Predicting Risk of Breast Cancer Relapse from Copy Number

Soumya Kundu
Computer Science Dept.
Stanford University
Stanford, CA 94305
soumyak@stanford.edu

Jose A. Seoane
Medicine and Genetics Depts.
Stanford University
Stanford, CA 94305

Christina Curtis
Medicine and Genetics Depts.
Stanford University
Stanford, CA 94305

1 Introduction

In this work, we develop machine learning models to classify breast cancer patients as having high or low risk of experiencing a distant relapse, which is a recurrence of their cancer at a location that is distant from the original site of the tumor, usually after many years following the original diagnosis. Specifically, we are interested in patients who have an immunohistochemical (IHC) subtype of $ER^+/HER2^-$, since late, distant recurrence of breast cancer and its treatment represent critical clinical challenges for this group of patients [1]. Although existing methods can cluster cancer patients into subtypes that have been shown to be enriched for high and low risk of distant relapse by using both gene expression and copy number, we aim to build a classifier for distant relapse risk using only copy number data, along with basic clinical variables, since both gene expression and copy number data may not always be available for a given cohort of patients.

The input to our models is copy number data from the tumors of the patients in the METABRIC cohort, along with their age at diagnosis, tumor grade, tumor size, and the number of tumor-positive lymph nodes. First, each patient is given a positive or negative label, corresponding to high or low risk of distant relapse, respectively, based on their estimated risk of distant relapse as predicted by the IntClust algorithm, which can cluster breast cancer patients into subtypes that are enriched for high or low risk of distant relapse among $ER^+/HER2^-$ patients using gene expression and copy number from their tumors. Next, we train supervised learning algorithms, including logisitic regression, support vector machines, and neural networks, to predict the IntClust risk label from just the copy number and basic clinical features, and we use 5-fold cross validation to generate unbiased predictions for the entire cohort of patients.

Finally, we use the actual clinical outcomes of the patients in the METABRIC cohort to conduct survival analysis. We construct cumulative incidence curves, which are 1 minus the Kaplan-Meier estimates, for the predicted high risk and low risk patients and conduct a log-rank test to determine whether the empirical risk of distant relapse is significantly greater for the group predicted to be at a high risk. Next, we use the predicted risk labels and our 4 clinical features as covariates in a Cox proportional hazards model that we fit first to the entire dataset and then independently to each of the 5 training folds; we compute the hazard ratio for the predicted risk label from the former step and generate predictions for time to distant relapse for the patients in each of the 5 test folds from the latter step. Then, we use all of the generated predicted times to compute the concordance index (C-Index) for our model and compare that to the C-index obtained from performing the same analysis using just the IntClust risk labels.

2 Related work

Identifying clinically meaningful breast cancer subgroups is crucial to targeted treatment and therapy. In their paper from 2012, Curtis et al. demonstrate that the IntClust clustering algorithm, which uses gene expression and copy number to subtype cancer patients into one of 10 different groups, can find subtypes that have distinct clinical outcomes [2]. In 2014, Ali et al. show that they can

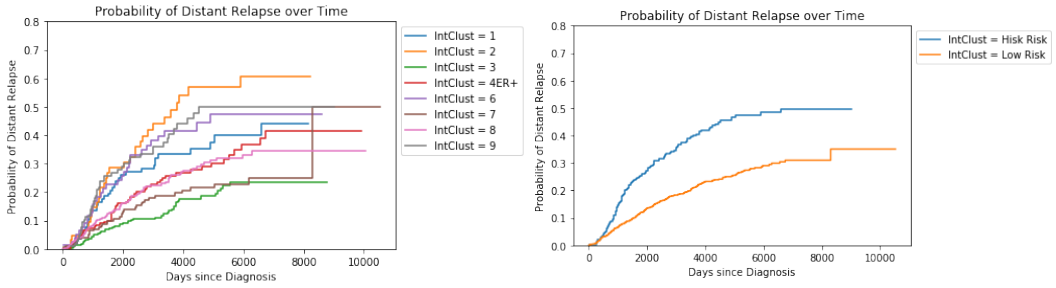
accurately recover IntClust subtypes using only gene expression, which is very useful when both gene expression and copy number data are not available [3].

In their recent paper published in March of 2019, Rueda et al. show that patients with $ER^+/HER2^-$ tumors in 4 of the IntClust subtypes are at a significantly higher risk for experiencing a distant recurrence of their breast cancer, while those in 4 other subtypes are at a much lower risk for such distant relapse [1]. Since these distant relapses are usually late relapses, which have posed a critical challenge in treating the $ER^+/HER2^-$ patients, the ability to successfully predict the risk of distant relapse for these patients will lead to significant improvements in the treatment of these patients. Although certain IntClust subtypes have been shown to be strong indicators of the risk of distant relapse, no model currently exists for accurately and directly predicting this risk of distant relapse for ER^+/HER^- breast cancer patients from genomic data. Furthermore, while there exists a method for recovering IntClust subtypes using only gene expression, there are no methods for doing so using only copy number. As a result, inferring the risk of distant breast cancer relapse using the IntClust subtypes is only possible when gene expression data is available, limiting the use of these methods.

3 Dataset and Features

3.1 Clinical Outcomes

Our molecular and clinical data come from the METABRIC dataset of 1980 patients with breast cancer [2]. First, we process the clinical data by filtering the dataset to keep only those patients with $ER^+/HER2^-$ tumors that have a known grade and size. Next, all patients belonging to the low risk IntClust subtypes, which are 3, $4ER^+$, 7, and 8, are given a negative label of 0, denoting low risk of distant relapse, while those belonging to the high risk IntClust subtypes, which are 1, 2, 6, and 9, are given a positive label of 1, denoting high risk of distant relapse; any patients not in either of these groups are removed from the dataset. This gives us 925 negative samples and 360 positive samples, for a total of 1285 samples. Plots of (1 minus the Kaplan-Meier estimates), also known as cumulative incidence curves, give us the probability of distant relapse as a function of time, based on the clinical outcome data. As seen in Figure 1, the low and high risk IntClust subtypes indeed separate the patients with low and high risk of distant relapse, confirming the findings in Rueda et al.



(a) IntClust Subtypes

(b) IntClust Risk Labels

Figure 1: IntClust labels capture risk of distant relapse

3.2 Molecular Dataset

Next, we process the molecular data, which consists of 1,191,855 segments spanning the entire tumor genomes of the 1285 patients in our dataset, with each segment denoting the average copy number in that region. While the high dimensionality of this data is an immediate challenge, an additional challenge arises in featurizing this data, as the boundaries of the copy number segments are not well aligned across the 1285 patients in our dataset. In order to both reduce the dimensionality and obtain consistent, useful features, we use the *CNRegions* function from the *iClusterPlus* R package to merge adjacent regions and obtain a final set of 4794 consistent copy number regions for each sample, with adjusted mean copy number values for each region. These can now be used as our features, alongside the clinical features of age at diagnosis, tumor grade, tumor size, and number of tumor-positive lymph nodes.

4 Methods

4.1 5-Fold Nested Cross Validation

We perform a 5-fold nested cross validation procedure to first choose the best parameters for our models and then test the chosen models. To account for the class imbalance, we divide the dataset into 5 stratified folds where each fold contains an equal proportion of positive and negative examples. Next, during each of five iterations, we designate one of the folds as a test set, and use the rest of the 4 folds as the training set. Then, we further divide each training set into 5 folds and during each of five iterations per outer training fold, we use one of the inner folds as a validation set while using the rest as the training set. We use the inner cross validation loop to train each of our models with different sets of parameters and choose the parameters that perform the best across the most validation folds. The metric we use is average precision, which is equivalent to the area under the precision-recall curve. Then, we fit the model with the best parameters to each of the training sets in the outer loop of the cross validation setup and report our performance on the test sets. This nested cross validation serves two purposes. First, it prevents us from over-fitting our models to the test sets since performance on the test set does not affect our choice of model parameters. Second, it allows us to evaluate our model's performance on the entire dataset, which is useful, since we only have 1285 samples in the entire dataset, and this allows us to check for any effect of a particular partitioning of the dataset on the model's performance.

4.2 Logistic Regression

The first machine learning model we consider is logistic regression, since it is one of the simplest models we have in our arsenal. Logistic regression models the distribution of the binary class label y given features x and parameters θ as follows [4]:

$$p(y = 1|x; \theta) = \frac{1}{1 + \exp(-\theta^T x)}$$

Since we have a lot of features and few samples, we use L1 regularization or Lasso regression, as that effectively results in feature selection through setting the coefficients associated with the less important features to 0. For parameter tuning, we try different values of C, the inverse of the regularization strength, including 0.001, 0.01, 0.1, 1, and 10; the one that works best is 0.1.

4.3 Support Vector Machine

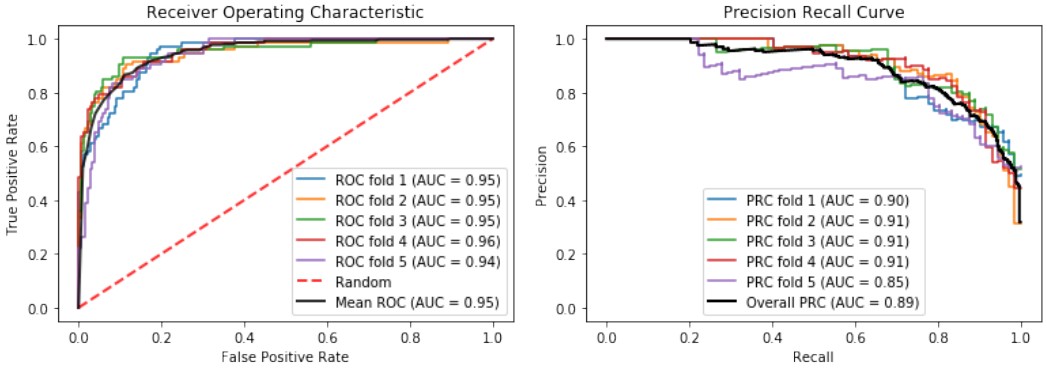
Next, we train SVMs to recover the IntClust risk labels. Here, we perform feature selection by choosing the top 500 features from the genomic copy number regions, according to the highest ANOVA F-values. We try both the Linear and the Gaussian kernel for the SVM as separate models, and search for the best parameters for each. For the Linear kernel, we search for the best value of C among candidates 0.0001, 0.001, 0.01, 0.1, 1, and 10, and find that 0.001 works best on the validation set. For the Gaussian kernel, we search for the best value of gamma, which defines the range of influence of a single training example, among candidates 0.01, 0.001, and 0.0001, and we search for the best value of C among candidates 0.1, 1, and 10. We find that for the Gaussian kernel, a combination of the gamma value of 0.001 and a C value of 1 works best on the validation set.

4.4 Neural Network

The final model we train is a neural network with 1 hidden layer. The neural network uses the Adam optimizer with a learning rate of 0.0002 and employs early stopping to keep the model parameter values from the epoch when the loss on the validation set is lowest, with a patience of 5 epochs. For the hidden layer, we use batch normalization, ReLU activation, and a dropout of 0.5, and for the output layer, we use sigmoid activation. For the loss function, we use binary cross entropy loss with L1 regularization. During cross validation, we try different values for the number of units in the hidden layer, including 100, 200, 300, 400, and 500 as candidates, and we find that 200 units offers the best performance on the validation set.

5 Results

5.1 Recovering IntClust Risk Labels



(a) AUROC (b) AUPRC

Figure 2: Simple neural network models can learn IntClust risk labels from copy number

Model	AUROC	AUPRC	TPR	TNR	FPR	FNR
Logistic Regression	0.94	0.88	0.82	0.90	0.10	0.18
SVM with Linear Kernel	0.93	0.85	0.87	0.86	0.14	0.13
SVM with Gaussian Kernel	0.94	0.86	0.86	0.87	0.13	0.14
Neural Network	0.95	0.89	0.89	0.84	0.16	0.11

Table 1: Performance of models in predicting IntClust risk labels

All of our models have an AUROC above 0.92 and an AUPRC above 0.84, which indicates that they are performing well at recovering the IntClust binary risk labels from just the copy number and clinical data. As shown in Table 1, the neural network has the strongest performance among the different models, with both the highest AUROC and the highest AUPRC. The ROC and PR curves for the neural network for all of the test folds are shown above in Figure 2.

5.2 Log-Rank Test

Using our predicted labels from all of the test folds, we construct cumulative incidence curves showing the observed relative risk for distant relapse for the predicted high risk and low risk patients, obtained from the actual clinical outcome data from METABRIC. The resulting curves, shown in Figure 3 for the neural network predictions, clearly demonstrate how the model predictions can separate the low risk patients from the high risk ones.

In order to formally validate that observation, we conduct a log-rank test to determine if the difference between the predicted low and high risk curves is statistically significant. We find that the predictions from all of the models result in a significant difference in observed risk of distant relapse between the predicted low and high risk groups, with all p-values under 0.005. The corresponding $-\log_2$(p-value) are shown in Table 2.

5.3 Cox Proportional Hazards Model & C-Index

While the log-rank test tells us that there is a significant difference between the risk of distant relapse of predicted low and high risk patients, in order to quantify the association of the risk label with the observed risk, we fit a Cox Proportional Hazards model to the dataset by using the predicted label and the clinical features as the covariates and the observed times to distant relapse as targets.

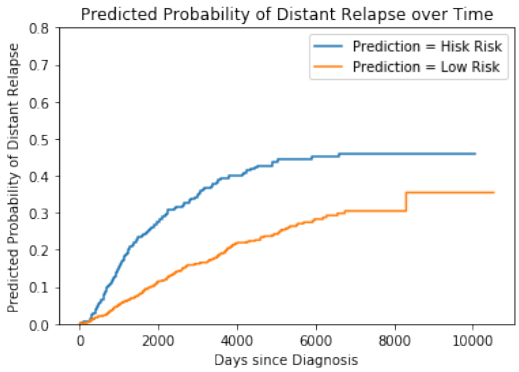


Figure 3: Cumulative Incidence Curves show higher risk for predicted high risk patients

Model:	IntClust	Logistic Reg.	Linear SVM	Gaussian SVM	Neural Net.
-log2(p-value)	36.12	33.73	36.43	37.17	39.88
Hazard Ratio	1.64	1.68	1.72	1.73	1.79
C-Index	0.6756	0.6785	0.6811	0.6809	0.6874

Table 2: Survival Analysis of Risk Predictions

The Cox PH model predicts the expected hazard of an event happening at time t as:

$$h(t) = h_0(t)\exp(b_1X_1 + b_2X_2 + ... + b_nX_n)$$

where X_i are the covariates and $\exp(b_i)$ are the hazard ratios for each covariate. Therefore, a higher hazard ratio for the predicted risk label indicates a stronger positive association of the prediction with the actual risk of distant relapse for a given patient. Table 2 shows that the hazard ratios for the predicted labels surpass those for the original IntClust labels. Furthermore, when we evaluate the accuracy of the Cox PH model using a 5-fold cross validation scheme, we calculate the concordance index (C-Index), which corresponds to the fraction of all pairs of patients whose predicted event times are correctly ordered among all patients that can actually be ordered [5]. Table 2 shows that the C-Index is higher for all of the trained models when compared to IntClust. Notably, the neural network performs best here as well. Overall, this indicates that training machine learning models to learn IntClust labels from copy number and clinical data can lead to successful predictors of distant relapse in breast cancer patients.

6 Conclusion

In this work, we have shown that machine learning models such as logistic regression, support vector machines, and neural networks can all learn to accurately recover the IntClust labels corresponding to the high or low risk of distant relapse in $ER^+/HER2^-$ breast cancer patients from just their copy number and basic clinical data. In doing so, we show that the binary risk predictions generated by these models can actually be used to construct accurate predictors of the risk of distant relapse and the time to distant relapse for a cohort of patients. Furthermore, we show that the risk estimators constructed using the model predictions outperform the ones fitted using the original IntClust risk labels, indicating that the predicted risk labels have a stronger association to distant relapse than the original IntClust risk labels.

Future work on this project will include evaluating these models on other datasets where copy number data is available and the clinical outcomes of patients are known. This will allow us to determine how generalizable these models are to different cohorts of patients.

The code for this project can be found at: https://github.com/soumyakundu/cs229.git

Contributions

S.K. carried out the project and wrote the paper. J.S. and C.C. provided advice.

References

[1] O. M. Rueda, S.-J. Sammut, J. A. Seoane, S.-F. Chin, J. L. Caswell-Jin, M. Callari, R. Batra, B. Pereira, A. Bruna, H. R. Ali, E. Provenzano, B. Liu, M. Parisien, C. Gillett, S. McKinney, A. R. Green, L. Murphy, A. Purushotham, I. O. Ellis, P. D. Pharoah, C. Rueda, S. Aparicio, C. Caldas, and C. Curtis, "Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups," *Nature*, vol. 567, pp. 399–404, Mar. 2019.

[2] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, and S. Aparicio, "The genomic and transcriptomic architecture of 2, 000 breast tumours reveals novel subgroups," *Nature*, vol. 486, pp. 346–352, Apr. 2012.

[3] H. R. Ali, O. M. Rueda, S.-F. Chin, C. Curtis, M. J. Dunning, S. A. Aparicio, and C. Caldas, "Genome-driven integrated classification of breast cancer validated in over 7, 500 samples," *Genome Biology*, vol. 15, Aug. 2014.

[4] S.-I. Lee, H. Lee, P. Abbeel, and A. Y. Ng, "Efficient l1 regularized logistic regression," in *AAAI*, 2006.

[5] V. C. Raykar, H. Steck, B. Krishnapuram, C. Dehing-Oberije, and P. Lambin, "On ranking in survival analysis: Bounds on the concordance index," in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, (USA), pp. 1209–1216, Curran Associates Inc., 2007.