CS6300 Final Project
Breast Cancer Diagnosis

Na Song

School of Computing, Clemson University

nsong@g.clemson.edu

***Abstract*—Using the machine learning, statistics, image processing and optimization methods, highly accurate diagnosis of breast can be done even by untrained users. Base on the the UCI Breast Cancer Wisconsin Data Set, random forest algorithm, SVM with linear kernel, SVM with radial basis kernel and recursive feature elimination method are used to classify the diagnosis result of Breast Cancer. The accuracy of random forest, SVM and recursive feature elimination method on the test set are 94.74%, 97.66%, 97.08% and 96.23%, SVM with linear kernel performs best on diagnosis of breast concer.**

***Keywords*— Random Forest, SVM, Recursive Feature elimination, Accuracy.**

I. Introduction

According to the Centers for Disease Control and Prevention(CDC), breast cancer is the most common type of cancer for women regardless of race and ethnicity(CDC,2016). Around 220,000 women are diagnosed with breast cancer each year in the United States (CDC, 2016).

The diagnosis of breast tumors has traditionally been performed by a full biopsy an invasive surgical procedure, in the recent years, by using and extending the results from the fields of machine learning, statistics, image processing and optimization, highly accurate diagnosis of breast can be done even by untrained users[1]. Using software system helps doctors to make accurate judgment is the tendency of the cancer diagnosis.

In this article, some machine learning algorithms will be applied to already processed images to get accurate judgment.

II. Dataset

The dataset is come from UCI Machine Learning Dataset, which called Breast Cancer Wisconsin (Diagnostic) Data Set[2-4]. There are 569 instances in the dataset, and for each instances, there are 32 attributes to describe it. All these 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. There are both qualitative and quantitative features in the dataset. All feature values (for quantitative features) are recorded with four significant digits. As for the class distribution, there are 357 benign and 212 malignant.

Here is a brief introduction of each attribute.

- ID number

The ID of each sample.

- Diagnosis

M means malignant, B means benign. This attribute is the output.

The rest of the attribute are features, and they are all quantitative parameters. 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, field 23 is Worst Radius.

- radius

mean of distances from center to points on the perimeter. radius_mean, radius_se, radius_worst are the mean, the standard error and the worst of the radius.

- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness

$$compactness = \frac{perimeter^2}{area} - 1.0$$

- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry

- fractal dimension ("coastline approximation" - 1)

III. Exploratory Data

A. *Overview Of Dataset.*

First of all, we count the diagnosis results as Fig.1.

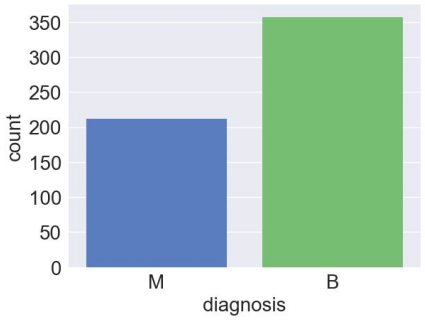

Fig.1 count for the diagnosis results.

There are 357 number of Benign and 212 number of Malignant.

To capture the structure of our data, and better understand the distinctions between categories of the dataset, I implemented two well-known algorithms: principal component analysis(PCA) and t-distributed stochastic neighbor embedding analysis(t-SNE). Figure 2 displays the projection of our dataset onto a two dimensional plane using the first two principal components which get from the standardised data, the left panel is the PCA scatter plot and the right panel is the t-SNE scatter plot.

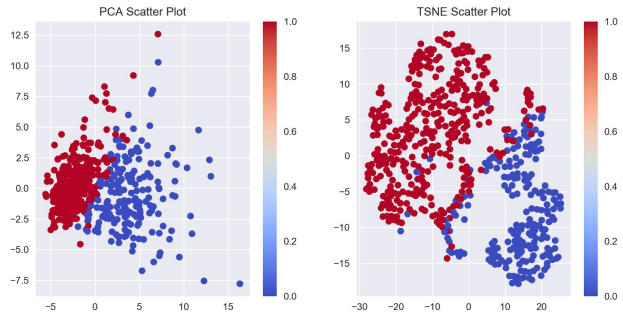

Fig.2 2D projection of the data with PCA and t-SNE

These two components explain 44.27% and 18.97% of the overall variance in the data, total 63%, PCA can only represent the structure of the data through linear subspaces. Alternatively, the t-SNE algorithm can capture interesting non-linear paths and hence. From Figure 2 we can see that t-SNE performs better.

In order to explore the influence of each feature on the output, at the first step, checking the difference in each feature value in different class.

Since the differences between values of features are very high to observe on plot, so before plotting, we need to normalization or standardization. Here, we use the function below to normalize the values.

$$data\_norm = \frac{data - data.mean}{data.std}$$

Then dividing the 30 features into 3 groups by mean, standard error and worst. And plotting the value and the classification of each group as Fig.2.

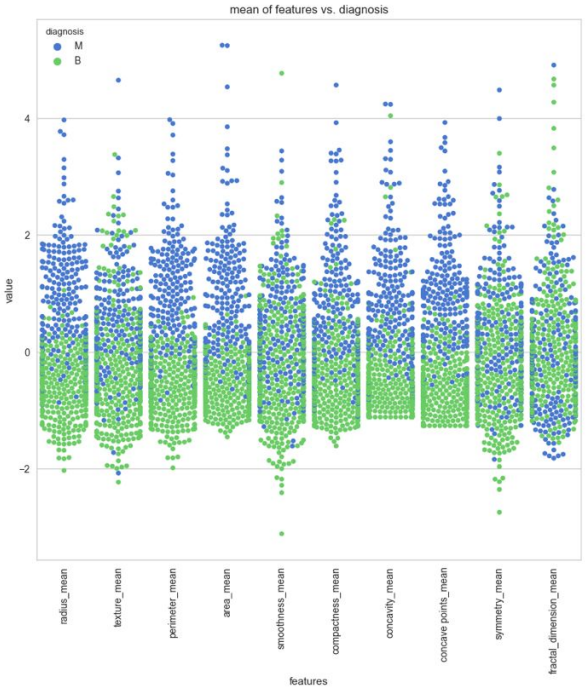

Fig.3 value distribution vs. the classification based on each feature.

From the plot above, we can find some useful information about features, for example, on the one hand, in perimeter_mean feature, median of Malignant and Benign looks like separated so it can be good for classification. However, in fractal_dimension_mean feature, mean of the Malignant and Benign does not looks like separated so it does not give good information for classification. On the other hand, we can see that the value distributions of radius_mean and perimeter_mean look similar with each other, which inspire us to explore whether they are correlated with each other or not.

And we can plot the same swarm plot for the other two groups: the standard error and the worst group.

B. *Correlation Of Features*

As we can see in Fig.2 that there are some features which are similar with each other. In order to compare these feature deeper, we use Pearson R value to access the correlation, where 1 is the highest. In Fig.3, we give an example.

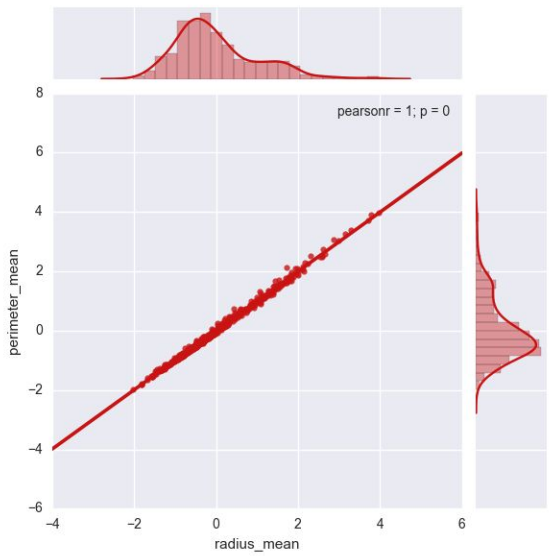

Fig.4 Pearson R value of radius_mean and perimeter_mean.

The pearson R value between radius_mean and perimeter_mean is 1, so they are high correlated.

One step further, in order to see the observe all correlation between features, the heatmap of the Pearson R value can be used to show the correlation between each features as Fig.4 shows.

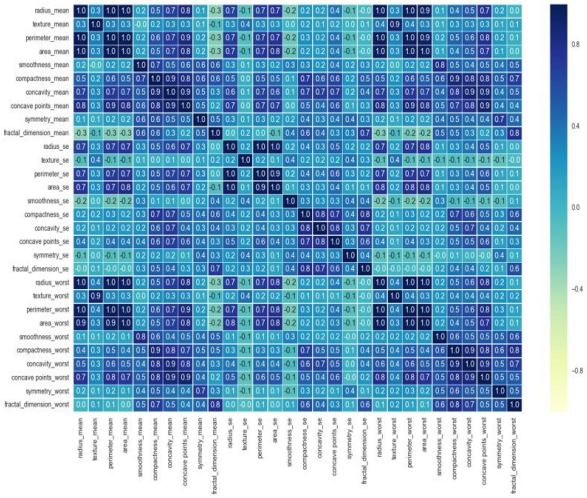

Fig.5 correlation heatmap for all features.

As it can be seen in heat map figure, **radius_mean**, **perimeter_mean** and **area_mean** are correlated with each other so we will use only one of them. Here, summarizing the correlated features in the table1.

features	correlation
radius_mean, perimeter_mean, area_mean	1.0
compactness_mean, concavity_mean **concavity_mean, concave points_mean** **compactness_mean, concave points_mean**	0.9 0.9 0.8
radius_se, perimeter_se, area_se	1.0
radius_worst, perimeter_worst, area_worst	1.0
Compactness_se, concavity_se **concavity_se, concave points_se** **compactness_se, concave points_se**	0.8 0.8 0.7
compactness_worst, concavity_worst **concavity_worst, concave points_worst** **compactness_worst, concave points_worst**	0.9 0.9 0.8
Texture_mean, texture_worst	0.9
Area_worst, area_mean	1.0

Table 1 high correlation value table

For the features that correlated with each other (the correlation value is 1.0), we can choose one of them to represent them, that means we can drop **radius_mean, perimeter_mean, radius_se, perimeter_se, radius_worst, perimeter_worst and area_worst.**

IV. Model Selection and Validation

Our main objective is to construct a highly accurate classifier that generalizes well on data from new individuals. For this purpose, we have tested the performance of different classifiers. Based on the features, we mainly divided these methode into two category. The one just simply drop all the other features that have high correlation with each other, and retain one of them. For the other category, we just use all the feature that could provide information, and training the model and selecting features at the same time.

A. *Models with simply dropping features.*

In this part of models, we just drop the features that have high correlation values with each other. **compactness_mean, concavity_mean** and **concave points_mean** are correlated with each other. So we only choose **concavity_mean**. Within **Compactness_worst, concavity_worst** and **concave points_worst** we use **concavity_worst**. Among **compactness_se**, **concavity_se** and **concave points_se** we use **concavity_se**. **texture_mean** and **texture_worst** are correlated and we use **texture_mean**.

In order to quantify the relationship between the distribution of the features' values with the diagnosis result(M/B), Fisher's score is used to show the distinction of different features. Since we have removed some of the features in the previous step, so that they will not be considered.

If we define the values of feature T from samples ,whose diagnosis is $i \in \{0, 1\}$ (0 stands for 'B' and 1 stands for 'M'), is $y_{i,T}$, then we have :

$$m_{i,T} = \frac{1}{N_{i,T}} \sum_{samples} y_{i,T} \quad (1)$$

$$S_{i,T} = \frac{1}{N_{i,T}} \sum_{samples} (y_{i,T} - m_{i,T})^2 \quad (2)$$

$$J_T = \frac{(m_{0,T} - m_{1,T})^2}{S_{0,T} + S_{1,T}} \quad (3)$$

Here J_T is the fisher score. As we can see from the definition, higher fisher score always means better performance of a feature to distinguish different diagnosis results. Figure 5 shows the fisher scores of all left features.

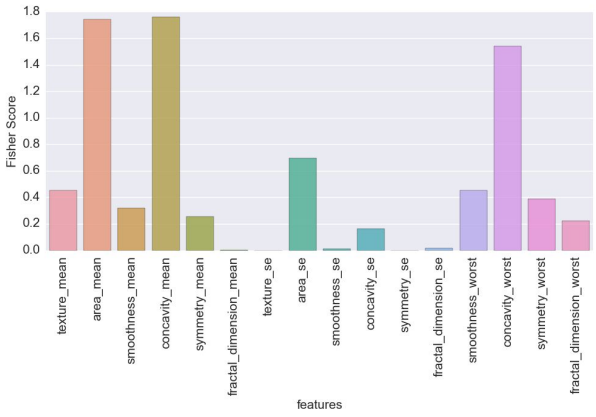

Fig.6 Fisher scores of all left features

There are several features whose fisher score are very low, which means they are less relevant to diagnosis results. As a result, we choose to drop half of the features whose fisher scores are lower than the median value of all. Finally 8 features remain : **texture_mean, area_mean, smoothness_mean, concavity_mean, area_se, smoothness_worst, concavity_worst** and **symmetry_worst**.

B. *Training and testing*

There are a lot of algorithms that can be used to do classification work, like SVM, random forest, bayesian algorithms, adaboost, neural network and so on. As we have already filtered out features with high relevance to cancer diagnosis, classical methods are suitable for this problem. Here we use SVM with radial basis kernel and linear kernel, random forest classifier and adaboost classifier. The algorithms are implemented in 'sklearn' python package.

We separate the whole dataset into two parts : training set (70%) and testing set (30%). All of the models in this passage share the same training and testing set.

- SVM(Support vector machines)[5]-As we can see in Figure. 2, the two cluster are partially overlap.Therefore, we would expect maximizing margins when separating these activities to result in good performance. We experimented with linear and radial-basis kernels, tuning each model and evaluating their performance.
- Random forest[6]-A random forest is a meto estimator that fits a number of decision tree classifiers on various sub-samples of the dataset, averaging them to improve the predictive accuracy and control over-fitting.
- Adaboost[7]- The AdaBoost classifier begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset, where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. That means the sample points near the border will get more attention in the classification, and this strategy could be a good treatment to the overlap part of Figure.2.

Figure 7 shows the classification results using the method mentioned above. In x-axis *0/1* means real *'B' /'M'*, while in y-axis *0/1* means predicted *'B' /'M'*. The accuracy of each method is shown above it's figure. We can see that in these four methods, SVM with linear kernel performs best, which gives 97.66% accuracy of the classification.

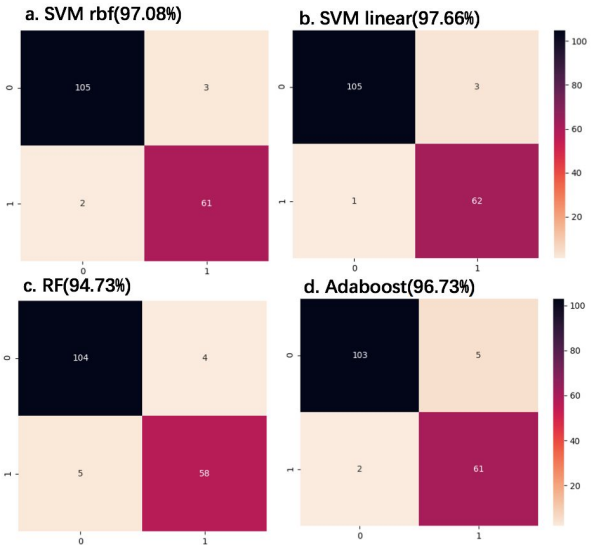

Fig.7 Classification results of SVM with radial basis kernel, SVM with linear kernel, random forest and adaboost algorithm.

C. *Models with feature selection*

In the previous part we use our own strategies to drop redundant features. However it is not guaranteed to be the most powerful. In this part, we use a proven feature selection technique to improve the performance of random forest classifier.

The feature selection algorithm is called *recursive feature elimination* (RFE). It evaluates the classification results using different combinations of features and choose the best one. On the basis of random forest classifier and the same dataset, RFE outputs the best feature selection choice, which raises the accuracy to 96.23%: **texture_mean, area_mean, concave points_mean, area_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, concavity_worst** and **concave points_worst**.

The cross validation score using different numbers of selected features are shown in figure 8.

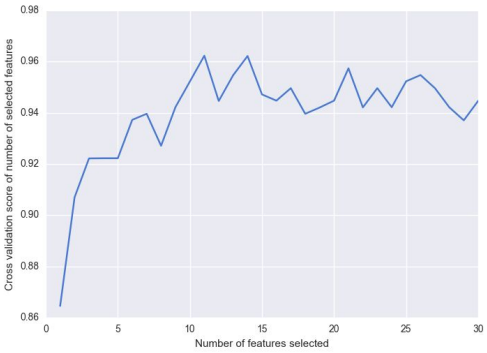

Fig.8 Cross validation score v.s. number of features selected

D. *Model Selection*

The train and test accuracy for each of our analysis are displayed below:

Machine Learning Algorithm	Test Accuracy
Random Forest	94.73%
SVM with Radial Basis Kernel	97.08%
SVM with Linear Kernel	**97.66%**
Recursive Feature Elimination	96.23%
AdaBoost Classifier	96.73%

Table 2 Accuracy of each algorithm

From the last part we can see that SVM with radial basis kernel performs best on the test set, the accuracy of it is more than 97%. The result could be helpful to assist doctor to diagnosis.

V. Conclusion

Lots of achievements have been made in recent years in machine learning field, and the algorithms are widely applied in all walks of life. The accuracy of medical imaging diagnosis obtained greatly ascend with the help of the machine learning algorithm, which provides a great help to the doctor.

In this article, Breast Cancer Wisconsin (Diagnostic) Data Set from UCI Machine Learning Dataset is used to train the diagnosis model and evaluate its accuracy. First, visualization, analysis and selection have be done on the features, so that we could find the correlation and the importance of each features. Then we use the random forest algorithm, SVM with linear and radial basis kernel and recursive feature elimination method to fit the train set. Here, I have also tried the SVM with polynomial kernel, however, since the radial basis kernel can be expanded to the infinite order polynomial, it is no wonder that polynomial kernel won't perform better that radial basis kernel. In this case, the accuracy of SVM with polynomial kernel is 94.68%. So, this method is not mentioned as one of the methods in "Model Selection" part. Finally, the test set is used to evaluate accuracy of each methods. And the accuracy of random forest, SVM with linear kernel, SVM with radial basis kernel and recursive feature elimination method on the test set are 94.74%, 97.66%, 97.08% and 96.23%. SVM with linear kernel method performs best, more than 97% accuracy will provide a great assistance to the diagnosis of the breast cancer.

REFERENCES

[1] Street, W. Nick, William H. Wolberg, and Olvi L. Mangasarian. "Nuclear feature extraction for breast tumor diagnosis." (1992): 861-870.

[2] Dr. William H. Wolberg, General Surgery Dept. University of Wisconsin, Clinical Sciences Center Madison, WI 53792.

[3] t, Computer Sciences Dept. University of Wisconsin, 1210 West Dayton St., Madison, WI 53706

[4] Olvi L. Mangasarian, Computer Sciences Dept. University of Wisconsin, 1210 West Dayton St., Madison, WI 53706

[5] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2015). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-7. https://CRAN.R-project.org/package=e1071

[6] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[7] Y. Freund, R. Schapire, "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting", 1995.