Using predictive models for Tumor Classification and Recurrence

Anand Bhave, Tony Seongwoo Han, Hongji Li

Introduction

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.

Our aim is to develop a model built from features 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 3 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.

Data Set

1. Wisconsin Breast Cancer (WBC) consists of visually assessed nuclear features of fine needle aspirates (FNAs) taken from patients. Tumor malignancy or benign diagnosis (Label attribute) is determined by performing a biopsy.
2. Wisconsin Diagnosis Breast Cancer (WDBC) datasets contains characteristics of the cell nuclei computed from a digitized image of a breast mass FNA. It has a label field which enlist whether a tumor is benign or malignant and a distinct set of features from the WBC dataset.
3. Wisconsin Prognosis Breast Cancer (WPBC) dataset contains information about whether a tumor is recurrent, the corresponding period of time elapsed and information pertaining to attributes like tumor size and lymph nodes.

References

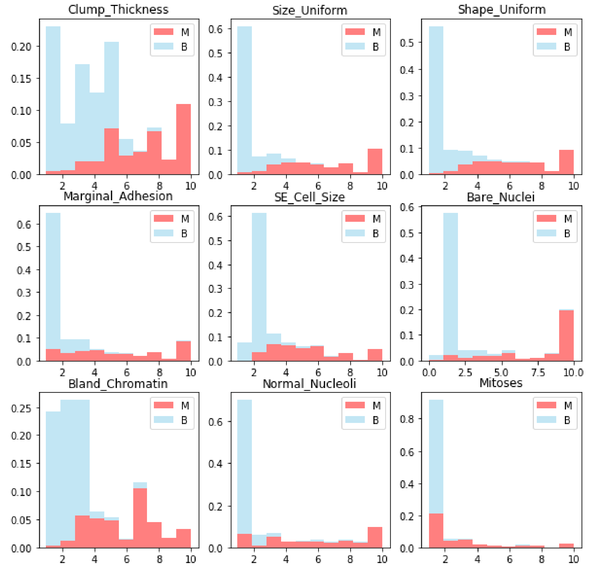
[1] Gouda I. Salama, M.B.Abdelhalim, and Magdy Abdelghany Zeid. Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers

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

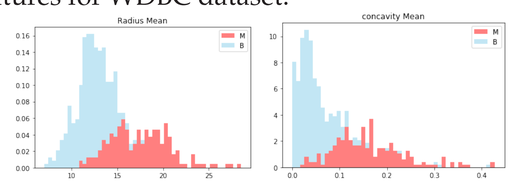
Models

For WBC, WDBC and WPBC datasets, based on the classification labels, histograms were plotted for different attributes for analyzing their value range for both malignant and benign tumors and recurrent and non-recurrent labels.

For WPBC dataset, Time, lymph nodes and tumor size are some of the important features. Following are some of the plots of important features for WBC dataset:

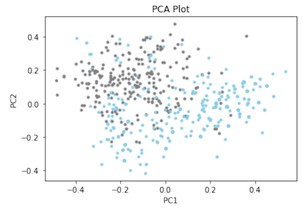


Following are some of the plots of important features for WDBC dataset:



For visualizing or target cluster labels across three datasets and to analyze how efficiently the existing features represent a particular label, we have performed PCA and following are the scatter plots.

Dataset 1



Dataset 2

Dataset 3

From these plots, datasets 1 and 2 show separation of benign and malignant cases based on existing features, while for dataset 3, clustering into distinct groups is not observed.

Now, through a rough idea about important features, we start with logistic regression by using a subset of features and then proceed to decision tree classifier where relative importance of features used is ascertained. Then we proceed towards gradient boosting classifier and random forests. Each of model incorporates 10 fold cross validation.

Future works

Firstly, it is possible to discover new features from fine needle aspirate(FNA) image.

Secondly, because detecting precision is definitely important in medical care field, we can develop more complicated ways to evaluate our system performance, such as evaluating detection rate in positive samples and negative samples.

Features

For the WPBC and WDBC datasets, each specified attribute has three measurements, namely the mean, standard error and the worst mean.

Th attributes in WPBC and WDBC datasets describe the characteristics of cell nuclei through features like radius, perimeter, area, texture, smoothness, concavity, compactness of tumor cell, etc.

Some of the attributes in WBC dataset are clump thickness, uniformity of cell size/shape, marginal adhesion, bare nuclei, bland chromatin.

By analyzing correlations, features like radius, area and perimeter are correlated the most while concavity, compactness and concave_points are correlated the most.

Results

Test accuracy for WBC data set:

Model	Accuracy
Logistic	93.1%
Decision Tree	93.7%
GBM	93.1%
RF	94.2%

Test accuracy for WDBC data set:

Model	Accuracy
Logistic	96.5%
Decision Tree	94.4%
GBM	97.2%
RF	97.2%

Test accuracy for WPBC data set:

Model	Accuracy
Logistic	92%
Decision Tree	86%
GBM	92%
RF	94%