CSIC5011 Mini-Project : Visualizing Random Forest Classification for Breast Cancer Diagnosis

WONG, Kwan Long Kyle

Department of Bioengineering, HKUST

Introduction

Breast cancer is the most common cancer among women in Hong Kong, and the number of diagnosed cases continues to increase every year [1].

Previous works have shown that machine learning (ML) techniques can achieve a diagnosis performance comparable to that of an expert clinician while automating the entire process [2]. However, real-world data sets often have multiple features, resulting in multi-dimensional data that becomes difficult for analysis and visualization.

The goal of this project is to use dimensionality reduction techniques to visualize the results of an ML classification model. In particular, a random forest classifier was found to easily achieve high accuracy on a high dimensional data set.

Data

The UCI ML Breast Cancer Wisconsin (Diagnostic) Data Set (downloaded from **https://goo.gl/U2Uwz2**) was used. It is a binary classification data set which consists of samples obtained from 569 patients, labelled either 'malignant' or 'benign'. Each sample contained 10 real-valued features computed from characteristics of cell nuclei of a breast mass.

Feature Information

The mean, standard error, and extreme ("worst") value of the following attributes were provided, resulting in 30 features per sample.

1. Radius
2. Texture
3. Perimeter
4. Area
5. Smoothness
6. Compactness
7. Concavity
8. Concave points
9. Symmetry
10. Fractal dimension

Methodology

Features were first standardized by subtracting the mean and dividing by the standard deviation. A random forest classifier was trained and evaluated on a test set with ten-fold cross-validation. Feature importance values were extracted and displayed.

To visualize the results of the classification, PCA was performed to reduce the features to two dimensions. The classification result was predicted for the new feature range and plotted along with the reduced data and labels highlighted.

Code is available at: https://github.com/klwong126/CSIC5011

Discussion/Results

The random forest classifier achieved a cross-validated mean accuracy of 0.94 and RMSE of 0.057. According to the computed feature importance values (feature importance plot found in code), the two most important features are 'worst area' and 'worst concave points'. However, it should be noted that random forest feature importance measures may not always be reliable, and further validation is needed to support this claim [3].

The new random forest trained on the PCA-reduced features achieved a similar cross-validated mean accuracy (~0.94) with an RMSE of 0.056 . This suggests that PCA is a promising technique for dimensionality reduction of this data set, which allows for a better and more intuitive visualization of the classification results (Figure 1).

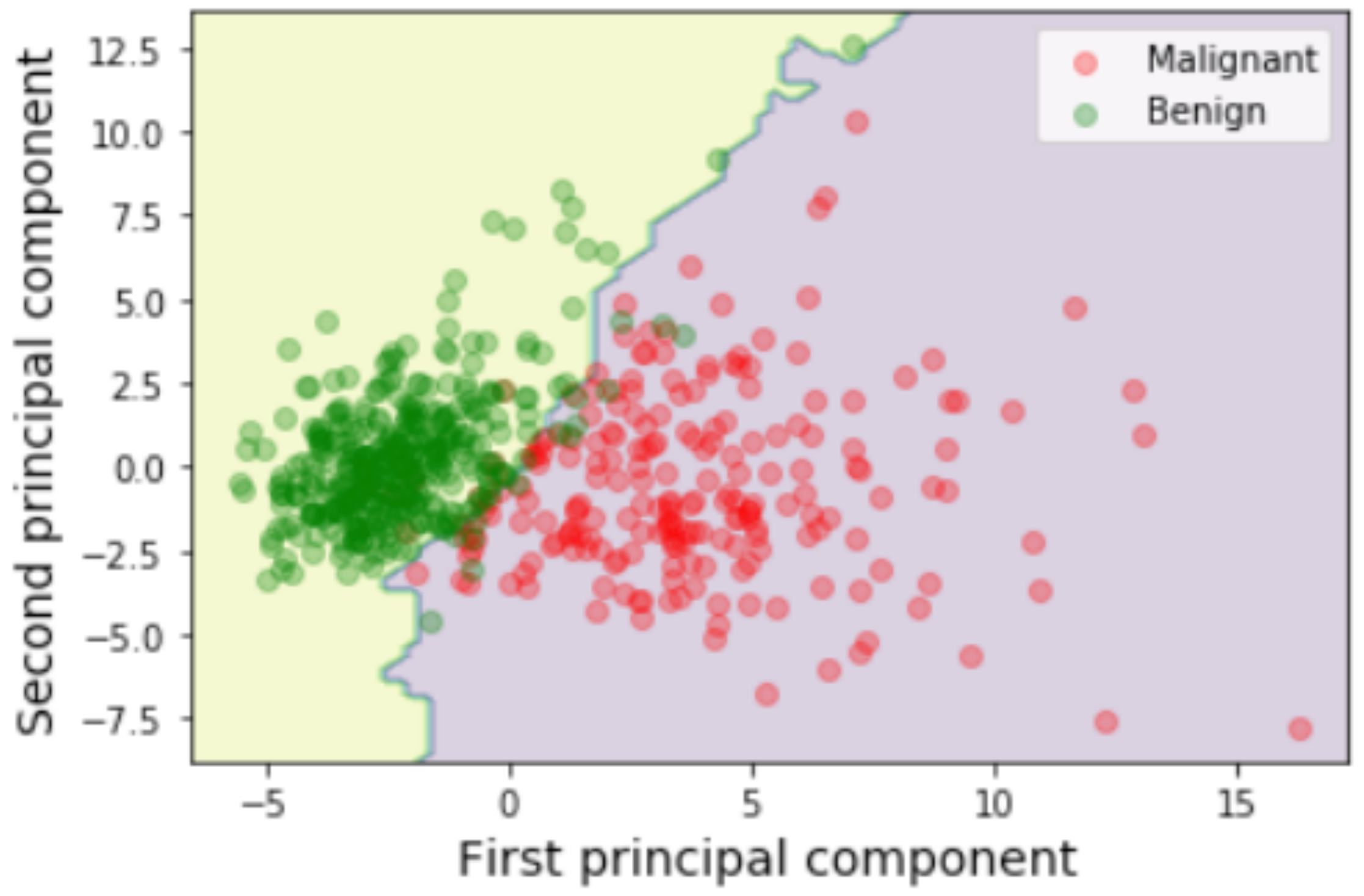

Figure 1: Results of Random Forest Classification in Principal Component Coordinates

Conclusion

PCA based dimensionality reduction using Singular Value Decomposition was applied to the problem of prediction of breast cancer diagnosis from multidimensional features of cell characteristics. The random forest classification result on the reduced 2D feature range achieved comparable accuracy to the original and provided a clear visualization of the decision boundary.

Future Work

It will be interesting to apply this methodology to visualize and compare the classification results and decision boundary of other models such as neural networks, SVMs, and linear models. Given the heterogeneous nature of medical data, further discussion is needed to consider the biological significance of certain features, to select an optimal model that requires minimal features while preserving accuracy.

References

[1] Breast Health | Hong Kong Breast Cancer Foundation - Local Statistics. (2020).

[2] Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer letters*, *77*(2-3), 163-171.

[3] Strobl, C., Boulesteix, A., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1). doi: 10.1186/1471-2105-8-25

[4] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.