FINDING THE BEST METHODS TO CLASSIFY BREAST CANCER GENE EXPRESSION

Idan Cohen, Alejandro Stawsky

Background

Breast Cancer accounts for around 25% of all cancers found in women. In relation to bioinformatics, research into breast cancer prompted the famous discovery of the BRCA1 gene's involvement in BC susceptibility in 1990 and inspired further research to explore the link between genes and cancer. Breast cancer screening is particularly flawed when compared to other cancers as shown by a 2013 Cochrane review of mammography. Therefore, the most accurate diagnoses come from biopsy, and it is using samples like these that we build high performing classifiers.

Goals

- Compare different types of machine learning classifiers on the datasets and see which is has the highest precision
- See if we can learn something about the cancer itself from the structure of the data

Data and Algorithms

For this project we used 3 datasets:

- BC-TCGA normal vs. tumor samples (17,814 genes by 590 samples)
- GSE2034 recurrence tumor vs. non-recurrence tumor samples (12,634 genes by 286 samples)
- GSE25066 Pathologic Complete Response vs. Residual Disease samples (12,634 genes by 492 samples)
- To make the most of our data, we decided to focus on 3 supervised learning algorithms: K nearest neighbors, Support Vector Machines (linear, rbf and polynomial kernels), and Decision Trees

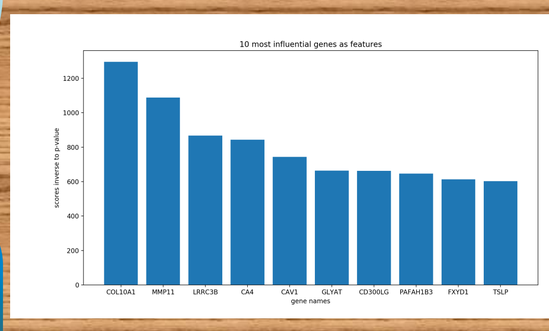
Conclusions

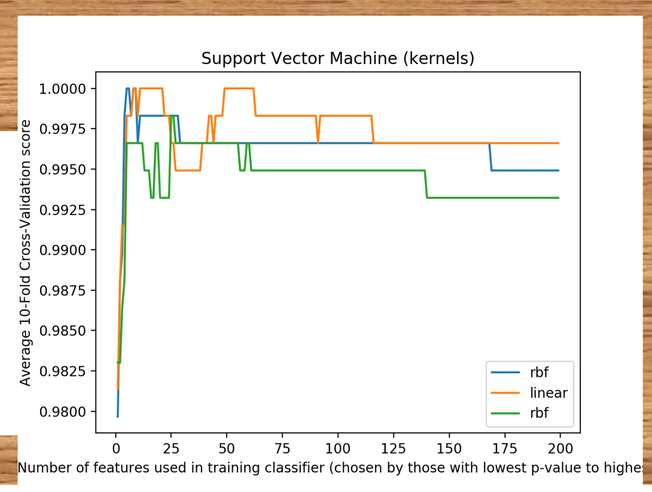
- With regard to the BC-TCGA data, we found very good performance by every algorithm since they were all above 96% precision, and only really needed the gene with the smallest p-value, COL10A1, to get there. (Best is linear SVM and KNN with 4-39 neighbors, and worst is Decision Tree).
- With the GSE2034 data, we found a "sweet spot" for the amount of features needed to include, so we selected the ones with the lowest p-values and added them until we got optimal parameters for all algorithms (Best is linear SVM with best 4,000 genes and the worst is Decision Tree with any combination of genes/features).
- With the GSE25066 data, we only found a significant change in precision w.r.t. number of features in the SVM models, of which the linear and polynomial kernels were the best performing models, with Decision Tree in second place and KNN in third, with a minimum precision of 67% precision.

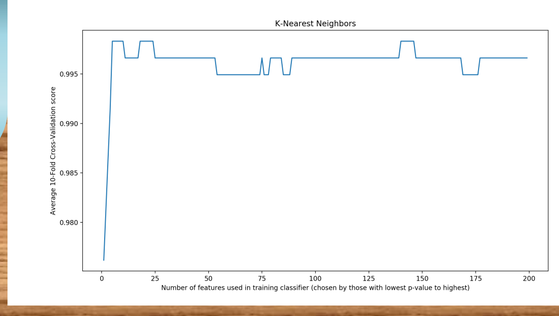
Results

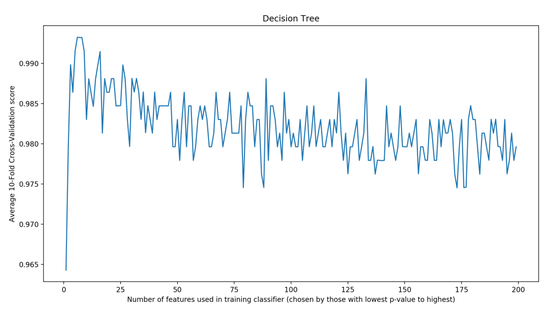
Best Precision	SVM, linear	SVM, Polynomial	SVM, rbf	Decision Tree	K Nearest-Neighbors
BC-TCGA	100%	100%	99.75%	99%	99.55%
GSE2034	84%	84%	76%	60%	58.5%
GSE25066	85%	85.5%	82%	73%	72%

BC-TCGA

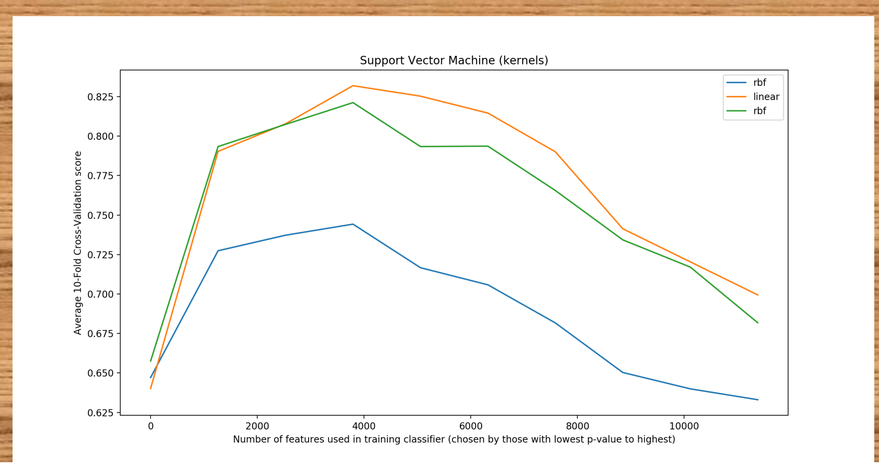


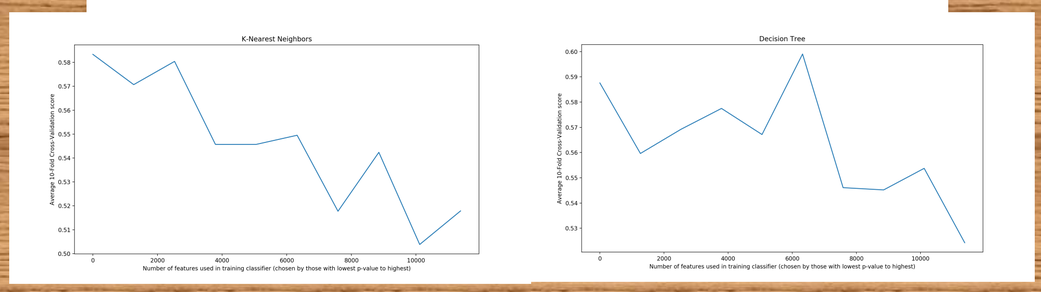






GSE2034





Gorilla gene enrichment: process, function, component

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)
GO:1903047	mitotic cell cycle process	1.51E-21	2.33E-17	3.12 (16534,554,927,97)
GO:0022402	cell cycle process	1.11E-20	8.56E-17	2.80 (16534,873,751,111)
GO:0051983	regulation of chromosome segregation	6.15E-17	3.17E-13	9.76 (16534,99,445,26)
GO:0051301	cell division	1.12E-14	4.34E-11	4.36 (16534,372,469,46)
GO:0051783	regulation of nuclear division	1.29E-14	3.98E-11	5.86 (16534,197,444,31)

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)
GO:0008017	microtubule binding	7.15E-10	3.27E-6	3.33 (16534,223,824,37)
GO:0005515	protein binding	2.61E-8	5.98E-5	1.10 (16534,10844,1716,1236)
GO:0015631	tubulin binding	3.53E-8	5.38E-5	2.74 (16534,300,824,41)
GO:0008092	cytoskeletal protein binding	6.05E-8	6.92E-5	2.04 (16534,907,652,73)

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)
GO:0044815	DNA packaging complex	6.62E-13	1.3E-9	3.80 (16534,99,1715,39)
GO:0000786	nucleosome	1.34E-12	1.31E-9	3.88 (16534,92,1715,37)
GO:0044427	chromosomal part	2.5E-11	1.63E-8	1.73 (16534,886,1715,159)
GO:0032993	protein-DNA complex	2.62E-11	1.28E-8	2.91 (16534,159,1715,48)

References

- Comparison among dimensionality reduction techniques based on Random Projection for cancer classification, Xie et al., 2016
- http://cbl-gorilla.cs.technion.ac.il/#ref
- https://doi.org//10.1002/14651858.CD001877.pub5