Classify and Predict Molecular Subtypes in Breast Cancer Tumors.

by

Li Liang, Man Shi, Yunhui Liu

Department of Mathematical Sciences
Statistics
GD Harpur

Binghamton University

Binghamton, New York

05/2020

Contents

1. Background Information, Data Cleaning and Visualization

1.1 Background

Breast cancer is cancer that develops in breast cells. Typically, the cancer forms in either the lobules or the ducts of the breast. Cancers developing from the ducts are known as ductal carcinomas, while those developing from lobules are known as lobular carcinomas. The diagnosis of breast cancer is confirmed by taking a biopsy of the concerning tissue. Once the diagnosis is made, further tests are done to determine if the cancer has spread beyond the breast and which treatments are most likely to be effective.

Outcomes for breast cancer depend on the cancer type, the extent of disease, and the age. Worldwide, breast cancer is the leading type of cancer in women, accounting for 25% of all cases. Incidence rates vary widely across the world, from 27 per 100,000 in Middle Africa and Eastern Asia to 92 per 100,000 in Northern

America. It is the fifth most common cause of death from cancer in women, with an estimated 522,000 deaths (6.4 % of the total). Overall survival rates for breast cancer vary worldwide, but in general they have improved. This is because access to medical care is improving in many nations and the majority of breast cancer cases are diagnosed at an earlier and localized stage.

As widely accepted, early detection of breast cancer has an enormous impact on patient's survival. Seeing that genome-wide expression patterns of tumors mirror the biology of the tumors, relating gene expression patterns to clinical outcomes sheds light on the biological diversity of the tumors. A molecular classification of breast cancer, with four main reproducible subtypes (basal-like, HER2 Enriched, luminal A, luminal B) was defined through gene expression profiling and microarray analysis. Therefore, the tumors could be classified into subtypes distinguished by pervasive differences in their gene expression patterns. Identifying the tumors correctly and efficiently would improve the clinical treatment greatly. Thus, the goal of this study is making use of the differential gene expression to identify, classify and predict the molecular subtypes of breast cancer tumors.

1.2 Dataset Basic Information

- **77 cancer proteomes CPTAC itraq.csv**

  The data set includes protein expression values and metadata on 12553 genes from 80 breast cancer patients and 3 healthy individuals.

- **clinical data breast cancer.csv**

  The data set contains clinical data and various breast cancer classifications from 105 breast cancer patients.

- **PAM50 proteins.csv**

  The data set contains the list of genes and proteins used by the PAM50 classification system.

1.3 Preprocessing the Initial Dataset

- Combining Dataset:
  - The first and second datasets have common keys "Complete TCGA ID", so we combined these two datset by using "stringr" in R. The Complete TCGA ID refers to the index of breast cancer patient.
  - As third dataset is the list of genes and proteins used by the PAM50 classification system, we only need "PAM50 mRNA" from the second dataset, which is the recorded molecular subtype.
- Missing Data:
  - We decide to only use the gene expressions that contains more than 99% of the data.

– Original dimension of combined dataset is 80×12554. After dealing with the missing data, the dimension lower down to 80×8018.

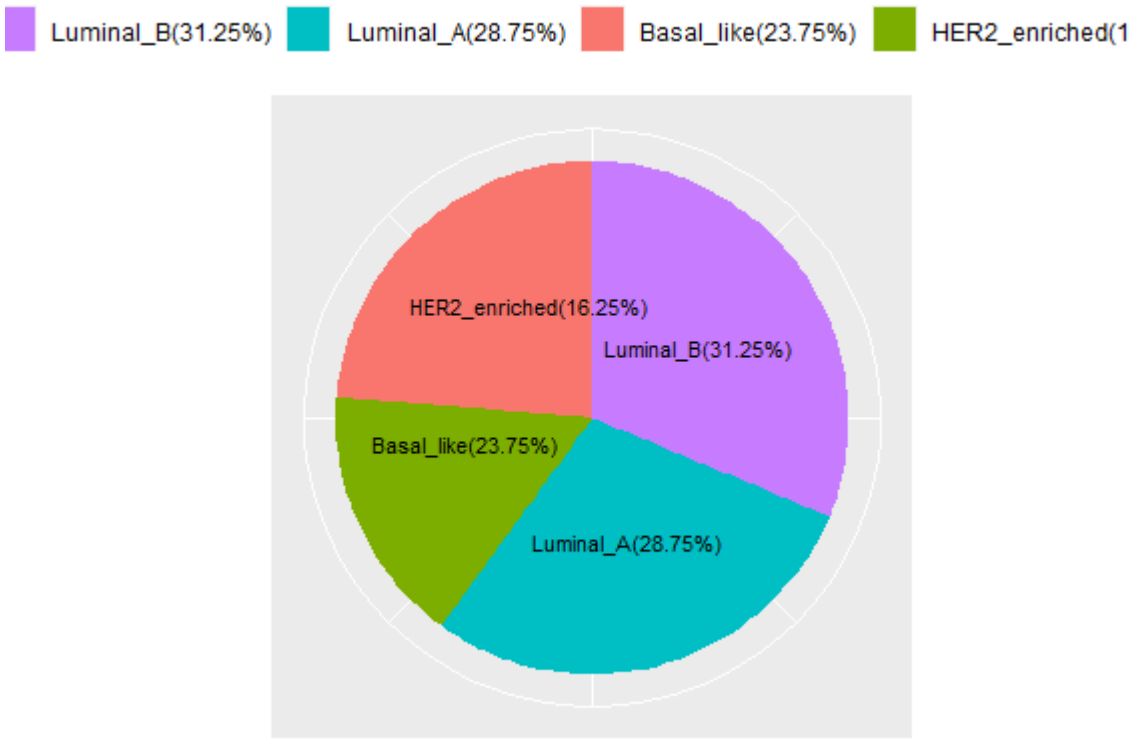

Figure 1.1: Visualization of the Disease Type of Combined Dataset

2. Variable Selection and Results

2.1 Initial Analysis

About the combined dataset, the number of the predictors is 8018 which is larger than the number of the observations that is 80. Therefore, there is no longer a unique least squares coefficient estimate. When the number of observations is very different from the number of predictors, high variance or over-fitting could happen. Therefore, we decided to use dimension reduction techniques "Lasso".

2.2 Lasso Variable Selection

Using a Shrinkage Method known as Lasso, we fit a model containing all 8018 predictors that shrinks the coefficient estimates towards zero. Variables with a coefficient equal to zero will be left out of our final model.

The results of the Lasso data reduction technique are shown below. The visualization of important variables are showing on the Figure 2.1.

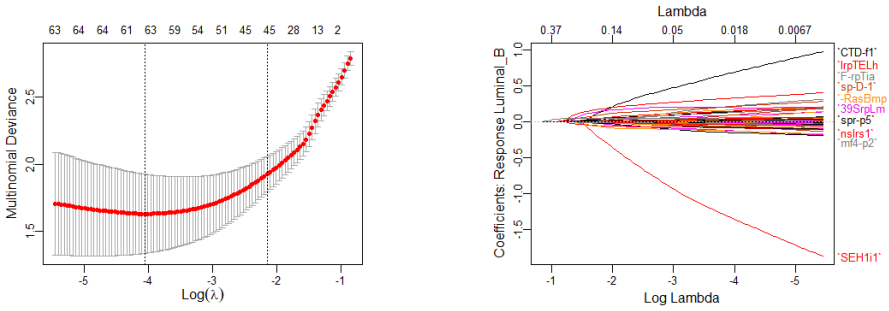

Figure 2.1: Left: Variable selection by LASSO; Right: Importance of Each Variable in LASSO

There are 61 genetic predictors with non-zero coefficients. We could use those 61 genetic predictors to classify the molecular tumor type. Those 61 genetic predictors and their coefficients are in the Appendix.

3. Prediction and Results

In this section, we will use Multinomial Method, Neural Network, K- Nearest-Neighbors, Support Vector Machines and Naive Bayes Methods to conduct predictions. We separate the whole data set into two sets including training set and test set by the ratio of 75% and 25%. Then we do prediction 1000 times get the average value about the result.

3.1 Multinomial Method

The result of prediction accuracy by Multinomial Method is shown on the graph below.

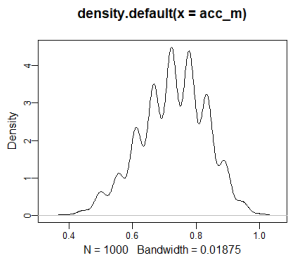

Figure 3.1: Accuracy by Multinomial Method

Table 3.1: Multinomial Method

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.3889	0.6667	0.7222	0.7276	0.7778	1.0000

3.2 Neural Network

The result of prediction accuracy by Neural Network is shown on the graph below.

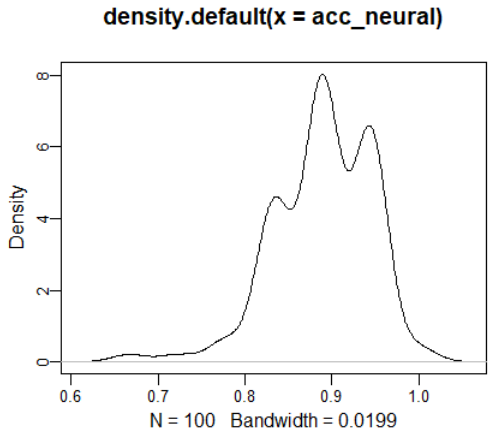

Figure 3.2: Accuracy by Neural Network

Table 3.2: Neural Network

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.6667	0.8333	0.8889	0.8894	0.9444	1.0000

3.3 K- Nearest-Neighbors

The result of prediction accuracy by K- Nearest-Neighbors is shown on the graph below.

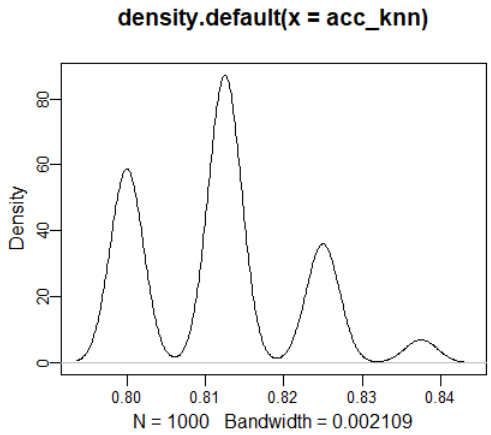

Figure 3.3: Accuracy by K- Nearest-Neighbors

Table 3.3: K- Nearest-Neighbors

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.8000	0.8000	0.8125	0.8119	0.8125	0.8375

3.4 Support Vector Machines

The result of prediction accuracy by Support Vector Machines is shown on the graph below.

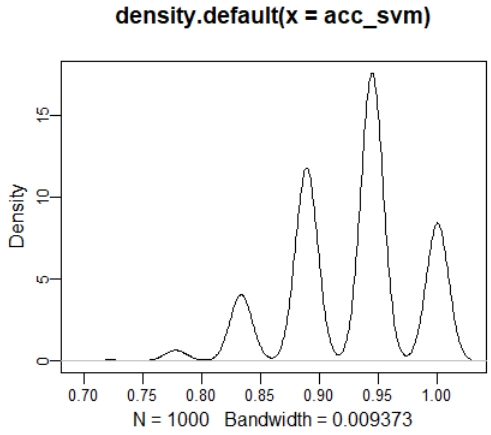

Figure 3.4: Accuracy by Support Vector Machines

Table 3.4: Support Vector Machines

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.7222	0.8889	0.9444	0.9268	0.9444	1.0000

3.5 Naive Bayes Methods

The result of prediction accuracy by Naive Bayes Methods is showing on the graph below.

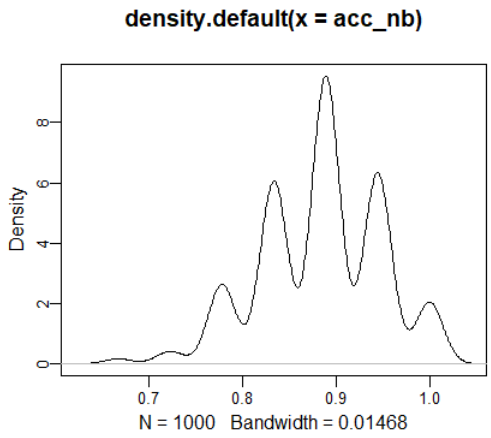

Figure 3.5: Accuracy by Naive Bayes Methods

Table 3.5: Naive Bayes Methods

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.6667	0.8333	0.8889	0.8832	0.9444	1.0000

4. Conclusion

- From the variable selection result of Lasso, 61 genetic predictors are selected. The 61 predictors could be used to classified Breast Cancer.
- From the result of predictions, Support Vector Machine method gives the highest accuracy of 92%.
- By using Lasso to conduct dimension reduction and by using SVM to conduct prediction, we could provide reliable result of Breast Cancer prediction.

5. Reference

- Breast Cancer Proteogenomics Landscape Study, Nature 2016. (n.d.). Retrieved May 03, 2020, from https://cptac-data-portal.georgetown.edu/cptac/s/S029
- (2019). Retrieved 2020, from https://www.healthline.com/health/breast-cancer
- Breast cancer. (2019, June 19). Retrieved May 03, 2020, from https://www.wcrf.org/dietandcancer/breast-cancer

6. Appendix

	LASSO	
2	Name of Variable	Coefficient
3	(Intercept)	-0.0432897547
4	[141] myoferlin isoform b	-0.0099842443
5	[342] cingulin	0.0071351948
6	[774] keratin type I cytoskeletal 23	0.1049062023
7	[1060] arginine--tRNA ligase cytoplasmic	-0.0718788112
8	[1069] TBC1 domain family member 1 isoform 2	0.0590605308
9	[1164] PREDICTED: myomegalin-like	-0.3184289867
10	[1280] KN motif and ankyrin repeat domain-containing protein 1 isoform a	0.3082655159
11	[1281] spermatogenesis-associated protein 5	0.0112467585
12	[1376] receptor tyrosine-protein kinase erbB-2 isoform a precursor	-0.0320687251
13	[1377] epidermal growth factor receptor isoform a precursor	0.0290982815
14	[1378] receptor tyrosine-protein kinase erbB-4 isoform JM-a/CVT-1 precursor	-0.0248213256
15	[1474] mitotic spindle assembly checkpoint protein MAD1	0.0826542518
16	[1570] transportin-2 isoform 2	0.0197184092
17	[1716] UTP--glucose-1-phosphate uridylyltransferase isoform b	0.0937011174
18	[1717] UTP--glucose-1-phosphate uridylyltransferase isoform a	0.1718050836
19	[1848] probable ATP-dependent RNA helicase DDX6	0.0283632793
20	[1916] L-lactate dehydrogenase B chain	0.0436553714
21	[2037] myelin expression factor 2	-0.005350687
22	[2117] interferon regulatory factor 2-binding protein 2 isoform B	0.0504328066
23	[2403] ADP-ribosylation factor-like protein 3	-0.0417912108
24	[2440] insulin receptor substrate 1	-0.1762135576
25	[2570] phosphorylated CTD-interacting factor 1	0.1136325149
26	[2634] SEC14-like protein 2 isoform 2	-0.0016993358

Figure 6.1: Variables Selected and Coefficients by LASSO

27	[2652] carnitine O-acetyltransferase isoform 1 precursor	-0.0555763549
28	[3438] arfaptin-1 isoform 2	-0.2228515334
29	[3609] F-box-like/WD repeat-containing protein TBL1X isoform a	-0.1586502027
30	[3629] oxysterol-binding protein-related protein 2 isoform 2	-0.0010926232
31	[3889] dual specificity mitogen-activated protein kinase kinase 4	-0.0165326223
32	[4088] telomere length regulation protein TEL2 homolog	0.1962590772
33	[4124] guanine nucleotide-binding protein G(i) subunit alpha-1 isoform 1	0.0026223098
34	[4125] guanine nucleotide-binding protein subunit alpha-12	-0.0020187826
35	[4211] nucleoporin SEH1 isoform 1	0.365567186
36	[4278] core histone macro-H2A.2	-0.0051466522
37	[4536] filamin-binding LIM protein 1 isoform b	0.1409398114
38	[4550] glutathione S-transferase P	0.007753581
39	[4657] A disintegrin and metalloproteinase with thrombospondin motifs 4 preproprotein	0.0169170694
40	[4816] cathepsin B preproprotein	0.0089614389
41	[4925] glutamyl-tRNA(Gln) amidotransferase subunit B mitochondrial precursor	0.0005819959
42	[5002] UPF0609 protein C4orf27	-0.0203571038
43	[5218] transcription elongation factor A protein-like 5	-0.0015055416
44	[5714] tRNA (guanine-N(7)-)-methyltransferase subunit WDR4 isoform 1	0.0017798728
45	[5784] armadillo repeat-containing X-linked protein 2	0.0456560568
46	[5819] partitioning defective 6 homolog beta	-0.0094141375
47	[5826] isopentenyl-diphosphate Delta-isomerase 1	0.012899961
48	[5900] cytochrome P450 20A1	0.0021688264
49	[5958] phenylethanolamine N-methyltransferase	-0.0072950657
50	[6333] mortality factor 4-like protein 2	-0.0297287134

Figure 6.2: Variables Selected and Coefficients by LASSO

51	[6662] eukaryotic translation initiation factor 6 isoform a	-0.0022960676
52	[6769] hepatocyte nuclear factor 3-gamma	-0.1177891292
53	[6873] 39S ribosomal protein L40 mitochondrial	-0.0554846216
54	[7053] ras-related protein Rab-25	0.0005951958
55	[7357] protein S100-A13	-0.00489703
56	[7378] stimulator of interferon genes protein	-0.0034614451
57	[7425] transcription initiation factor TFIID subunit 8	0.0211630216
58	[7512] phosphoribosyltransferase domain-containing protein 1	0.0683427075
59	[7515] sorting nexin-24	-0.0188252795
60	[7597] ATP synthase subunit s-like protein isoform 1	-0.0002159
61	[7827] mediator of RNA polymerase II transcription subunit 21 isoform 1	0.0010682398
62	[7836] 39S ribosomal protein L14 mitochondrial	0.0520109864
63	[7841] glutamyl-tRNA(Gln) amidotransferase subunit C mitochondrial	-0.0012035518
64	[7924] eutrophil defensin 3 preproprotein	0.0544790276

Figure 6.3: Variables Selected and Coefficients by LASSO