UMassAmherst School of Public Health & Health Sciences

Leukemia Classification Based on Gene Type Using Lasso Regression and Elastic Net

Chuong Huynh, Evan Ciccarelli, Katrina Wang, Nicole Sanchez Flores

UMassAmherst School of Public Health & Health Sciences

INTRODUCTION

Leukemia classification has been challenging partly because it has traditionally depended on specific biological insights rather than systematic and unbiased methods for identifying tumor subtypes [1]. In this project, we aim to address this challenge by building a model to classify Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) based on gene expression data. Using Lasso regression for feature selection and logistic regression for classification, our approach seeks to improve the accuracy and reliability of leukemia subtype prediction. Data is derived from Golub et al., a sample of 72 people, 7,135 possible predictors. 47 people have Acute lymphocytic leukemia (ALL), while 25 have Acute myeloid leukemia (AML) type of cancer.

RESEARCH QUESTION

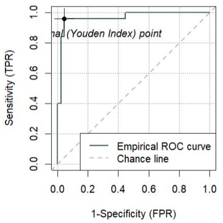
How can we use classification algorithms to create the "class predictor" for AML and ALL cancer type based on the given gene expression?
How can we determine what genes are important/relevant in the distinction of AML and ALL?

METHODS

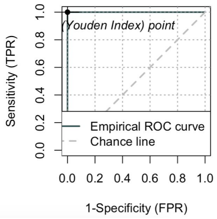
LASSO Regression Model

We conducted two ways to perform the LASSO test. In our first method, we use the 5-fold cross validation to partition the data, the dataset is split into five parts. In each iteration, four parts are used for training and one part is used for testing. In the other method, we use the 3-7 data partition. Due to the small sample size we have, using 70% of data for training provided a substantial amount of data to learn the pattern and relation, and reserving 30% of data allows for a robust evaluation of the model's performance.

Cross Validation



3-7 Data Partition



METHODS cont.

Metrics	X5.Fold.Cross.Validation	Data.Splitting.3.7
Accuracy	0.9167	0.9048
95% CI	(0.8274, 0.9688)	(0.6962, 0.9883)
NIR	0.6528	0.6667
P-Value [Acc > NIR]	1.946e-07	0.01283
Sensitivity	0.9574	1.000
Specificity	0.84	0.7143
Positive Predictive Value	0.9184	0.8750
Negative Predictive Value	0.9130	1.000
Prevalence	0.6528	0.6667
Detection Rate	0.6250	0.6667
Detection Prevalence	0.6806	0.7619
Balanced Accuracy	0.8987	0.8571

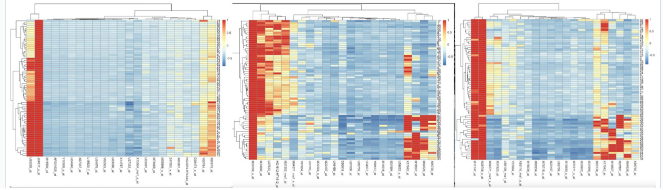
Testing the model at different alpha levels and summarizing the intersection of genes from different alpha.

Different alphas implies the combination of percentage in using Lasso and ridge regression. Below is the reported length of each gene expression at each of four alphas We can observe that as we increase alpha, the closer the number of length of gene expression comes to the number of length in Lasso regression model. Regarding the tuning parameter, as we increase alphas, we can see the penalized lambdas increase and number of predictors decreases and resemble the Lasso regression model.

Alpha.levels	Number.of.genes	Aml	All
0.1	192	112	80
0.25	77	50	22
0.5	51	32	19
0.75	42	30	12
Intersection of genes	40	30	10

RESULTS

Random 70/30 LASSO Full LASSO



We compared the genes we found in the intersection of the various alpha-levels with the genes in the paper that were found to be most relevant to the distinction between ALL and AML. The below table illustrates the genes which were common between our analysis and the paper.

Gene_Code	Name	Our.Coefficient	Our.Findings	Paper.Findings
x59417	Proteasome iota	-0.0000057	ALL	ALL
M31211	Myosin Light Chain	-0.0001601	ALL	ALL
M31523	E2A	-0.0000341	ALL	ALL
M31303	Op 18	-0.0000117	ALL	ALL
Y08612	Rabaptin-5	-0.0002845	ALL	ALL
M55150	Fumarylacetoacetate	0.0000735	AML	AML
X95735	Zyxin	0.0001315	AML	AML
U50136	LTC4 Synthase	0.0000720	AML	AML
M16038	LYN	0.0000853	AML	AML
U82759	HoxA9	0.0003481	AML	AML
M23197	CD33	0.0003306	AML	AML
M84526	Adipsin	0.0000246	AML	AML
M27891	Cystatin C	0.0000226	AML	AML
X17042	Proteoglycan I	0.0000186	AML	AML
Y00787	IL-8 Precursor	0.0000020	AML	AML
M80254	CyP3	0.0000503	AML	AML
M62762	ATPase	0.0000213	AML	AML
M63138	Cathepsin D	0.0000344	AML	AML
X85116	Ebp72	0.0001490	AML	AML

LIMITATIONS & CONCLUSIONS

The largest limitation of our dataset was the low number of participants especially when compared to the number of potential predictors (genes). There were only 72 participants but over 7,000 genes. As a result we chose Lasso and Elastic Net models because they are strong predictive tools and provide insight into which genes are important in distinguishing ALL and AML. The low number of participants made it difficult to test our model. In order to do so we needed to set aside portions of the dataset, exacerbating the issue of limited data. Finally, the analysis of various different alpha-levels for elastic net models helped reveal which genes were important, and 19 of those genes were presented as important in the Golub paper as well.

ACKNOWLEDGEMENTS

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., … Lander, E. S. (1999). Molecular classification of cancer: Class Discovery and class prediction by Gene Expression Monitoring. Science, 286(5439), 531–537. doi:10.1126/science.286.5439.531

Special thanks to Weidong!