Comparing the Performance of Outlier Detection Algorithms in Detecting Antigenic Transitions in Influenza A

Sofia McDonough[1], Omid Arhami[2], Alpha Forna[2], Pejman Rohani[2]

[1]Florida State University, [2]University of Georgia

Background

- The influenza virus mutates over time (antigenic drift), resulting in variants in different clusters. Clusters are identified based on their antigenicity (ability of antibodies to neutralize the virus).
- It is important to identify the emergence of new antigenic clusters in order to determine what viruses to include in seasonal flu vaccines, since individuals previously infected with a virus from one cluster will not be immune to a virus from a new cluster.
- Antigenic cluster transitions are usually identified with hemagglutination inhibition assays, which require a lot of time and resources.
- **NIAViD (Novel Influenza A Virus Detector)** is an unsupervised machine learning model that was developed to identify new antigenic clusters in Influenza A.

Objective

Compare the effectiveness of different outlier detection algorithms in identifying antigenic shifts in the flu virus using the **NIAViD** model.

Methods

- NIAViD is an outlier detection algorithm to detect transitions based on physicochemical properties, calculated from HA gene sequence.
- The physicochemical properties are:
  - Hydrophobicity
  - Charge
  - Boman's Index
  - Instability
  - Isoelectric Point
- Once the five physicochemical properties are calculated for each sequence, an anomaly detection algorithm is employed to identify whether a new virus is an outlier or is part of an existing cluster.
- NIAViD has previously employed Isolation Forest and One Class SVM as outlier detection methods.
- Here, I tested the performance of two new methods, **ECOD** and **HDBSCAN**:

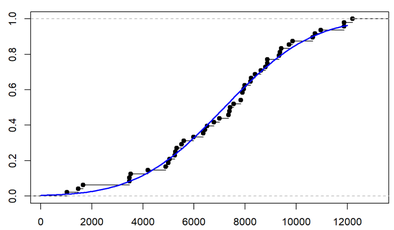

Figure 1: Empirical cumulative distribution function, which ECOD uses to detect outliers. From University of Virginia Library 2020 .

ECOD is a distribution-based algorithm that fits an empirical cumulative distribution function to each of the five properties.

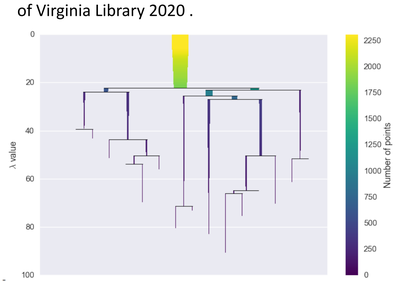

Figure 2: Condensed cluster tree used in HDBSCAN. From McInnes et al. 2017.

HDBSCAN is a hierarchical clustering algorithm that finds the minimum spanning tree for the data.

Results

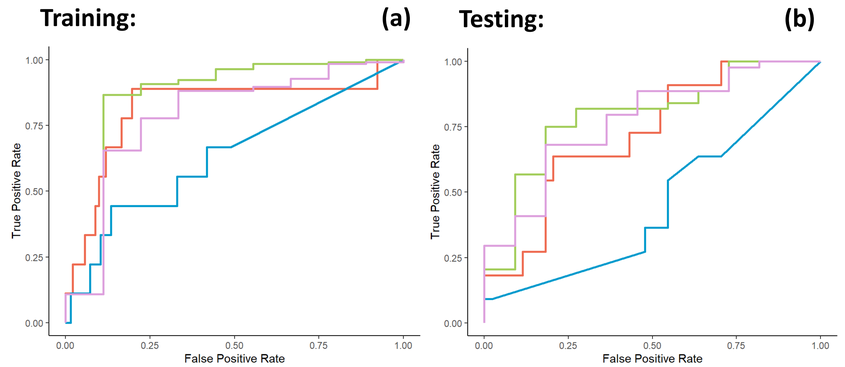

Figure 3: ROC (receiver operating characteristic) curve for each of the four algorithms in the training (a) and testing (b) phases.

	True Positives		False Negatives		False Positives		True Negatives	
Algorithm	Training	Testing	Training	Testing	Training	Testing	Training	Testing
ECOD	5	4	4	7	59	4	165	40
HDBSCAN	1	1	8	10	5	1	189	43
Isolation Forest	8	8	1	3	57	8	137	36
One Class SVM	8	9	1	2	112	18	82	26

Table 1: Predicted and actual classification for each algorithm expressed as true positives, false negatives, false positives, and true negatives.

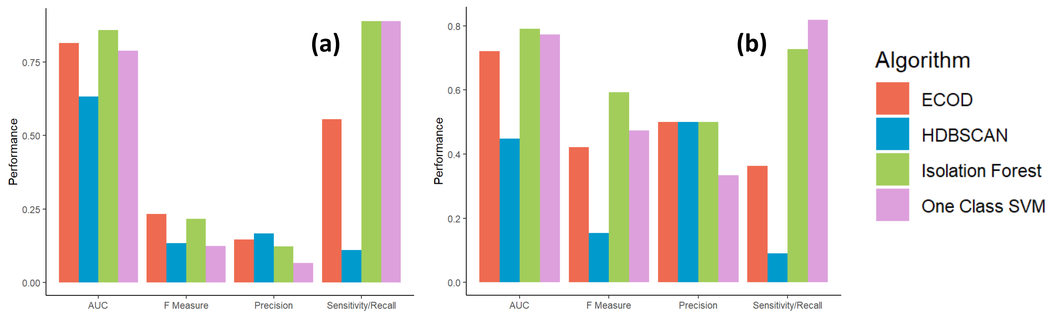

Figure 4: Performance measures for each of the four algorithms in the training (a) and testing (b) phases, with an outlier threshold of 0.5 for HDBSCAN

- ECOD, Isolation Forest, and One Class SVM had similar ROC AUC in both the training and testing phases.
- Isolation Forest and One-Class SVM had the highest sensitivity in both phases, followed by ECOD.
- HDBSCAN had lower AUC and much lower sensitivity than the other three algorithms.

Conclusions and Future Directions

- NIAViD's performance is comparatively higher using Isolation Forest and One Class SVM than using ECOD and HDBSCAN.
- In the future, additional algorithms could be included in our pipeline to improve robustness so that NIAViD can be used to aid in vaccine development by predicting antigenic cluster transitions.

Acknowledgements

Thank you to the members of the Rohani Lab, the coordinators of the PopBio Program, and my fellow REU students. Support for this research was provided by the National Science Foundation (grant #1659683) through the Population Biology of Infectious Diseases Undergraduate Research program.

References

Ford, C. (2020). Understanding empirical cumulative-distribution functions. *UVA Library,* https://data.library.virginia.edu/understanding-empirical-cumulative-distribution-functions/

Li, Z. et al. (2015). ECOD: Unsupervised outlier detection using empirical cumulative distribution functions. *Journal of LaTeX Class Files 14*(8):1-13.

McInnes, L et al. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software, 2*(11):205.

Smith, D. et al. (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science, 305*(5682):371-376