LCS-DIVE: Detecting and Characterizing Signal in Complex Datasets

Robert Zhang (SEAS 2022, Wharton 2022) with mentorship from Dr. Ryan Urbanowicz (Penn Medicine - Dept. of Biostatistics, Epidemiology, and Informatics)

funded by Grants for Faculty Mentoring Undergraduate Research. Emails (robertzh@wharton.upenn.edu, ryanurb@pennmedicine.upenn.edu). Site: www.med.upenn.edu/urbslab

What is LCS-DIVE?

LCS-DIVE (LCS Discovery and Visualization Environment) is a general, automated method for any data researcher to discover and visualize the signal in a wide variety of complex classification problems. Namely, LCS-DIVE is able to extract and visualize human interpretable information about instance heterogeneity, epistatic patterns of association, and feature importance from the data for further exploration and knowledge discovery. LCS-DIVE uses the ExSTraCS Learning Classifier System as its core modeling algorithm. ExSTraCS is a powerful classification algorithm that has been shown to perform extremely well on problems containing complex patterns and noise [1] [3].

Project Objective

Having built LCS-DIVE, we tested its performance on a wide variety of real world, as well as simulated data. Simulated datasets were meant to simulate specific, predetermined complex patterns, such as main effects, heterogeneity, epistasis, and heritability. In this poster, we will present the results from a few n-bit Multiplexer (MUX) problems, as well as from a real world pancreatic cancer dataset.

The Multiplexer Benchmark Problem

Highly epistatic (data features are not independent) and heterogeneous (multiple, distinct subsets of data features give the same class) problems are challenging to solve using standard ML algorithms (e.g. deep learning). However, these types of problems are prevalent in day to day life: notably in epidemiology for the diagnosis and treatment of chronic conditions. The Multiplexer Problem is a classification problem that involves a high amount of epistasis and heterogeneity. Hence, it is an excellent benchmark for ML algorithms. The attributes in themultiplexer problem are split into *n* address bits and *2^n* register bits.

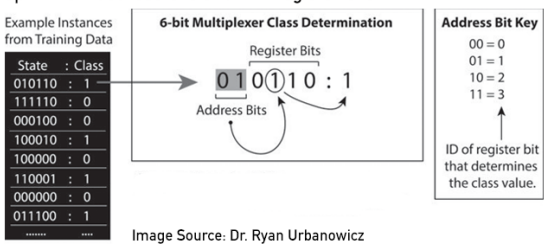


Image Source: Dr. Ryan Urbanowicz

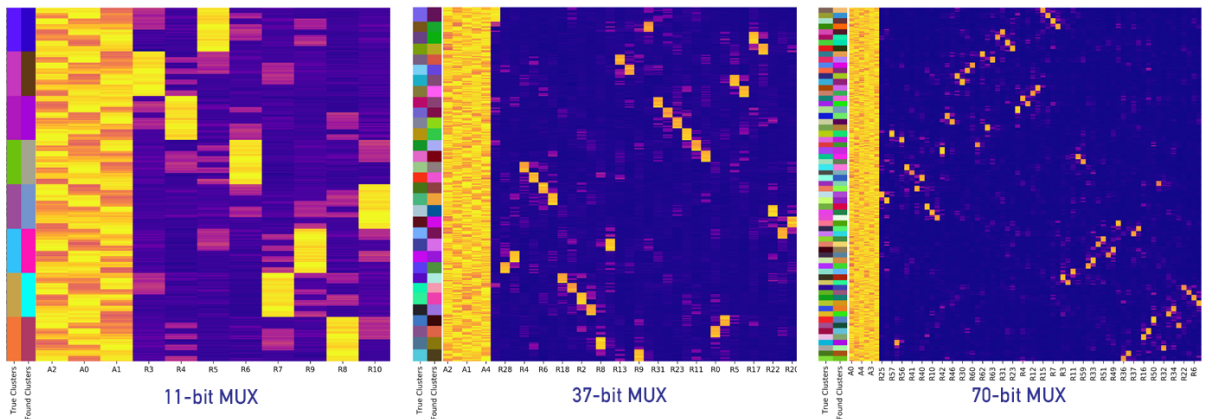
Citations:

[1] Urbanowicz, Ryan J, and Jason H Moore. "ExSTraCS 2.0: Description and Evaluation of a Scalable Learning Classifier System." Evolutionary Intelligence, U.S. National Library of Medicine, Sept. 2015

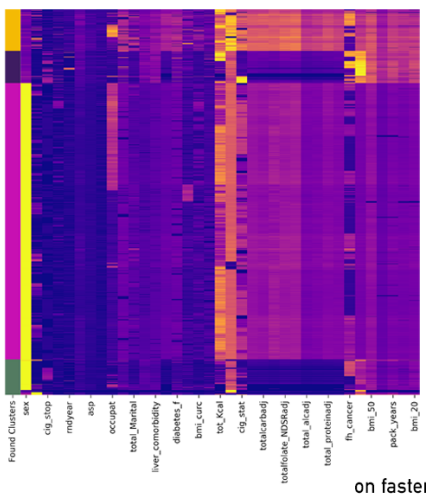
[2] Urbanowicz, Ryan J, and Jason H Moore. "Instance-Linked Attribute Tracking and Feedback for Michigan-Style Supervised Learning Classifier Systems" Aug. 2012

[3] Urbanowicz, Ryan J, and Ambrose Granizo-Mackenzie. "An Analysis Pipeline with Statistical and Visualizaion-Guided Knowledge Discovery for Michhigan-Style Learning Classifier Systems", IEEE Comput Intell Mag. Nov. 2012

LCS-DIVE Results from the Multiplexer Problem (11-bit to 70-bit MUX Problems)



The above diagrams show a core output of LCS-DIVE: AT score visualization. In the process of training, ExSTraCS generates 'attribute tracking scores' for each feature of each instance, which represent a 0 - 1 value of how predictive that each feature was for that instance [2]. Higher AT scores indicate a feature with higher predictive value. LCS-DIVE then clusters these AT-scores to find instances with similar predictive features. The effectiveness of AT visualization is shown with the MUX problems. LCS-DIVE finds that the address bits were predictive for all instances, with the addition of a single predictive register bit. Our framework is thus able to **perfectly** characterize the heterogeneity in these problems.



LCS-DIVE Results from the Pancreatic Cancer Dataset

The Pancreatic Cancer Dataset is a real-world dataset containing the dietary habits, sex, race, and other environmental information of patients. It also contains a binary label for whether each patient has pancreatic cancer. LCS-DIVE finds that **sex** and **race** are the most predictive features in the dataset. Both features are main effects, but also have a small epistatic effect. DIVE finds that for 80.6% of the patients, sex and race alone can nearly perfectly characterize pancreatic cancer. However, the remaining 19.4% of patients follow the opposite association from the primary sex main effect, and thus make up the vast majority of predictive error. We conclude that the features in his dataset did not have enough heritability for the model to properly detect between these two patient subgroups. LCS-DIVE recommends that we add new features that can provide the model with the proper predictive information to characterize these subgroups.

Conclusion and Acknowledgements: LCS-DIVE is shown to be highly effective in characterizing the signal found in a trained ExSTraCS model. Future work mainly focuses on faster clustering methods, more precise interpretation of dataset noise and epistasis, and further testing on real-world datasets. I would like to thank Dr. Ryan Urbanowicz for his insight that greatly assisted the development of this project.