Zachary Boroughs

Computational Molecular Medicine

JHED: zboroug1

The Importance of Feature Selection in Predicting Breast Cancer Progression Through HER2 Expression

I. Background and Dataset

The dataset to be discussed consists of normalized, log-2 scaled values for gene expression taken from the primary tumor of patients with breast cancer prior to initiating neoadjuvant chemotherapy by either fine needle aspiration or core biopsy. The phenotype of interest (that being the phenotype exhibited by the binary "Y" variable is that of human epidermal growth factor (HER2) status. The two classes are represented by positive, "Y=P," and negative, "Y=N," which was determined via immunohistochemistry. The dataset contains gene expression information from 140 different patients covering 13,299 different genes. The goal of this study is to determine whether one can accurately predict HER2 expression status utilizing a classifier of the gene expression values extracted and measured. A single dataset was given for further separation into training and testing sets.

II. Initial Data Examination and Analysis

The first step in data preprocessing was to transform the labels given into a binary "0-1" format for ease of use and later classification. During this preprocessing, we notice a large imbalance in the HER2 expression status with only 30 positive values recorded compared to 111 negatives. There are pros and cons associated with rebalancing classes for training purposes with

2 major, contradicting philosophies dominating the field of machine learning. Some argue that training on imbalanced data skews the classifiers, biasing the model towards the majority class. Others argue that, despite the potential for bias, training on rebalanced or up-sampled data gives an inaccurate picture of reality. This argument is especially relevant in relation to cancer prediction and expression, where it is argued that even if a model picks up bias when training on uneven classes, that bias is an inherently important component of the data itself. For example, if cancer expression rate is low, it would be desirable that the model is aware of this imbalance so that it is able to make predictions conservatively. This second approach has been adopted in the analysis presented in this paper as, although we desire our model accurately predict positive expression values, predicting negative expression is equally important in the context of a patient with tumor expression. A thorough analysis of sensitivity and specificity will be performed following any classification results to ensure that the model in question can accurately predict both possibilities.

A correlation matrix was first computed to determine the genes with the most positive correlation to HER2 status in addition to those with the strongest negative correlation to HER2 status. These Pearson correlation coefficients were calculated through taking the covariance between each gene and the HER2 expression values and then dividing by the product of the individual standard deviations.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Figure 1: The formula utilized to determine correlation between two random variables, In this case HER2 expression and a given gene

Of the 13,299 gene values recorded, those with the strongest positive correlation with HER2 expression included GRB7, STARD3, PGAP3, PSMD3, and MIR6884///MED24 in descending order, with GRB7 topping the list with a positive correlation of .696. Upon further research of the gene, we see that "growth factor receptor-bound protein 7 (GRB7) gene is located adjacent to the HER2 gene on the 17q12-21 amplicon, is often co-amplified with HER2 in a subset of breast cancers, and has been implicated in resistance to anti-HER2 and antiestrogen therapy."[1] As such, a strong positive correlation coefficient makes sense in the context of HER2 expression due to its proximal location and co-amplification within breast cancer cases. In analyzing the most negatively correlated genes, we see that ASB13, NUDT6, HCFC2, SERPINA3, and SEC22B emerge as having the strongest negative correlation with HER2 expression status, with ASB13 having a Pearson correlation coefficient of -.438. Upon further analysis of ASB13, "clinical data analysis reveals that ASB13 expression is positively correlated with improved overall survival in breast cancer patients. These findings establish ASB13 as a suppressor of breast cancer metastasis by promoting degradation of SNAI2 and relieving its transcriptional repression of YAP."[2] So here, we observe that ASB13 is strongly correlated with survival in breast cancer patients through suppressing the metastasis of the cancer. As such, this negative correlation follows logically.

---

[1] Bivin, William W. MD; Yergiyev, Oleksandr MD; Bunker, Mark L. MD; Silverman, Jan F. MD; Krishnamurti, Uma MD, PhD GRB7 Expression and Correlation With HER2 Amplification in Invasive Breast Carcinoma, Applied Immunohistochemistry & Molecular Morphology: September 2017 - Volume 25 - Issue 8 - p 553-558
doi: 10.1097/PAI.0000000000000349

[2] Fan, Huijuan, et al. "Abstract 162: ASB13 Inhibits Breast Cancer Progression and Metastasis through SNAI2 Degradation and Transcriptional Regulation of YAP." *Tumor Biology*, 2019, doi:10.1158/1538-7445.am2019-162.

In further analyzing these two strongly correlated genes, both the mutual information and conditional entropy between the separate genes and expression values were calculated.

$$I(X;Y) = D_{\mathrm{KL}}(P_{(X,Y)} \| P_X \otimes P_Y)$$

Figure 2: The formula utilized to determine mutual information between 2 random variables, In this case HER2 expression and a given gene

The mutual information of GRB7 and HER2 expression was found to be .3 while the mutual information between ASB13 and HER2 expression was .162. A mutual information value of zero would mean complete independence, so although small, the nonzero nature of these results further illustrate the dependence and correlation of these genes with the HER2 expression status. The conditional entropy, in both cases, was markedly lower than the individual entropies as well, with the conditional entropy dropping to 4.63 from 4.93 for GRB7 status and HER2 expression and to 4.70 from 4.94 for ASB13 status and HER2 expression, again confirming our previous hypothesis of some form of dependence between the variables.

Lastly, a Wilcoxon Rank Sum Test with level alpha of .05 was performed to get a broader view of how many of the genes could be reasonably linked with HER2 expression. Out of the 13,299 genes, only 1,603 of those genes returned a p-value of less than .05 following the test and were able to reject the null hypothesis of independence. However, due to the chance of significance with large random testing, a Family Wise Error Rate (FWER) correction was performed, further limiting the number of statistically significant genes to only 12 with a p-value of less than .05 that successfully reject the null hypothesis. As such, we now have a number of

differing gene sets on which to perform our classification to determine the ideal method and hyperparameters for testing and predicting on our dataset.

III. K-Nearest Neighbors Classification

The first form of classification performed was a K-Nearest Neighbors Classifier, implemented utilizing the Scikit-Learn python library. This methodology takes in the training data and makes predictions based on the testing data's distance to a predetermined number of neighbors. We surmised that such a model would perform poorly on the entirety of the dataset due to the large number of uncorrelated features. However, upon using corrected hypothesis testing to greatly narrow the number of features, we predict the model will perform extremely well due to the relatively small number of genes shown to be strongly correlated. To determine the optimal value of K that would return the highest balanced accuracy for the dataset. The balanced accuracy of a classifier is given by the equation,

$$Balanced\ Accuracy\ = \frac{Sensitivity + Specificity}{2}$$

Figure 1: Equation utilized to calculate balanced accuracy

Where, in figure 1, sensitivity is our true positive rate and specificity is our true negative rate (or the false positive rate subtracted from 1). This is a much more reliable measure of accuracy when working with imbalanced data as we would like to average our classifier's ability to accurately predict both positive and negative potential outcomes. The data was first split according to a randomly selected training and testing sets with a training fraction of .75 and 10-fold cross validation was performed to confirm validity of classifier accuracy. We then performed feature selection for the entire genetic dataset before running K-Nearest Neighbors for K values from 1

through 30. Our initial metric was that of the balanced accuracy across the varying K values. After performing 10-fold cross validation with a .75 training fraction, the following results were observed.

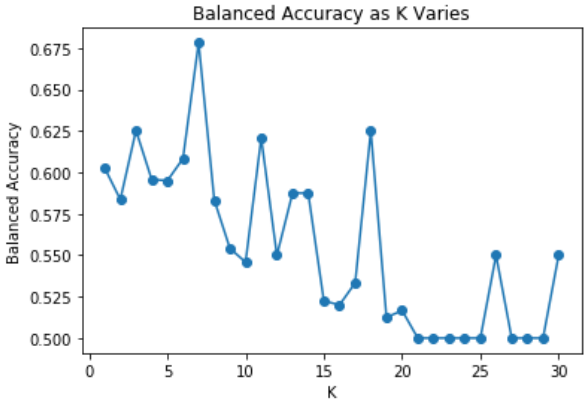

Figure 2: Balanced accuracy as K varies for Nearest Neighbors trained on the entirety of the dataset

Initially, when trained on the entirety of the dataset, we see that the balanced accuracy generally trends downwards as the number of neighbors increases following a brief peak at 7-Nearest Neighbors with a balanced accuracy of .675. The classifier was then run again with this hyperparameter in mind and the following ROC curve was generated.

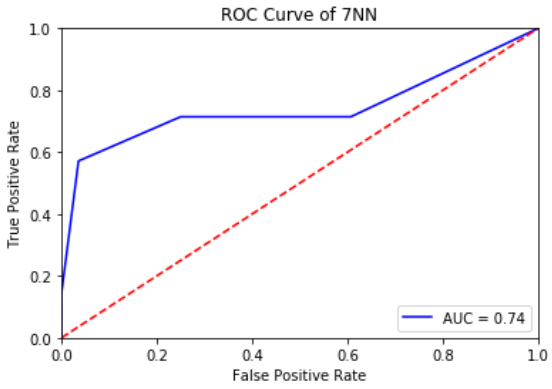

Figure 3: Receiver Operating Characteristic Curve for 7 Nearest Neighbors trained on the entirety of the dataset

This curve gives very similar information to that of the balanced accuracy value with an AUC of roughly .7. In further analyzing the information presented, we see that, after an initial rise, our true positive rate remains relatively constant as our false positive rate increases, signalling that the overwhelming number of genes in the dataset could be misleading the classifier from identifying our true negatives in addition to the true positives.

The same methodology was then applied to those genes that successfully rejected the null hypothesis via an uncorrected Wilcoxon Rank Sum Test. As explained prior, this test narrows down the gene pool from 13,299 to 1,603. The Wilcoxon Test was performed within each iteration of the cross validation on the training set selected, ensuring that the testing set remained independent from the feature selection process.

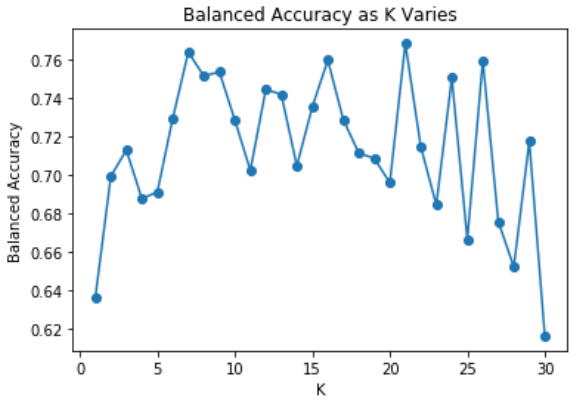

Figure 4: Balanced accuracy as K varies for Nearest Neighbors trained on the features selected through uncorrected hypothesis testing

For these genes, we see no real trend across the varying k values. This could likely be due to the number of randomly significant genes that remain in the dataset due to the uncorrected hypothesis testing performed. However, we do observe that at very low and very high values of K, the balanced accuracy suffers, although to not much lower than the peak accuracy observed across the entirety of the dataset. For comparison, both the peaks at 7 and 21 Nearest Neighbors will have ROC curves examined.

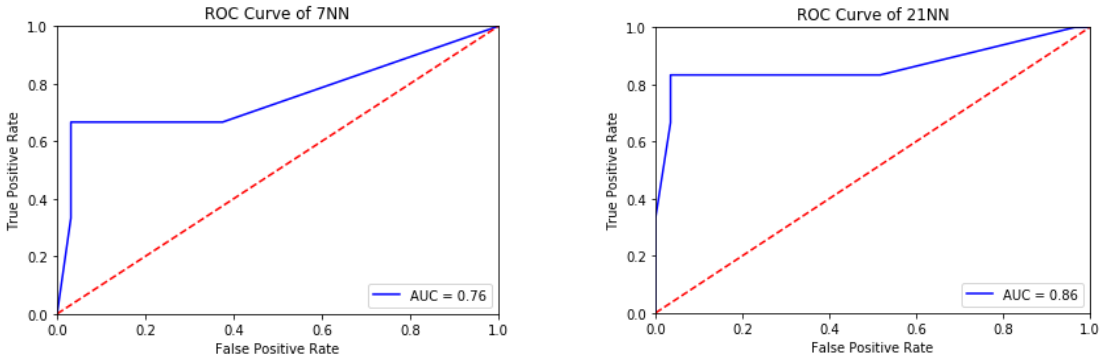

Figure 5 and 6: Receiver Operating Characteristic Curve for 7 and 21 Nearest Neighbors, respectively, trained on the features selected through uncorrected hypothesis testing

In both cases, we see improved performance over that of the entire dataset. However, similarly in both cases we observe the same leveling out of the true positive rate as the false positive rate increases as observed before. It is of note that increasing the K value to that of 21 would mean a largely increased bias to this specific set of data. As such, 7-Nearest Neighbors would be a more applicable choice as a classifier to be applied in a real world context.

Finally, this same process was applied to only those genes that passed the corrected hypothesis testing. Similar to the uncorrected genes, these genes were selected within each iteration of cross validation by performing hypothesis testing on the training set.

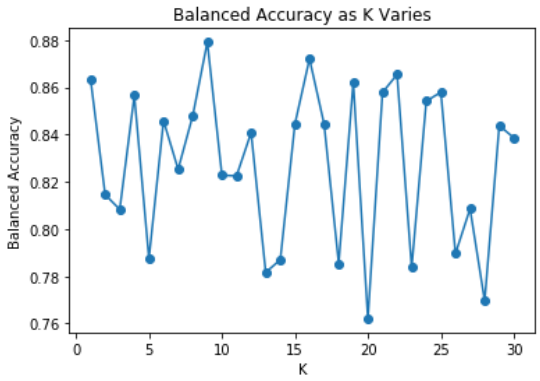

Figure 7: Balanced accuracy as K varies for Nearest Neighbors trained on the features selected through FWER corrected hypothesis testing

Again, we see large variance across the K range tested, but across all K values we see increased balanced accuracy relative to the uncorrected genes tested prior. In this case, the K value of 9 was selected for further analysis. When this K value was trained on 75% of the sample size and then tested on the entirety of the dataset, an average balanced accuracy of .845 was returned, showing slightly lowered but still extremely high performance in comparison to the performances of the previous feature selections examined.

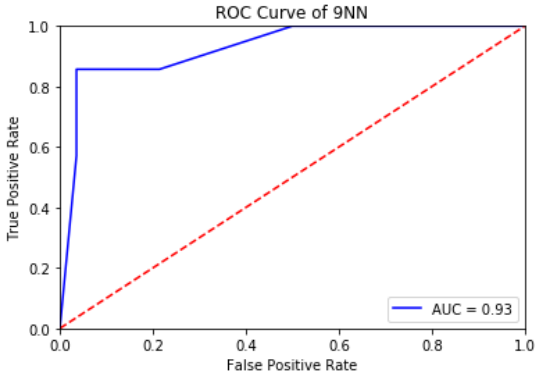

Figure 8: Receiver Operating Characteristic Curve for 9 Nearest Neighbors trained on the features selected through FWER corrected hypothesis testing

This ROC curve shows extremely promising results. The false positive rate is a measure of the specificity of our data subtracted from 1, which is a measure of the model's ability to correctly predict the negative class. The true positive rate is a measure of sensitivity, or our model's ability to correctly predict the positive class. To achieve a minimum sensitivity of .8, this model has a specificity of .964, calculate using interpolation. This means that when our model accurately predicts progression 80% of the time, the model also predicts non progression 96.4% of the time. As discussed prior in regard to the imbalance of the class data, it is important that our model be able to predict both potential outcomes of cancer expression. In the analysis of the ROC curve and the sensitivity and specificity it conveys, despite the class imbalance our classifier remains able to predict both potential outcomes, which is desirable in a clinical setting. It is worth noting that further increasing the selectivity of our hypothesis test (decreasing our level alpha) had no substantial effect on classifier performance.

IV. Random Forest

The Scikit-Learn Random Forest classifier was then tested on the data to determine the potential for an even higher success rate with the data given. The Random Forest method combines numerous decision trees, trains each one on a slightly modified set of the observations, and splits nodes in each tree based on a limited number of the features. The final predictions are then made through averaging the predictions on each individual tree. Normally, with an extremely large and varied number of features to input, a shallow decision tree would be desirable without overfitting to extraneous data. However, in the case of a more limited feature

pool, a deeper tree could be beneficial in determining the ideal nodes and splits to develop an accurate classifier.

The main hyperparameters for a random forest classifier are that of the number of estimators and the maximum depth of the tree. The number of estimators can be likened to the number of trees in the forest and is the number of different measurements that will then be averaged together, as part of an ensemble method. In the case of very noisy datasets, having a large number of trees can cause overfitting on the dataset. However, in application with our previously found gene pool that successfully rejects the null hypothesis following corrected Wilcoxon Testing, a large number of trees should not be an issue. As a result, a larger number of estimators in this case should only increase computational power required. As such, the default value of 100 will be used. In analyzing the ideal maximum depth of the tree, a more nuanced scenario arises. A low max depth has high bias and low variance as we allow only one node to form in our classifier's decision making, resulting in underfitting. On the other hand, a high max depth would have a much lower bias and higher variance, a result of potential overfitting on our specific dataset.

As such, to determine the ideal max depth for our decision tree, the same methodology as that of part 3 will be used. We will be training and validating our Random Forest Classifier at max-depth values in increments of 10, starting with 10 and increasing to 100, followed by an evaluation at the default maximum depth of none to determine the ideal hyperparameter. 10-fold cross validation will be utilized with feature selection of the corrected, hypothesis tested gene pool occurring within each iteration as we perform Wilcoxon Rank Sum Testing and FWER correction on the training set. The balanced accuracy will be evaluated at each hyperparameter recorded. The following results were obtained.

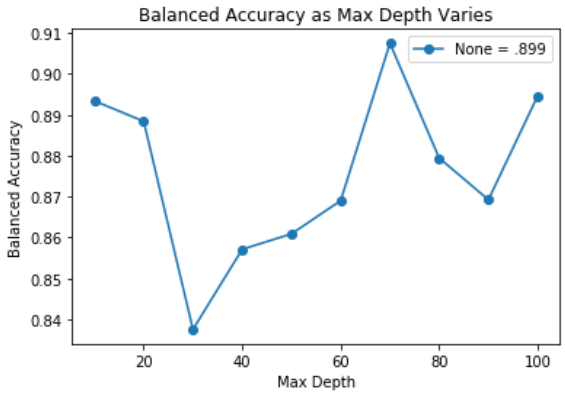

Figure 1: Balanced accuracy as max depth varies for Random Forest trained on the features selected through FWER corrected hypothesis testing

From the chart given, we can observe that across the board our classifier performs very well on the testing data. Although there is some variation, we can observe a substantial decline in balanced accuracy at a max depth of 30 with a noticeable peak at 70. Even lacking a max depth entirely produces a fairly high balanced accuracy, although lower than that of 70 and potentially introducing unnecessary variance. As such, a max depth of 70 was used to produce the following ROC curve.

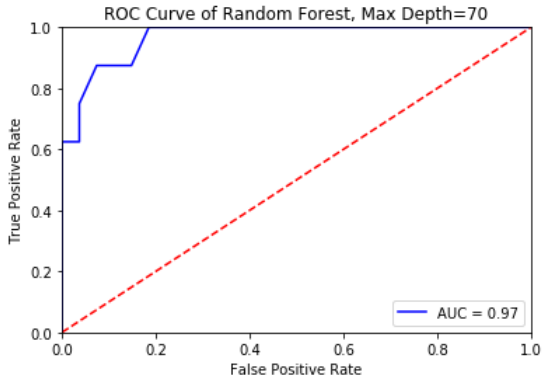

Figure 8: Receiver Operating Characteristic Curve for max depth of 70 Random Forest trained on the features selected through FWER corrected hypothesis testing

The ROC curve observed is equally impressive as that of the Nearest Neighbors classifier, having a .8 sensitivity threshold at .948 specificity, meaning when the classifier can predict HER2 expression 80% of the time, we can additionally predict cancer nonprogression roughly 93% of the time. Even when trained on our testing data and then tested on the entirety of the dataset, 10-fold cross validation produces an average balanced accuracy of .964. For reference, the same Random Forest classifier trained without limited feature selection produces an average balanced accuracy of only .564.

V. Discussion and Future Directions

Throughout this report, a number of varying classifiers were applied to various subsets of the data. Although both classifiers performed fairly comparably at the end of hyperparameter selection and feature selection, the largest effector on balanced accuracy and ROC curve performance was that of the feature selection. To recap, the data initially consisted of 13,299 genes. Upon initial hypothesis testing through a Wilcoxon Rank Sum Test with level alpha equal to .05, we were able to successfully narrow down that number of genes to only 1,603. However, due to the extremely large size of our dataset coupled with the large class imbalance meant that the odds of random significance were extremely high and, as such, a number of the genes even within this limited subset would be extraneous in training our classifier. As such, Family Wise Error Rate (FWER) correction was performed, eliminating the potential for this random significance discussed. Following this process, the gene pool for feature selection was narrowed down to an astonishingly low 12 genes for a level alpha of .05.

Gene Number (in Dataset)	Gene Name	Gene Number (in Dataset)	Gene Name
633	PSMD3	8247	MIR6884///MED24
1752	STARD3	8730	LOC102724788///PRODH
4865	PNMT	9277	GSDMB
6473	PLXNB2	10734	ASB13
7441	GRB7	11903	NUDT6
7489	ERBB2	13083	PGAP3

Figure 1: The features selected through FWER corrected hypothesis testing

As confirmation of the success of this corrected hypothesis test, both the most strongly positively and negatively correlated genes with the HER2 expression status can be found on the table in figure 1, along with the remainder of the top 12 most strongly correlated genes with HER2 expression found from the correlation matrix computed prior. Each of these genes has independently been researched for its impact on HER2 expression and breast cancer propagation (although for the sake of brevity and focus on the classification itself, most research has been omitted from this report). As such, it logically follows using only these most strongly and relevantly correlated genes would remove a large amount of the extraneous randomness from our dataset, allowing the classifier to focus on only the important factors. Although this approach does introduce a large amount of bias into our methodology, this sort of bias is acceptable as it allows us to narrow our focus to create a more accurate classifier with fewer genetic readings and less computational power, saving time on both the clinical and analytical sides of the process.

In isolating a classifier for recommendation, the analyses in this paper have shown that a Random Forest classifier with 100 estimators and a max depth of 70 consistently produced the

best results on our limited dataset, producing a balanced accuracy on the entirety of the feature selected data set of .964 with very high sensitivity and specificity readings, even recording a specificity rate of 80% at a sensitivity of 100%. However, the most important takeaway from this dataset is the feature selection itself, as simply limiting feature selection to those genes that are most strongly correlated with HER2 expression allows for a much more accurate and less computationally intensive classifier. This information is also extremely beneficial in a clinical setting as the readings of many extraneous genes could be left out altogether for the sake of HER2 expression prediction, focussing only on those genes that are relevant and aid the classifier in performance.

Although in the context of the classifiers examined in this report, further narrowing of feature selection was not beneficial, it could be a relevant vein of future inquiry to see how it would impact other classifiers. Additionally, it could be interesting to perform further clinical examination as to how pairs of these highly statistically significant genes could be related to each other. This greatly narrowed down gene pool could open many interesting fields of inquiry into how certain genes either amplify or deamplify the effects of one another in the context of HER2 expression and breast cancer growth. Lastly, if this tool was to ever be implemented in a clinical setting, the incorporation of some sort of quantification of results may be useful. If certain specific bins could be output by the classifier based on reliability, the applicability of the tool in a live, clinical setting could be greatly improved. If the doctor and patient knew, not only the prediction, but how certain the classifier was of HER2 expression, it could help with a variety of decisions regarding patient care and potential treatment options.