STAT5330 Final Project

----Breast Cancer Diagnosis in Wisconsin

Jingying Pan jp9pbd
4/23/2020

Abstract

This project studies the Wisconsin Prognostic Breast Cancer. The aim is to predict breast cancer diagnosis using the digitized image of a fine needle aspirate (FNA) of a breast mass. This project also focuses to find the optimal classification model that maximize the accuracy of the prediction. LDA, QDA, Logistic, SVM, Random Forest and Boosting are used for the classification analysis.

1. Introduction

Breast cancer is cancer that develops from breast tissue. Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, a newly-inverted nipple, or a red or scaly patch of skin. In those with distant spread of the disease, there may be bone pain, swollen lymph nodes, shortness of breath, or yellow skin. Most types of breast cancer are diagnosed by microscopic analysis of a sample - or biopsy - of the affected area of the breast. Also, there are types of breast cancer that require specialized lab exams. This project provide a method to diagnose breast cancer with the digitalized image of FNA of a breast mass. With this project we can construct a method that can be completed by computer, other than human beings.

2. Data Description

2.1 Data Source and Collection

The last 30 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. A few of the images can be found at

http://www.cs.wisc.edu/~street/images/

This database is also available through the UW CS ftp server:

ftp ftp.cs.wisc.edu

cd math-prog/cpo-dataset/machine-learn/WDBC/

Also can be found on UCI Machine Learning Repository:

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

2.2 Variables

This dataset contains 31 variables among which there are 30 continuous predictors and 1 categorical response. 659 observations are collected and we decided to use the first 500 observations as the training dataset and the rest as the testing dataset. The variables information in this dataset are as follows:

- diagnosis : The diagnosis of breast tissues (M=malignant, B=benign)
- radius_mean : mean of distances from center to points on the perimeter
- texture_mean : standard deviation of gray-scale values
- perimeter_mean : mean size of the core tumor
- area_mean
- smoothness_mean : mean of local variation in radius lengths
- compactness_mean : mean of perimeter^2/area-1.0
- concavity_mean : mean of severity of concave portions of the contour
- concave points_mean : mean for number of concave portions of the contour
- symmetry_mean
- fractal_dimension_mean : mean for "coastline approximation"-1.0
- Column 12-31 : the standard error and "worst" or largest of these features computed for each image

2.3 Fine Needle Aspiration (FNA)

Fine-needle aspiration (FNA) is a diagnostic procedure used to investigate lumps or masses. In this technique, a thin (23–25 gauge), hollow needle is inserted into the mass for sampling of cells that, after being stained, are examined under a microscope (biopsy). The sampling and biopsy considered together are called fine-needle aspiration biopsy (FNAB) or fine-needle aspiration cytology (FNAC) (the latter to emphasize that any aspiration biopsy involves cytopathology, not

histopathology). Fine-needle aspiration biopsies are very safe minor surgical procedures. Often, a major surgical (excisional or open) biopsy can be avoided by performing a needle aspiration biopsy instead, eliminating the need for hospitalization. In 1981, the first fine-needle aspiration biopsy in the United States was done at Maimonides Medical Center. Today, this procedure is widely used in the diagnosis of cancer and inflammatory conditions.

3. Exploratory Data Analysis

3.1 Response consistance

The response contains 63% malignant and 37% benign, shown as follows:

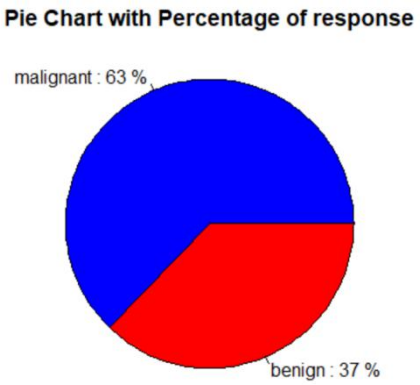

Figure 1 percentage of response

3.2 Correlation between variables

Then I'd like to explore the correlations between variables. The variables in columns 2 to 11 is the mean value of the features, as shown in section 2.2, columns 12 to 31 is the standard error and "worst" or largest of these features computed for each image. So it is reasonable to seek the correlation between variables 2 to 11, 12 to 21 and 22 to 31.

In order to check the correlations between the mean value of the features we removed the obvious correlated variables(such as radius, perimeter and area, we moved the perimeter and area). The correlations between mean value of the features are as follows:

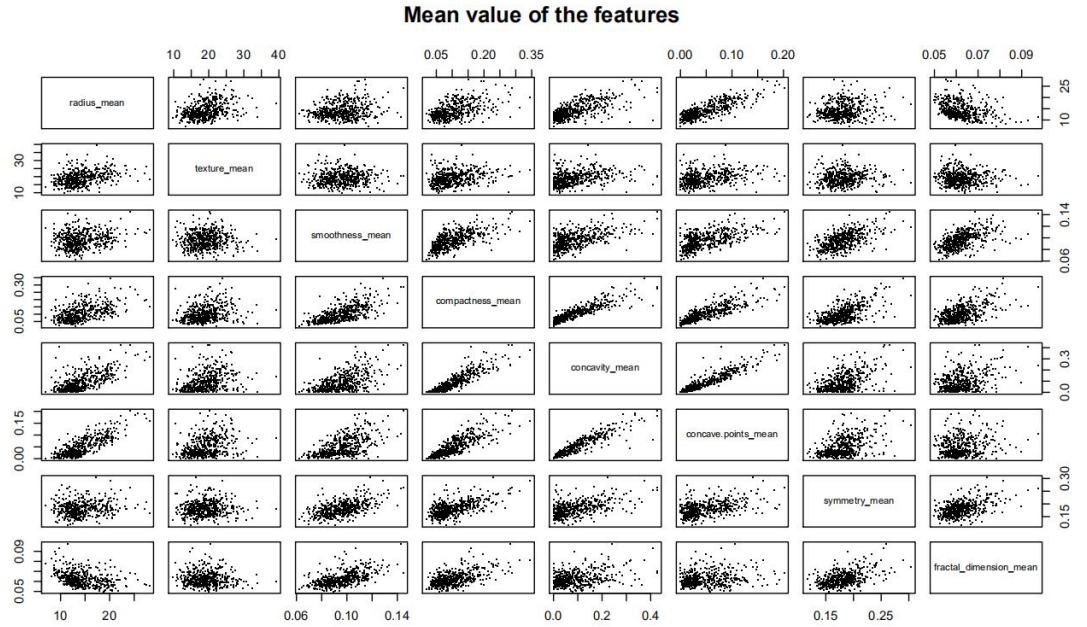

Figure 2 correlations of mean of features

From the plots we can see that,the compactness_mean, concavity_mean and concave.points_mean have strong relationship. Other than that,the concave points and radius also have strong relationship.

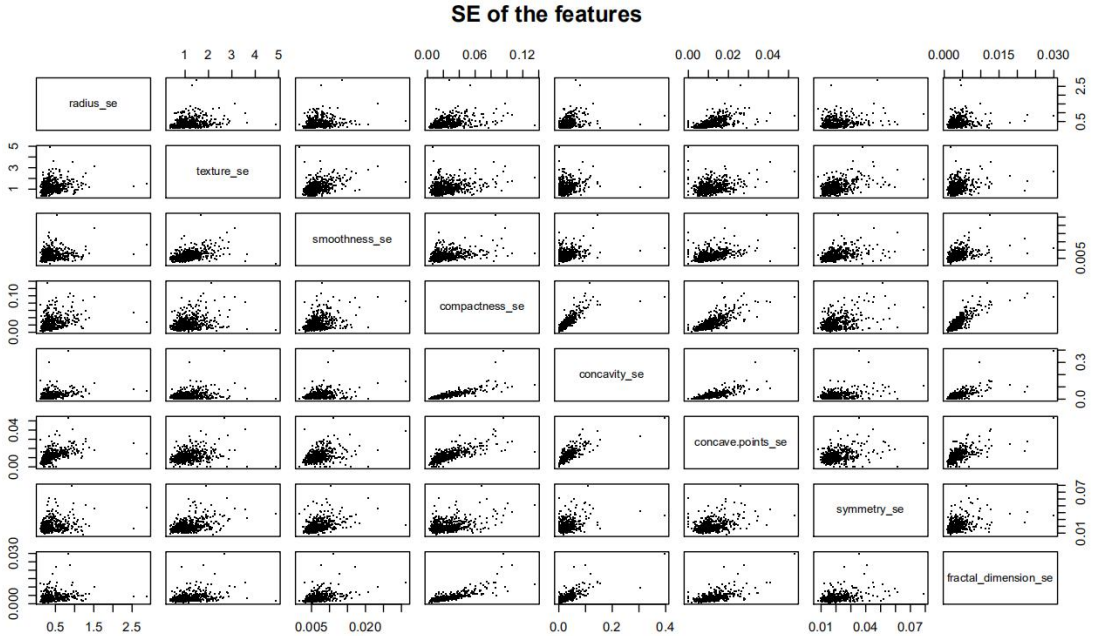

Figure 3 correlations of se of features

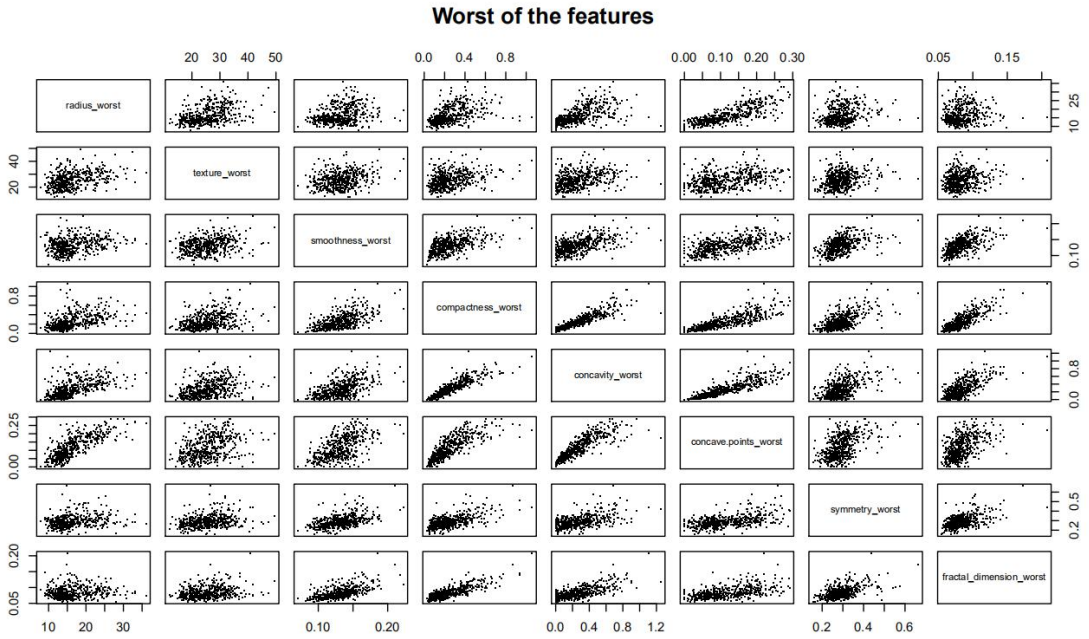

Figure 4 correlations of worst of features

From the pairs plots of the SE and the worst of the features(again,perimeter and area removed),we can see that the compactness_se, concavity_se and concave.points_se have correlations. The compactness se and fractural_dimension also have correlations. The worst of the features also got similar results.

4. Regression

To simplify the classification, we have converted the response from malignant - benign into 1 - 0. By doing that, we can make our model easier to interpret. Other than that, we have removed the perimeter and area for all the features (radius retained). All of our predictors are continuous and our response is binary. So in our regression analysis, we have 24 continuous predictors and 1 categorical response. There are many classification methods, we used six of them to construct our model: LDA, QDA, Logistic, SVM, Random Forest and Boosting.

4.1 LDA

The Linear Discriminant Analysis (LDA) classifier results from assuming that the observations within each class come from a normal distribution with a class-specific mean vector and a common variance σ^2, and plugging estimates for these parameters into the Bayes classifier. We

firstly used LDA to fit our data, the accuracy is 95.65%, the sensitivity is 96.23%, the specificity is 93.75%.

4.2 QDA

LDA assumes that the observations within each class are drawn from a multivariate Gaussian distribution with a classspecific mean vector and a covariance matrix that is common to all K classes. Quadratic discriminant analysis (QDA) provides an alternative quadratic discriminant Analysis approach. Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes' theorem in order to perform prediction. However, unlike LDA, QDA assumes that each class has its own covariance matrix.

We then apply the QDA classification on our dataset, the accuracy is 95.65%, the sensitivity is 100%, the specificity is 85%.

4.3 Logistic

For the breast cancer prediction data, logistic regression models the probability of X. In this project, the probability of default given balance can be written as

$$\Pr(Y = 1|X).$$

The values of $\Pr(Y = 1|X)$, which we abbreviate $p(X)$, will range between 0 and 1. Then for any given value of X, a prediction can be made for default. With 0.5 as our threshold, we might predict $Y = 1$ for any individual for whom $p(\text{balance}) > 0.5$. Logistic classification model assumes equality of conditional probability and linear combination of independent variables transformed by the logistic function. The use of transformation can help guarantee that linear combination of the independent variables would fall in range (0, 1). We perform a logistic regression on our training dataset and predict the model on the testing dataset. The accuracy is 97.1%, the sensitivity is 100%, the specificity is 89.47%.

4.4 SVM

Support Vector Machine (SVM) is a supervised learning model that incorporates associated learning algorithms that optimizes classification problems. SVM was developed in the computer science community in the 1990s and that has grown in popularity since then. SVM have been shown to perform well in a variety of settings, and are often considered one of the best "out of the box" classifiers. The support vector machine is a generalization of a simple and intuitive classifier called the maximal margin classifier. We have tried both linear SVM and non-linear SVM in our analysis.

With linear SVM, the accuracy is 96.1%, the sensitivity is 98.08%, the specificity is 94.12%. With non-linear SVM, the accuracy is 97.89%, the sensitivity is 98.65%, the specificity is 92.18%. From the result we can see that effect of our predictors show signs of nonlinearity.

4.5 Random Forest

After tuning with cross validation, we find that the accuracy is highest when nodesize is 4(as shown in Figure 5). We then ran a random forest model with tree number 1000 and nodesize 4, the model reaches accuracy of 94.80%, sensitivity of 93.47%, and specificity of 92.14% in the testing set.

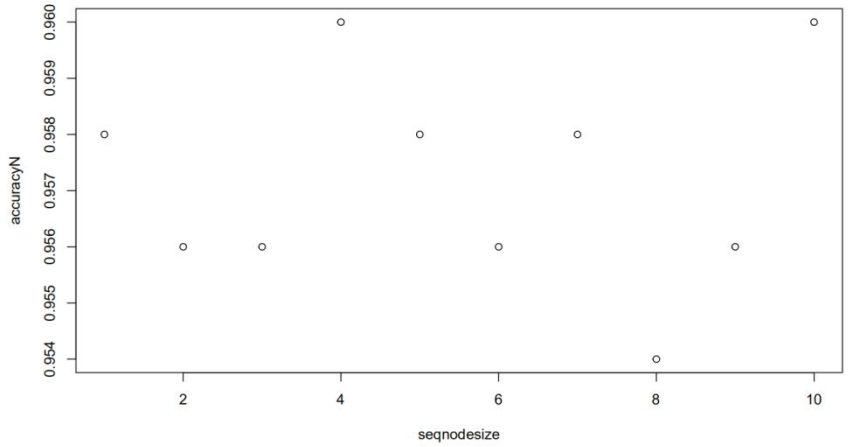

Figure 5 effect of nodesize

After fitting the data with random forest model, we also checked the importance of each variable, as shown in Figure 6, radius_worst is the most important factor that would affect the diagnosis. And ave.points_worst and ave.points_mean also have important effect on the diagnosis. In all the

features, smoothness does not have much effect on diagnosis (the mean, se, worst of smoothness). Symmetry_mean, symmetry_se and concave.points_se have little effect on the diagnosis.

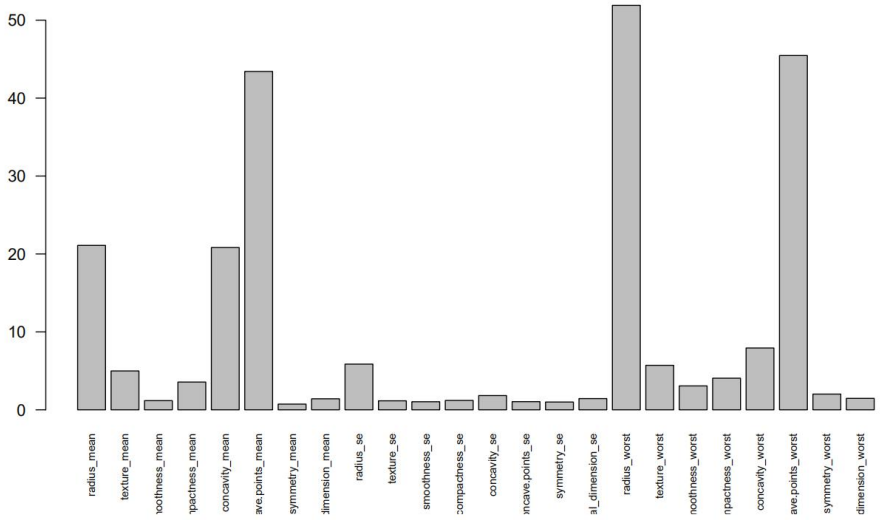

Figure 6 importance of each variable

4.6 Boosting

Boosting is another approach for improving the predictions resulting from a decision tree.Like bagging, boosting is a general approach that can be applied to many statistical learning methods for regression or classification. Here we restrict our discussion of boosting to the context of decision trees. We firstly tune the number of trees and shrinkage factor by five-folder cross-validation. The tuned number of trees is 648 (as shown in Figure 7). We set 0.5 as the threshold, and with this model we get the accuracy is 98.55%, the sensitivity is 98.11%, the specificity is 99.88%.

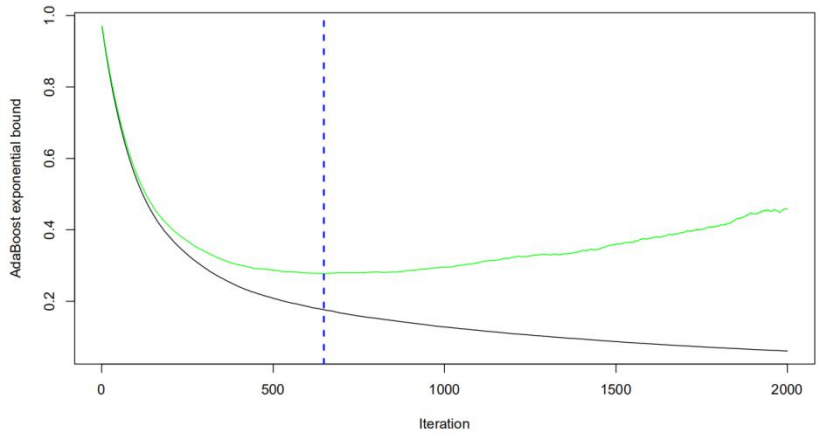

Figure 7 effect of number of trees

5. Conclusions and Limitations

In all the classification methods above, the Logistic Regression got the highest accuracy and the highest sensitivity. The boosting regression got the highest specificity. From the importance analysis we can see that most variables have small effect on the diagnosis. All of the accuracy are above 95% which indicates diagnosis prediction from the digitalized image is reliable and efficient.

With this method, there are still some cases that can not be diagnosed correctly. The worst case is that one got malignant breast cancer but is diagnosed as benign. This would mislead one to miss the best time for treatment.

6. Future Work

Breast cancers are classified by several grading systems. Each of these influences the prognosis and can affect treatment response. In the future, we can try to classify the breast mass into different stages, which can help doctors to decide the treatment efficiently.