Predicting Breast Cancer Diagnosis from Fine Needle Aspiration using PLS-DA and Logistic Regression

Ali Farhat1 and Nikan Namiri1

1. Department of Bioengineering, UCLA

Introduction

Breast cancer mortality is top five among all cancers and will consist of 266,120 new invasive cases in 20181. It is essential that efficacious feature detection is created for early diagnosis, so that the mortality rates decrease.

Initial evaluation of breast cancer is typically performed using a mammogram, which utilizes X-rays to search for lesions within the breast. If a suspect mass is identified, it is more accurately quantified using fine needle aspiration (FNA). FNA provides a cost-effective, minimally invasive method that allows pathologists to investigate the cellular properties of the suspect tissue, lending higher resolution in identifying the lesion2. Various factors are used to determine whether the cells viewed through the microscope are cancerous, including cell size and number of nuclei. In light of these components, our team decided to implement machine learning techniques on data received from breast cell images in order to understand predictive factors of malignant cancer cells.

Previous machine learning studies using FNA have included Multisurface Method-Tree, which iteratively adds separating planes into feature space until distinct regions are created for malignant and benign cells; the classifier is then used as a decision tree for categorization3. An additional model from the literature includes classification using support vector machines, which provides levels of malignancy for cells and was shown to reach an accuracy of 94.24%. However, this model was not externally validated on data not used for model generation4. Our project aims to utilize two methods known as partial least squares discriminant analysis (PLS-DA) and logistic regression (LR) on FNA data received from the UCI Machine Learning Repository.

Problem

The goal of this study is to analyze the most significant phenotypic features of cells from FNA in predicting tumorigenesis. The phenotypic features of cells must be identified and measured in a thorough manner without losing characteristic information of the cells. This principle makes it often difficult to simulate a biological system using engineering principles, as biological phenomena, such as cancer cells, possess many parameters, which must be selected and subjectively measured by human instruments. In the case of FNA, nine factors of cells were clinically quantified into a data set, although it is evident that these parameters suffer from physician bias and may not be the most optimal for fully simulating a cancer cell.

However, these factors, such as cell size and number of replications, may still provide accurate results, as they are characteristic of significant contributors to cell pathology. Additionally, there is a tradeoff between complexity and predictive capability of a model, and thus even though we may not have an infinite number of parameters, we may be able to have high predictive capacity according to analysis methods and using different assumptions, such as covariance and independence.

Methods

PLS-DA was used to identify parameters that have the greatest covariance with the cancerous state of the cell. The scores and loadings matrices were obtained using sklearn.cross\_decomposition.PLSRegression. In this algorithm, the desired number of principal components were determined by calculating the explained variance (R2Y) for various numbers of components and comparing them on a plot (Figure 2). PLS-DA is essentially analogous to partial least squares regression (PLSR), with the exception that the PLS-DA is used for binary outcomes. The package is fit on X (input) and Y (binary output) data and computes the scores and loadings matrices, which provided the values of each point in principal component space and the weights (loadings) of each variable, respectively. Overlaying the loadings allowed for visualizing which parameters were most covariated in predicting the cell type.

LR, the second implementation method, is commonly used to predict binary outcomes from multiple parameters. LR is similar to ordinary regression, in that it assigns weights to each model parameter for prediction. However, it differs in that it prioritizes parameters that maximize the observed likelihood of the sample and it assumes independent variables. The sklearn.linear\_model.LogisticRegression module was implemented for this calculation, and the sub-modules of LogisitcRegression.fit and LogisitcRegression.predict were used to fit training folds of the input and output matrices to the LR model and then compare the subsequent predictions of the left-out fold to its measured values, respectively. In order to determine which of the variable coefficients were important for the predictive capacity of the model, bootstrapping was conducted, which revealed distributions for each parameter. The significance of a parameter was dependent on its distribution’s distance from zero.

Five-fold cross validation was used to evaluate the predictive capacity for both PLS-DA and LR models, by fitting models to all but one of the folds and using the left-out fold to predict its output. The predictions were then used to create a receiver operating curve (ROC), which displays the true positive rate as a function of false positive rate. ROC curves were then compared for both PLS-DA and LR to evaluate which method produced the better predictions for cell cancer. This was done by viewing which ROC curve had a closer position to the upper left corner, as well as a higher area underneath the curve. From sklearn, we imported confusion\_matrix, which also gauged the accuracy of both our PLS-DA and LR predictions on cell type.

Results

PLS-DA resulted in fewer false negatives but slightly higher false positives than LR. The decreased number of false negatives in PLS-DA is also reflected by its ROC curve, in which the curve is closer to the upper left corner than the LR curve, as seen in Figure 1. This reflects that for the given thresholds in the graph’s frame, PLS-DA was able to maintain a higher true positive rate and result in less false negatives (1-TPR=FNR).

False negatives are of greater importance for determining which test is more appropriate because given a false negative result, a patient will not undergo further testing and will allow the cancer to metastasize. To minimize the false negatives, and serve clinicians better, the PLS-DA model should be implemented over LR, despite the former’s slightly higher false positive rate, as seen in Table 1.

Table 1. Correct and incorrect test results from PLS-DA and LR confusion matrices.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **True Negative** | **True Positive** | **False Positive** | **False Negative** |
| **PLS-DA (2 PCs)** | 431 | 233 | 13 | 6 |
| **LR** | 433 | 224 | 11 | 15 |

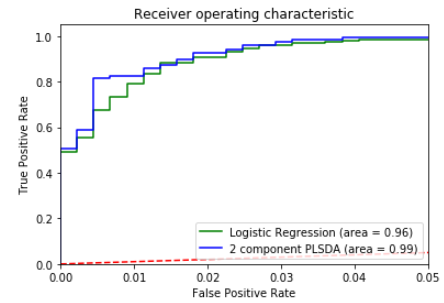


Figure 1. Zoomed in ROC curves for PLS-DA and LR.

The PLS-DA curve in Figure 2 demonstrates a relatively high R2Y with one component, with only an approximately 2.5% increase in explained variance with the addition of a second principal component. This suggests that our data is heavily covariated and can be largely explained along one major axis (i.e. PC1).

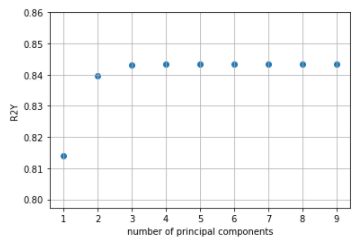


Figure 2. R2Y of each number of principal components in the PLS-DA model.

This notion is further seen in our loadings plot shown in Figure 3, in which all parameters’ loadings points are seen to be highly covariated with Y loadings along PC1, conveying that each had an important impact on the prediction of Y.

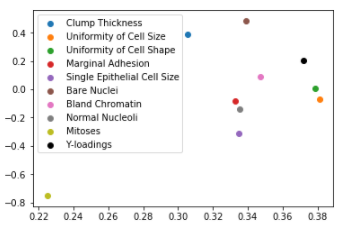


Figure 3. Plot of both X and Y loadings.

LR provided a comparison for PLS-DA, and Uniformity of Cell Shape was found to have a beta weight of 0.298, and a loadings very close to that of Y in Figure 3. As seen in Table 2, this parameter possessed the second highest weight.

Table 2. Beta weights of parameters from LR.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Clump Thickness | Uniformity of Cell Size | Uniformity of Cell Shape | Marginal Adhesion | Single Epithelial Cell Size | Bare Nuceli | Bland Chromatin | Normal Nucleoli | Mitoses |
| 0.263 | 0.070 | 0.298 | 0.169 | 0.003 | 0.343 | 0.105 | 0.175 | 0.209 |

Uniformity of cell shape was an effective parameter in determining cancer, which is also supported by recent findings stating that the shape of the cell, as determined by the stresses and strains of its surroundings, directly influences genetic expression and leads to cancer activation5-6. In accordance with our implementation results, the cell shapes of FNA lesion images may possess more valuable information that other cell parameters in determining cancer diagnosis in the early stages.

The PLS-DA loadings of Mitoses had a negative correlation with the y loadings along PC2, and the smallest loadings along PC1.This was unexpected, as the number of mitoses of a cell has been shown to result in cancer; disrupted cell division can lead to uneven chromosome separation in daughter cells, which is a cause of tumorigenesis7. On the other hand, LR gives a relatively high weight to Mitoses in the model, which emphasizes the differing assumptions (i.e. covariance vs. independence) of the two models. Nonetheless, Mitoses is still a significant predictor of cancer, as the X loadings is still positive. All in all, LR highlights this parameter through a high weighting, while PLS-DA does not ostensibly suggest the same.

We see that all parameters in PLS-DA, while most in LR, contribute to tumorigenesis, which helps solve our initial question of the some of the most significant predictors for breast cancer metastasis. Moreover, the different assumptions about collinearity of the data account for differences in PLS-DA and LR performance, which calls for future external validation on new data to reveal which model performs better in the clinical setting.

References

1. American Cancer Society: Cancer Facts and Figures 2018. Atlanta, Ga: American Cancer Society, 2018.

2. Wu, M., & Burstein, D. E. (2004). Fine needle aspiration. *Cancer investigation*, *22*(4), 620-628.

3. 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.

4. Jeleń, Ł., Fevens, T., & Krzyżak, A. (2008). Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies. *International Journal of Applied Mathematics and Computer Science*, *18*(1), 75-83.

5. Sailem, H. Z., & Bakal, C. (2017). Identification of clinically predictive metagenes that encode components of a network coupling cell shape to transcription by image-omics. *Genome research*, *27*(2), 196-207.

6. Alizadeh, E., Lyons, S. M., Castle, J. M., Foss, J. I., & Prasad, A. (2017). Is Shape of Cancer Cell Correlated with its Invasiveness?. *Biophysical Journal*, *112*(3), 124a-125a.

7. Weaver, B. A., & Cleveland, D. W. (2005). Decoding the links between mitosis, cancer, and chemotherapy: The mitotic checkpoint, adaptation, and cell death. *Cancer cell*, *8*(1), 7-12.