Clustering and Classifying Autism

Rachael Aikens, (raikens@stanford.edu) and Brianna Kozemzak (kozemzak@stanford.edu)
Stanford University Department of Biomedical Informatics, Wall Lab

The iHart Consortium has helped to collect one of the largest Autism Spectrum Disorder (ASD) datasets ever, including genetic and behavioral data for several thousand ASD Cases and Controls.

This offers us unprecedented opportunity to take Machine Learning Approaches to two major Autism Research problems:

Aim 1: Clustering Autism Subtypes

Goal: : Develop a cluster validation tool kit and use it to analyze clustering results

The Problem:

ASD can manifest over a broad spectrum of symptoms, from great intellectual and communication disability to near-normal 'high-functioning' forms. As a result, it is often asked whether ASD is in fact composed of some number of Autism 'sub-types' that are best diagnosed, studied, and treated in different ways.

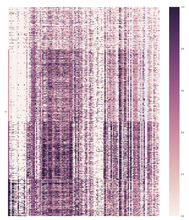
Data:

- 13,493 individuals
- 123 features from ADOS and ADI-R instruments
- Diagnostic, medical, demographic, etc. labels

Prior Work in Wall Lab:

- Imputed missing values and clustered data using generalized low rank model with logistic loss
- Crisp and soft k-means clusterings were created for k = 1, 2, ..., 6.

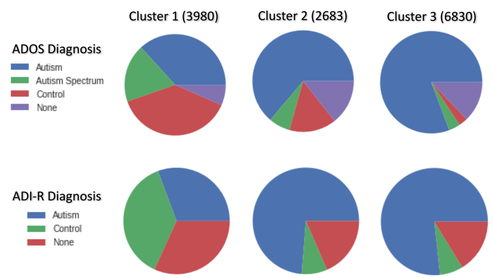
Feature Heat Maps:



Features on x-axis and examples on y-axis, sorted by cluster. This was too complex to be useful, so we looked only at the centroids of the cluster (a low rank representation of examples) instead.

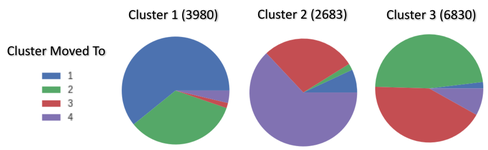
Features on the x-axis and centroids on the y-axis. Lighter feature values usually indicate more neurotypical behavior. We see separation of neurotypical individuals from atypical individuals and then a mixed cluster.

Label Pie Charts:



Pie charts were generated for 29 different labels including diagnostic, demographic, and computed ADOS/ADI-R labels. The control group appears to separate from the ASD individuals.

Individual Movement:



Movement between clusters was not random. This indicates some common underlying features driving cluster formation for all k values.

Conclusions and Future Work:

Conclusions

- "Best" clustering result was soft k-means with k=3, where each individual is assigned to a single cluster based on maximum partial membership
- Why? Clusters are separated by diagnosis, medical history, and computed ADOS/ADI-R labels without creating indistinguishable extra clusters

Future work will:

- Employ methods to work directly with the soft clustering results by using enrichment tests developed for soft clustering [1] and implementing weighted membership for pie charts
- Apply other clustering methods to data set and compare with k-means and soft k-means results

Aim 2: An Autism Genetic Risk Score

Goal: Build a genetic risk predictor for ASD

The Problem:

Autism is a *complex disease* – it is determined about 50% by genetics and 50% by a person's environment

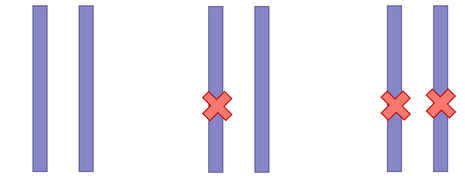
$$\text{Genotype} + \text{Environment} = \text{Phenotype}$$

As a result, ***it is impossible to perfectly predict autism from genetics***.

However, an imperfect classifier can:

- give us a measure of a person's genetic risk of autism
- provide intuition about which genetic features are most predictive of disease.

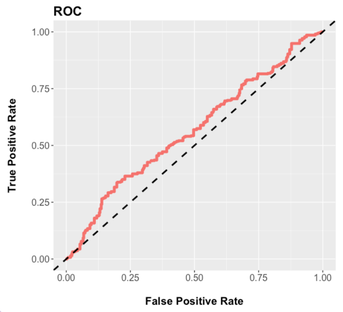
The Feature Set:



Each genome is shown as a 1109 × 1 binary describing where each person has a loss-of-function in a gene.

A Logistic Regression Classifier:

We first trained a Logistic Regression Classifier because these models are often simple to interpret.

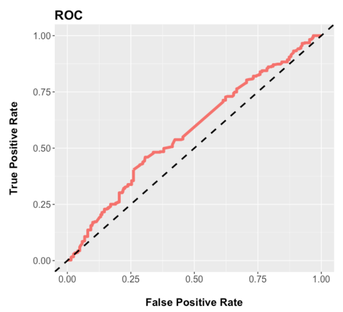


F1 score: **0.634**

Area under ROC: **0.565**

A Gradient Boosted Classifier:

We also trained a gradient boosted tree classifier to capture non-linear gene-gene relationships.



F1 score: **0.647**

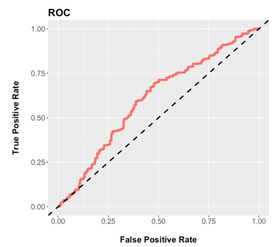
Area under ROC: **0.580**

Conclusions and Future Work:

Our best performance is achieved from averaging the predictions from the two classifiers above (see right). ***This classifier outperforms previous methods (best AU-ROC = 0.54 [2])***, showing promise as a genetic risk score predictor for ASD.

Future work will:

- Continue to optimize ensemble and non-linear classification models
- Analyze feature importance to infer which genetic variants are most predictive



F1 score: **0.642**

Area under ROC: **0.602**

Acknowledgements

This work would not have been possible without the help of Dennis Wall, Kelley Paskov, and the other members of the Wall lab, as well as the funding and computing resources of the Wall Lab and the Stanford University School of Medicine. Thanks also to the Machine Learning instructors and TAs.

1. Phillips, R. D. et al. Enrichment Procedures for Soft Clusters: A Statistical Test and its Applications. (2010). 2. Unpublished work from the Wall lab, Stanford University