard error and worst mean values for radius, perimeter, area, concavity, compactness, and concave points. Tumor size and lymph status are additional features used in WPBC dataset.

For WDBC dataset, the searching result is similar to the WBC dataset. They both reach more than 95% test score. For WPBC data set, since it does not have enough number of samples, prediction accuracy is not good as previous two. We experiment with feature extraction and using it in our model leads to marginal increase in accuracy for WPBC dataset. Also, we use the features derived from feature extraction, and it brings marginal increase of test accuracy.

3.2 Random Forest

Random forest is built by a set of decision tree classifiers, and each classifier predicts classes from input samples. Random forest combines relatively independent trees together in an average way. The word "random" indicates that when creating each decision tree, it selects random subset of features for splitting node so that the algorithm increases randomness (i.e., reduce dependencies among trees). By computing average output of all the trees, this algorithm reduces variance and generalize the output. We use random forest for each dataset.

We find that best test score is 0.9656, but the train score is 1.0. This is a typical overfit because maximum depth of tree is 40. Although the train score is less, the forest's maximum depth is 5 which indicates this forest has better performance in practice.

3.3 Gradient Boosting Machine

Gradient boosting machine is a technique for regression and classification problems that generates precise prediction model by ensemble of "weighted" weak classifiers like decision trees. A big advantage is that it allows arbitrary differentiable loss function, which makes the model generalized easily and predictable stably.

4 Results

Based on our classification model for each of the dataset and the feature analysis, important features for each of the dataset are ascertained. For the WBC dataset, the features paramount for a good accuracy are clump thickness, size and shape of nuclei, bare nucleoli and normal nuclei. For the WDBC dataset, the radius, concavity and compactness across all three measurements are important. For the WPBC dataset, tumor size and lymph status are important to some extent, so we experiment with both feature extraction and log transforming our data. Feature extraction shows marginal increase in accuracy while log transformations do not result in any appreciable increase of accuracy.

Model	**Accuracy**
Logistics Regression	93.7%
Decision Tree	93.1%
Gradient Boosting Machine	93.1%
Random Forest	93.1%

Table 1: WBC dataset

Model	**Accuracy**
Logistics Regression	96.5%
Decision Tree	93.0%
Gradient Boosting Machine	95.8%
Random Forest	97.9%

Table 2: WDBC dataset

Model	**Accuracy**
Logistics Regression	74.2%
Decision Tree	62.8%
Gradient Boosting Machine	77.1%
Random Forest	80.0%

Table 3: WPBC dataset

5 Conclusion

As the features for WBC and WDBC datasets represent the respective clusters well, which are analyzed in exploratory analysis and through the histograms and PCA plots, the accuracy rate for these datasets are high, and random forest gives the best results. For the WPBC dataset, he features do not represent the respective categories efficiently, decreasing the accuracy. We utilize feature extraction from existing features for WPBC dataset but it only brings in marginal increase of the accuracy. One of the main reasons for this less accuracy and less marginal increase is the small size of dataset and the test dataset. Also, another reason is some recurrent records have feature values that are almost the same as the values in nonrecurrent cases, leading to incorrect predictions.

Acknowledgments

We would like to acknowledge support for this project from DS-GA 1004 Machine Learning and Computational Statistics class by Professor David Rosenberg and multiple help from Brian D' alessandro and Brett Bernstein.

References

[1] Mayo Clinic Staff, “Breast cancer.” Mayo Clinic. June 09, 2011

[2] Sujana Movva,“What Causes Breast Cancer?” WebMD Medical Reference. April 26, 2015

[3] Lyon IAfRoC: World Cancer Report. International Agency for Research on Cancer Press 2003:188-193.

[4] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998

[5] Gouda I. Salama, M.B.Abdelhalim, and Magdy Abd-elghany Zeid, Breast Cancer Diagnosis on Three Different Datasets Using Multi Classifiers. International Journal of Computer and Information Technology (2277 – 0764) Volume 01– Issue 01, September 2012

[6] W. N. Street, O. L. Mangasarian, and W.H. Wolberg, An inductive learning approach to prognostic prediction. In A. Prieditis and S. Russell, editors, Proceedings of the Twelfth International Conference on Machine Learning, pages 522–530, San Francisco, 1995. Morgan Kaufmann.

[7] W.N. Street, W.H. Wolberg and O.L. Mangasarian, Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.

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

[9] W.H. Wolberg, W.N. Street, and O.L. Mangasarian, Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994) 163-171.

[10] W.H. Wolberg, W.N. Street, and O.L. Mangasarian, Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Analytical and Quantitative Cytology and Histology, Vol. 17 No. 2, pages 77-87, April 1995.

[11] 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.

[12] 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.