A Novel Query Efficient Algorithm for Active Covering

Evan Archer, Electrical Engineering
Mentor: Dr. Gautam Dasarathy, Assistant Professor
School of Electrical, Computer, and Energy Engineering

Background and Introduction

- Active Covering is a machine learning problem where the goal is to find all positive cases in a set of data, in as few queries as possible. Active Covering appears in clinical trials, drug discovery, etc. Where the goal is to find positive cases in as few tested candidates as possible.
- Three different algorithms are tested against each other, first the Active Explore-then-Commit Learner, which initially samples the data then queries the closest node to a positive node (2). Next there is S^2 which uses label prediction and the graphs cut edges to isolate the positive cluster and label all the points (1). Lastly, the Improved algorithm uses S^2 and the epsilon neighborhood factor from Active Explore to decrease query cost, by not sampling the known negative nodes 3 closest nodes.

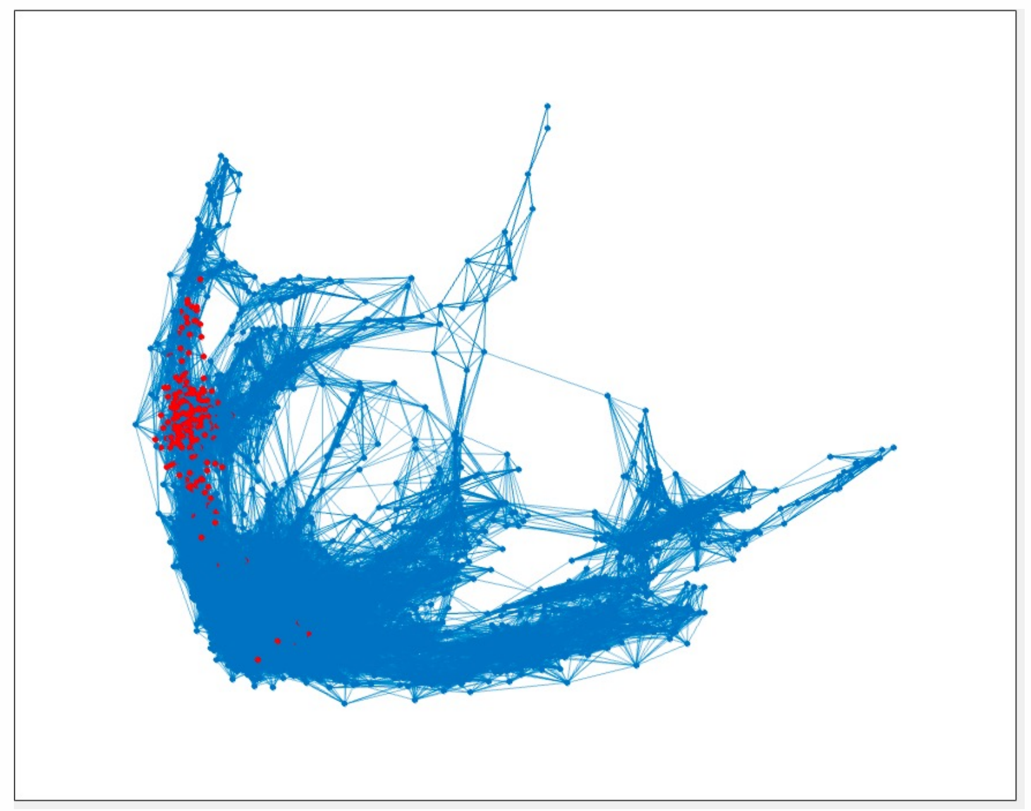
Fig 1. Data with 10 connections with red dots representing positives

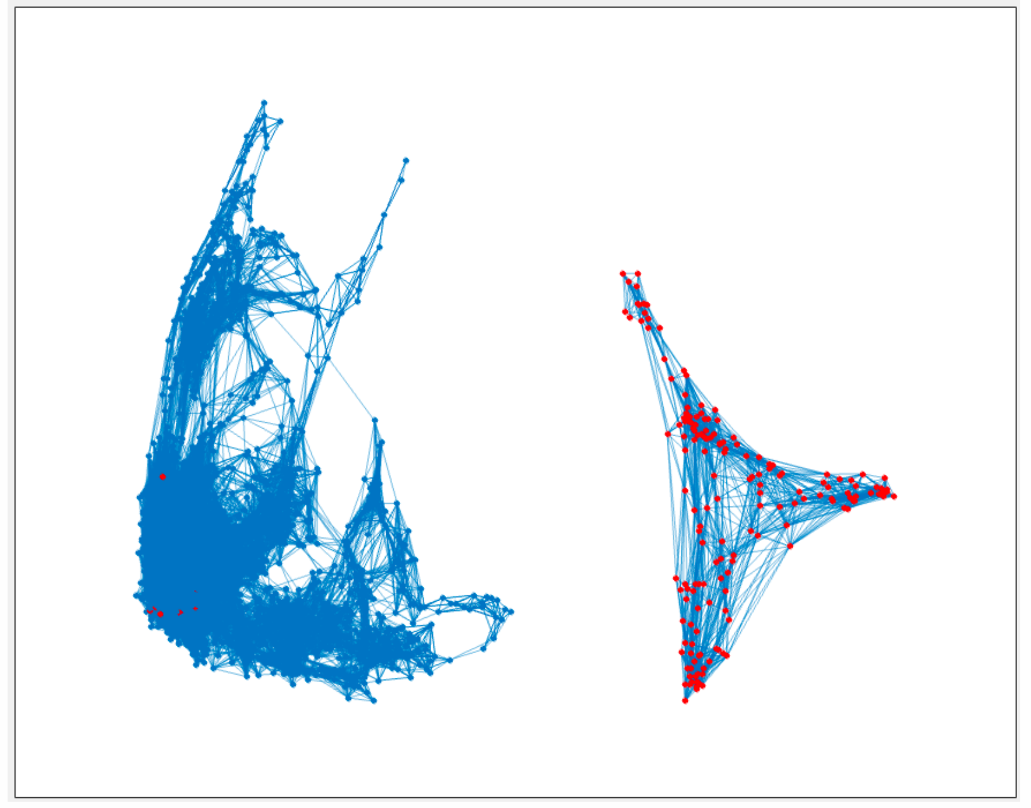
Fig 2. Output of S^2 and Improved algorithm showing the isolated positives.

Problem Set Up

- Each algorithm was tasked with finding 80% of positive nodes in the UCI Letters Recognition data set. This allows us to ignore outlier cases.
- Letter data was processed into adjacency matrices, for the algorithms to use.
- Test runs were done with changing number of nodes and number of connections for S^2 and the Improved Algorithms.
- 17 connected nodes is the minimum connections that result in a one connected cluster of nodes.
- Active Explore Initially samples 5% of the data.

Results

Information on Run				Query Cost Data			
Algorithim	# of Runs	# of Nodes	Connected Nodes	Average	Min	Max	STD DEV
Active	25	20000	20000	1601.84	1589	1611	6.1825
S2	25	20000	17	1075.16	881	1704	287.948
Improved	25	20000	17	1061.12	858	1860	279.871
Active	20	20000	20000	1599.65	1590	1612	6.81542
S2	20	20000	5	358.1	304	632	73.433
Improved	20	20000	5	363.8	285	690	107.627
Active	100	5000	5000	411.44	393	1035	63.1721
S2	100	5000	10	331.01	262	654	72.7456
Improved	100	5000	10	323.96	259	716	72.6742
Active	100	5000	5000	404.28	395	417	4.18047
S2	100	5000	5	213.23	157	444	51.7209
Improved	100	5000	5	204.35	161	335	36.3633

Conclusions

- S^2 and the Improved algorithm are 2-4x more efficient than Active Explore depending on conditions.
- S^2 and the Improved algorithm benefit as data size, positives nodes, cluster size increase.
- There is a relationship between number of nodes and connections for the performance of S^2 and the Improved algorithm.

Future Work

- Further explore the connectedness factor to make it an active connectedness factor
- Look into the optimal number of connected points for S^2 and the Improved algorithm.
- Optimizing S^2 for low density cluster cases

Acknowledgments

I would like to thank Dr. Gautam Dasarathy for his guidance and input on this research project. I would also like to thank ASU Research Computing for allowing me to run my code on the Agave Supercomputer.

[1] Dasarathy, Gautam & Nowak, Robert & Zhu, Xiaojin. (2015). S2: An Efficient Graph Based Active Learning Algorithm with Application to Nonparametric Classification.

[2] Jiang, Heinrich & Rostamizadeh, Afshin. (2021). Active Covering.

MORE
Masters Opportunity for Research in Engineering

ASU Ira A. Fulton Schools of Engineering
Arizona State University