Predicting Breast Cancer Prognosis Using Genetic Data

Ziyuan Shen

September 25, 2019

Abstract

Machine learning techniques have a wide range of applications in healthcare and disease studies. With the advancement of biochips, genetic expression data can be obtained from thousands of genes. In this report, we design and implement classifying estimators to predict whether a patient diagnosed of breast cancer will have good prognosis or bad prognosis. Traditional classifiers as well as a novel approach are compared and cross validation is employed to evaluate the performance. Such prediction conclusion has the potential to aid oncologists with personalized and precision medicine.

1 Introduction

Over the past few decades, machine learning techniques have been increasingly applied to disease studies [1]. The development of single-cell technology greatly facilitates the availability of multi-omics data [2], including genetic data. Accordingly, delicately designed data mining and machine learning techniques are desired to explore the biological information underlying the large datasets. Relevant studies can aid disease diagnosis and precision medicine. Cancer classification is one of the top promising topics [3]. By far, clinical datasets, genetic datasets [4, 5] as well as integrated datasets have been used for cancer classification studies. Improving the classification performance is undoubtedly meaningful as human beings' overall health conditions will be improved by these efforts.

Although clinical data, such as patient information and tumor appearance, can certainly be employed for cancer classification, a biological intuition is that cancer subtypes are more dependent on patients' genetic components. In fact, patients with similar clinical appearance do have completely different therapy responses [6]. Therefore, genetic datasets should play an important role in cancer class discovery and prediction. Different from other kinds of data, genetic data is cursed by high dimensionality due to the large number of gene types. Prediction models with high accuracy are aimed to aid oncologists with diagnosis and prognosis. Our work is based on the paper "Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers" [7]. The goal is to create a good predictor with samples from known classes, and possibly identify hidden gene expression patterns related to each class, without many biological priors.

2 Data Description

High through-put sequencing (HPS) data of Breast Invasive Carcinoma (BRCA) is downloaded from Broad Institute GDAC Firehose [8]. The entire dataset is composed of 189 patient samples. All the samples are labelled good prognosis if the patient survived more than 5 years, otherwise they are labelled poor prognosis. Gene interaction network data is downloaded from Reactome [9, 10].

2.1 Summary Statistics

The original dataset has 189 samples, 90 of which are poor samples and 99 of which are good samples. Each sample has 12027 gene expression values (gene types). The gene interaction network consists of

12175 different genes. Every two related genes are connected by an edge and there are 229285 edges in total. As the two datasets are collected separately, they include different sets of gene types. In order to incorporate the information, preprocessing the data requires finding common genes in the expression dataset and the network dataset. Statistics are summarized in Table 1.

Table 1: Summary Statistics of Breast Cancer Genetic Datasets.

Datasets Statistics	Original	Restricted
#Total Samples	189	
#Poor Samples	90	
#Good Samples	99	
#Genes	12027	8819
#Network Genes	12175	8819
#Network Edges	229285	150168

2.2 Data Visualization

As the data is very high dimensional, some dimension reduction technique is required for visualizing the data. Here we employ principle component analysis (PCA). Note that normalization should be conducted before PCA. The first three PCA components are visualized in Figure 1.

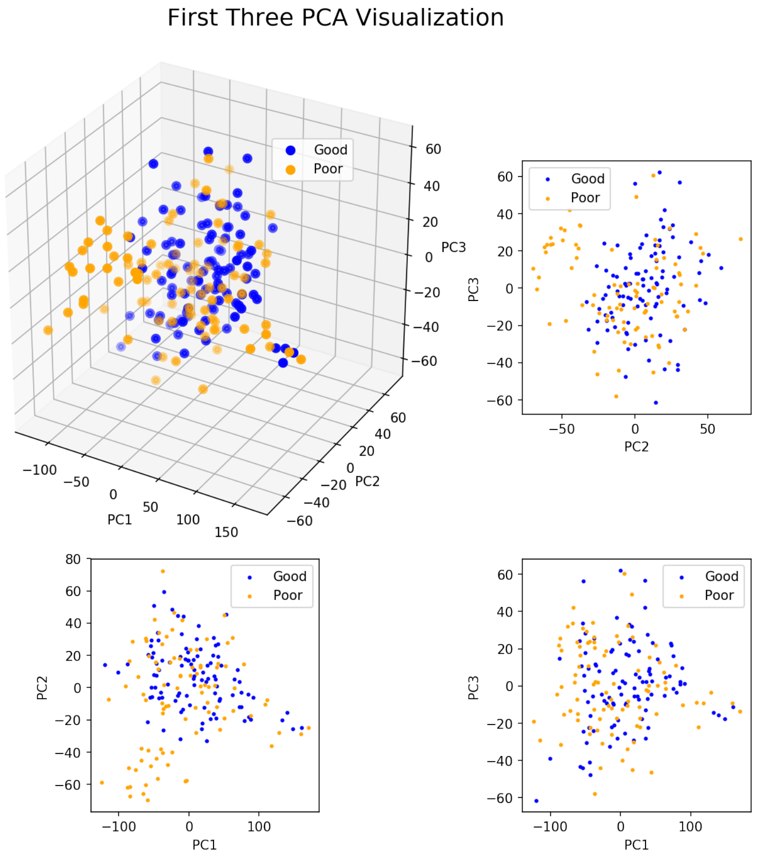

Figure 1: First three principle components visualization.

From the visualization result, the data is almost inseparable, which enhances the difficulty of classification.

3 Approach/Methods

3.1 Preprocessing

For our data, preprocessing includes finding common genes as mentioned, as well as normalization and dimension reduction.

3.1.1 Normalization

z-scoring is a common normalization approach before classification, as many estimators like KNN utilize distance between data points to perform classification. z-scoring helps scale all data features to zero mean and same variance (usually 1) to prevent that one or several features dominate the performance. The transformation formula is stated in Eq. (1).

$$z = \frac{x-u}{s}, \tag{1}$$

where u is the mean of the training samples and s is the standard deviation of the training samples.

3.1.2 Dimension Reduction

As shown in Table 1, the dimension of the data is 8819. Although most classifiers are able to be trained by the original data, such a high dimension can largely decrease model interpretability. In this report, we use both feature selection (Lasso) and feature generation (PCA) to realize dimension reduction.

Lasso Feature Selection [11]: It is not uncommon that Lasso serves for feature selection in high dimensional gene expression data. As a linear regression model with $l1$ regularization, the objective is to minimize:

$$\frac{1}{2n}||Xw-y||_2^2 + \alpha||w||_1, \tag{2}$$

where n is the number of samples. After the model is trained, non-zero coefficients serve to select most useful features out of all features.

Principle Component Analysis (PCA) [12]: As a feature generation method, PCA aims to identify components that explain the most variance of the data information. The transformation is in the form:

$$\boldsymbol{y} = \mathbf{A}^T\boldsymbol{x}, \tag{3}$$

where $\boldsymbol{x}$ is the input vector, $\boldsymbol{y}$ is the output vector, and $\mathbf{A}$ is the transformation matrix. PCA requires that $E\{\boldsymbol{x}\} = 0$, thus $E\{\boldsymbol{y}\} = 0$. As $\boldsymbol{y}$ should have uncorrelated components, the correlation matrix should be a diagonal matrix:

$$R_y = E\{\boldsymbol{y}\boldsymbol{y}^T\} = E\{\mathbf{A}^T\boldsymbol{x}\boldsymbol{x}^T\mathbf{A}\} = \mathbf{A}^T E\{\boldsymbol{x}\boldsymbol{x}^T\}\mathbf{A} = \mathbf{A}^T R_x \mathbf{A}. \tag{4}$$

By eigendecomposition, $\mathbf{A}$ is chosen to have orthonormal eigenvectors of R_x. Data variance explained by the first 30 components as well as cumulative variance is plotted in Figure 2. As cross-validation is employed to evaluate performance in this report, PCA is trained by training data first and then testing data is projected onto the trained components. As the number of samples is limited, we keep all the components in this report. Note that only 2 components are kept in CPR.

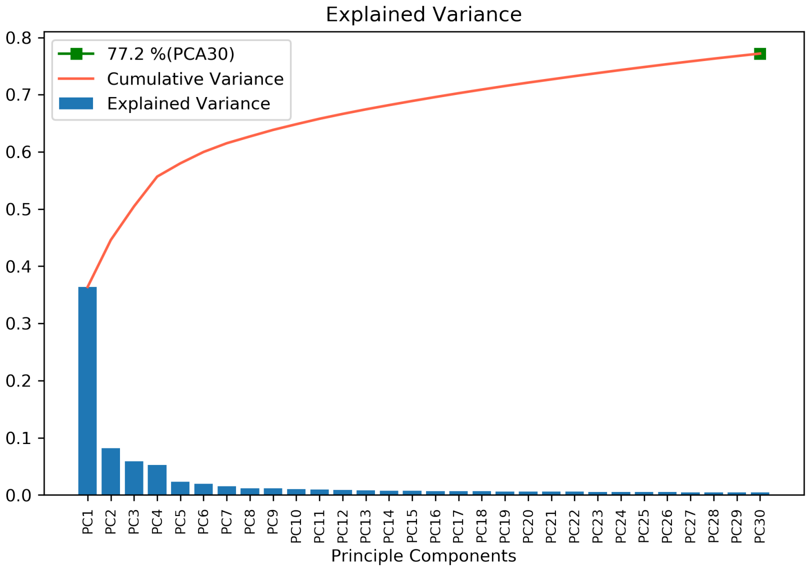

Figure 2: Explained variance for 30 PCA components.

3.2 Lasso for Classification

Although Lasso is formed as a regression algorithm, we can also use it to conduct classification with labels set as 0 and 1. To achieve the goal of classification, we use a threshold of 0.5, namely, predict a 1 if the decision statistic is greater than or equals 0.5 and predict a 0 otherwise.

3.3 Conventional Classifiers

K Nearest Neighbors [13]: KNN adopts information from the surrounding points to make classification decisions. The decision statistic is computed as follows:

$$\lambda(x) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{I}(y_i = 1), \tag{5}$$

where $\mathbb{I}$ is the indicator function that holds:

$$\mathbb{I}(y_i = 1) = \begin{cases} 1, \text{if} \quad y_i = 1, \\ 0, \text{if} \quad y_i = 0. \end{cases} \tag{6}$$

Naive Bayes: Naive Bayes classifier assumes feature independence, which is not true of gene expression data. However, PCA generates uncorrelated data so this classifier is worth trying. Here we assume Gaussian distribution:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\text{exp}\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right). \tag{7}$$

Generally, Naive Bayes estimates data class that holds:

$$\underset{y}{\text{argmax}} P(y)\prod_{i=1}^{n} P(x_i|y) \tag{8}$$

Logistic Regression [14]: Regularized logistic regression solves:

$$\underset{w,c}{\text{argmin}}||w|| + C\sum_{i=1}^{n} \log(exp(-y_i(X_i^T w + c)) + 1) \tag{9}$$

Both $l1$ and $l2$ norm are tested in our report for evaluating performance.

Decision Tree [15]: A decision tree classifier splits the feature space into several non-overlapping regions to achieve the goal of classification. It finds splits $R_1, ...R_J$ that minimize the residual sum of squares:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_J})^2. \tag{10}$$

Random Forest [16]: Random forest is one of the ensemble methods that combine predictions of multiple decision trees. Each tree is learnt from a bootstrap sample of the original training data. Such randomness can relieve the model's overfitting due to averaging.

3.4 Clustering and Pagerank (CPR) [7]

3.5 Modified CPR

4 Results

4.1 Algorithmic Pipeline

The entire designed algorithm is illustrated in Figure. A random test train split is firstly conducted to prepare cross-validation. The normalization and dimension reduction are used for preprocessing. Classification and CPR algorithms are tested, combined and compared. Parameter tuning is employed to finalize identifying the best estimator.

4.2 ROC Results

ROC results of lasso classification is shown in Figure 3. The classification accuracy is 68.4%.

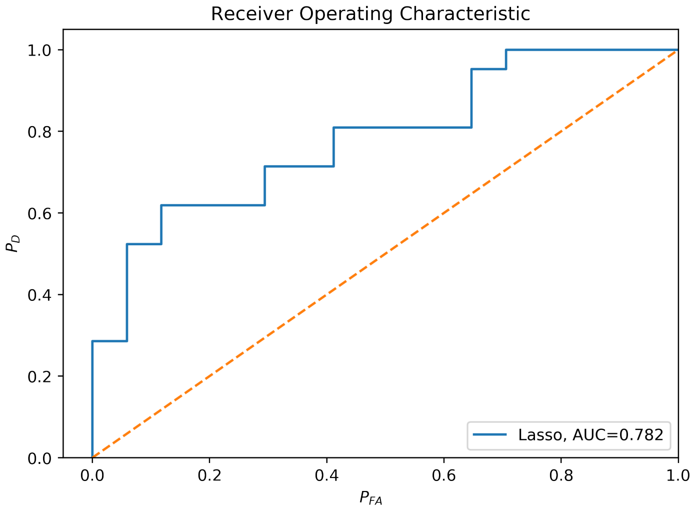

Figure 3: ROC results of Lasso classification.

ROC results of all mentioned traditional classifiers combined with PCA feature generation is shown in Figure 4. ROC results using Lasso feature selection is shown in Figure 5. From the results, logistic regression shows the best performance.

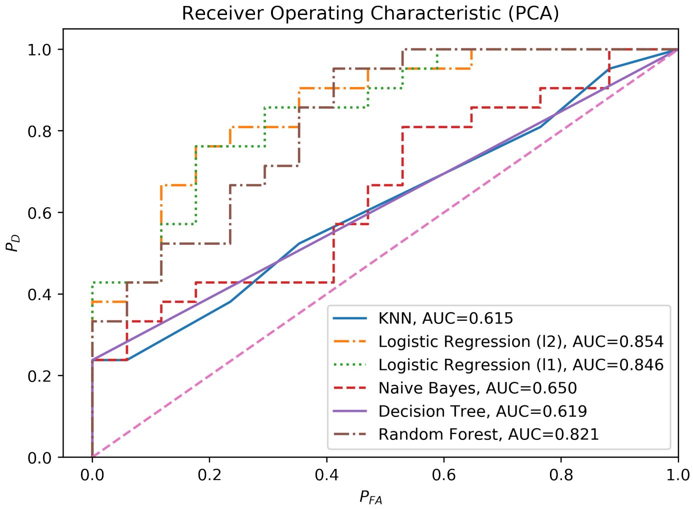

Figure 4: ROC results of conventional classifiers using PCA feature generation.

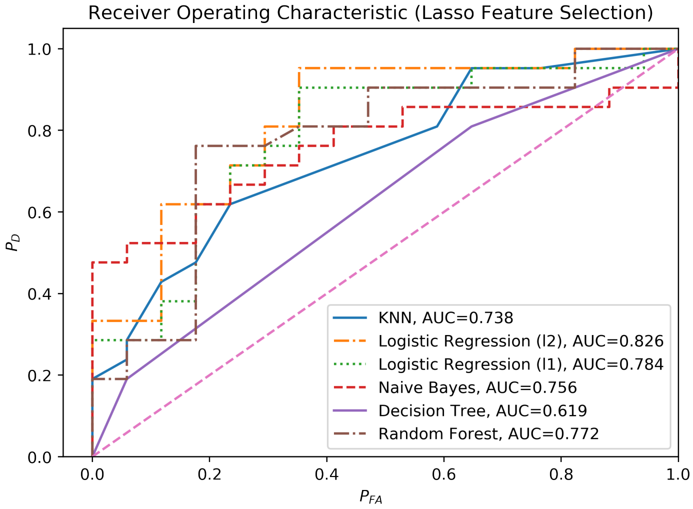

Figure 5: ROC results of conventional classifiers using Lasso feature selection.

4.3 Classification Accuracy

Classification accuracy of all conventional classifiers is summarized in Table 2.

Table 2: Classification accuracy of all conventional classifiers.

Dim Reduction **Classifier**	**PCA**	**Lasso**
Logistic Regression (l2)	68.4%	68.4%
Logistic Regression (l1)	71.1%	68.4%
KNN	55.3%	63.2%
Naive Bayes	63.2%	68.4%
Decision Tree	57.9%	60.5%
Random Forest	63.2%	71.1%

5 Conclusion

References

[1] P. Sajda, "Machine learning for detection and diagnosis of disease," *Annu. Rev. Biomed. Eng.*, vol. 8, pp. 537–565, 2006.

[2] M. Ibrahim, K. Chakrabarty, and U. Schlichtmann, "Synthesis of a cyberphysical hybrid microfluidic platform for single-cell analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2018.

[3] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, p. 117693510600200030, 2006.

[4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.

[5] X. Huang and W. Pan, "Linear regression and two-class classification with gene expression data," *Bioinformatics*, vol. 19, no. 16, pp. 2072–2078, 2003.

[6] Y. Liu, X. Yin, J. Zhong, N. Guan, Z. Luo, L. Min, X. Yao, X. Bo, L. Dai, and H. Bai, "Systematic identification and assessment of therapeutic targets for breast cancer based on genome-wide rna interference transcriptomes," *Genes*, vol. 8, no. 3, p. 86, 2017.

[7] J. Choi, S. Park, Y. Yoon, and J. Ahn, "Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers," *Bioinformatics*, vol. 33, no. 22, pp. 3619–3626, 2017.

[8] Dataset, "Analysis-ready standardized tcga data from broad gdac firehose 2016 01 28 run," *Broad Institute TCGA Genome Data Analysis Center (2016)*, 2016.

[9] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 42, no. D1, pp. D472–D477, 2013.

[10] A. Fabregat, K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 44, no. D1, pp. D481–D487, 2015.

[11] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[12] J. Lever, M. Krzywinski, and N. Altman, “Points of significance: Principal component analysis,” 2017.

[13] H. Trevor, T. Robert, and F. JH, “The elements of statistical learning: data mining, inference, and prediction,” pp. 463–471, 2009.

[14] ——, “The elements of statistical learning: data mining, inference, and prediction,” pp. 119–127, 2009.

[15] ——, “The elements of statistical learning: data mining, inference, and prediction,” pp. 305–313, 2009.

[16] ——, “The elements of statistical learning: data mining, inference, and prediction,” pp. 587–601, 2009.