Binary and Multi-Class Biomedical Data Classification using Automated ML Pipelines

Riju Datta, Ryan J. Urbanowicz, Ph.D.

Perelman School of Medicine, University of Pennsylvania

Department of Biostatistics, Epidemiology & Informatics

Large-scale SNP Dataset Analysis using AutoMLPipe-BC

Background:

- Automated machine learning (AutoML) pipelines provide robust mechanisms to conduct rigorous machine learning analyses in their entirety (from data filtering and data preprocessing to model training and performance evaluation)
- With upsurge of "big data" availability in biomedical informatics, automated machine learning methods have become invaluable in extracting predictive features from large data sets and identifying associations between such features
- Analysis of complex "omics" data, among other dataset categories, can potentially offer a plethora of insights into clinical relevance of different medical targets
- AutoMLPipe-BC analyzed large-scale single nucleotide polymorphism (SNP) data collected from 810 patients with congenital heart disease (CHD) in order to determine which SNPs were most closely correlated with disease phenotypes

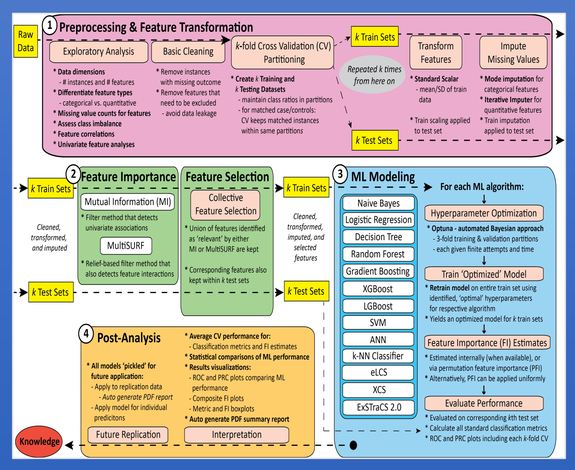

Fig. 1 – Schematic of AutoMLPipe-BC Pipeline[1]

Methods:

- Original SNP dataset contained 184,527 unique SNP features; thus, exploratory analysis was conducted to identify SNPs with a statistically significant association between feature value and class outcome, yielding 3,452 features
- Sci-kit Learn API's[2] *IterativeImputer* function, which imputed missing feature values for a given instance using a function of all other features for that instance (more useful in smaller feature spaces), was replaced by univariate *SimpleImputer* function to minimize computational expense during data preprocessing
- AutoMLPipe-BC was then run using reduced dataset and with 8 out of 13 algorithms in the ML modeling phase (omitting Logistic Regression, XGBoost, SVM, Artificial Neural Networks, and K-Nearest Neighbors because of their inability to scale computationally to larger feature spaces)

Analysis and Results:

- AutoMLPipe-BC statistical analysis on CHD dataset generated high evaluation metrics on select ML models, with average ROC_AUC scores of **0.990**, **0.978**, and **0.955** from the Naive Bayes, Random Forest, and LGBoost ML algorithms, respectively (Fig. 2)
- Precision-Recall curve indicated high positive predictive power across large range of sensitivity values, with Naive Bayes, Random Forest, and LGBoost algorithms having average precision scores of **0.991**, **0.979**, and **0.955**, respectively (Fig. 3)
- 3 SNPs, **rs1406002_A**, **rs2765283_T**, and **rs4522809_G**, received highest feature importance scores on composite feature importance bar plot, indicating a correlation between those SNPs and CHD phenotypes (Fig. 4)

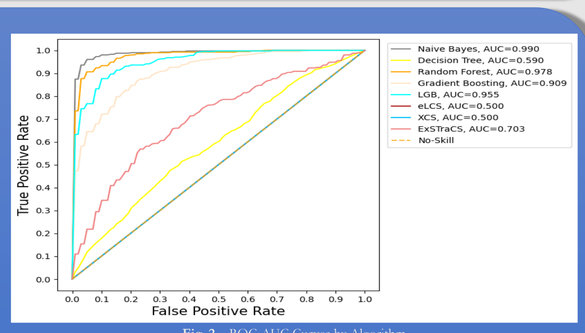

Fig. 2 – ROC-AUC Curves by Algorithm

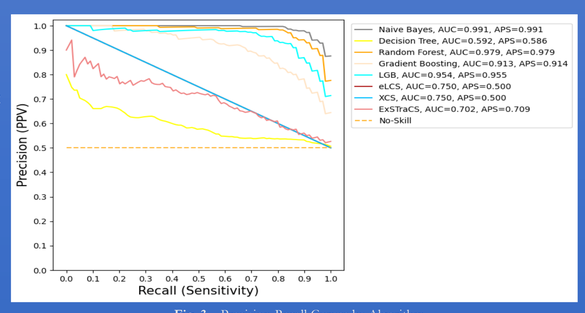

Fig. 3 – Precision-Recall Curves by Algorithm

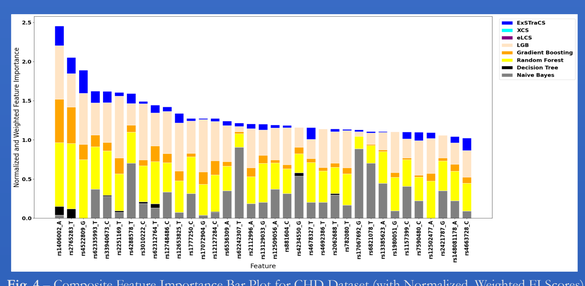

Fig. 4 – Composite Feature Importance Bar Plot for CHD Dataset (with Normalized, Weighted FI Scores)

Conclusion:

- AutoMLPipe-BC successfully conducted a full statistical analysis of SNP data of patients with CHD, not only demonstrating AutoMLPipe-BC's viability to filter and analyze large datasets, but also its potential in the biomedical domain to extract key features from patient, EHR, and/or "-omics" data (among other sources) and generate meaningful clinical insights
- Going forward, areas to continue research include:
  - Developing novel preprocessing and data wrangling methods to analyze noisier, more complex or even non-tabular datasets
  - Implementing more computationally expensive, multivariate imputation steps using iterative runs through batches of data instead of a full dataset

AutoMLPipe-MC – A Multi-class Extension of AutoMLPipe-BC

Background:

- Limitation of AutoMLPipe-BC – can only conduct binary data classification; AutoMLPipe-MC is a novel statistical analysis pipeline extending original pipeline using a One-vs-Rest multi-class strategy (i.e., a classifier is trained to recognize and predict one and only one class)
- One-vs-Rest allows AutoMLPipe-MC to accommodate multi-class supervised learning analyses, generate relevant visualizations that compare the performance not only across different ML algorithms, but also across classifiers themselves based on how accurately each classifier predicts class values
- Statistical analysis pipeline was tested on both sample 3-class and 9-class SNP datasets[3], as well as real-world dataset[4] containing information on protein expression levels in the cerebral cortices of 8 different categories of mice
- Mice were tested on their associative learning abilities following an electric stimulus and subsequent drug treatment; 3 different criteria for mice categories: genotype (normal or trisomic), treatment (placebo or drug recipient), and behavior (electrically stimulated or left undisturbed); goal of statistical analysis was to determine which proteins would differentiate between the distinct classes of mice

Methods:

- Primary edits to AutoMLPipe-BC pipeline were in ModelJob.py and StatsJob.py (and their corresponding "Main" files, through which the job files were run on our lab's I2C2 Computing Cluster)
- Multi-class strategy in ModelJob.py implemented using Label Binarizer (creating positive labels for values of one class and negative labels for all others), OneVsRestClassifier wrapper (training fitted multi-class models that could recognize distinct classes)
- Kept majority of feature importance metrics in StatsJob.py (Fig. 5); introduced several new performance evaluation metrics suited for multi-class classification, including Confusion Matrix plot and multi-class extension of ROC curve plot

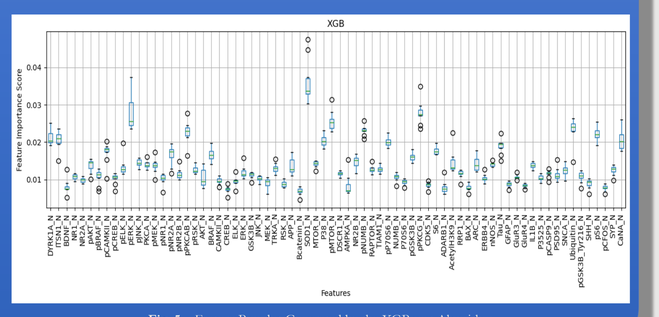

Fig. 5 – Feature Boxplot Generated by the XGBoost Algorithm

Analysis and Results:

- AutoMLPipe-MC pipeline reported extremely high performance accuracy on the mice protein expression dataset, across all evaluation metrics (including balanced accuracy, roc_auc score, and TP/FP distribution)
- Multi-class ROC curve plots (Fig. 6) showed, even for strict decision thresholds where positive values are only assigned to a minute proportion of labeled data, each classifier rapidly distinguished all values of its class from other values
- Normalized confusion matrix plot (Fig. 7) highlighted high degree of accuracy in each classifier's true/false positive labeling
- Composite feature importance bar plot (Fig. 8) demonstrated clinical relevance of AutoMLPipe-MC statistical analyses; gave highest scores to key proteins involved in mice learning, including:
  - **SOD1** (whose overexpression has been linked to learning and memory deficits in mice)
  - **pPKCG** (one of 21 proteins discriminating between class of mice unable to accurately learn and three classes of mice that are able to perform successful associative learning[5])

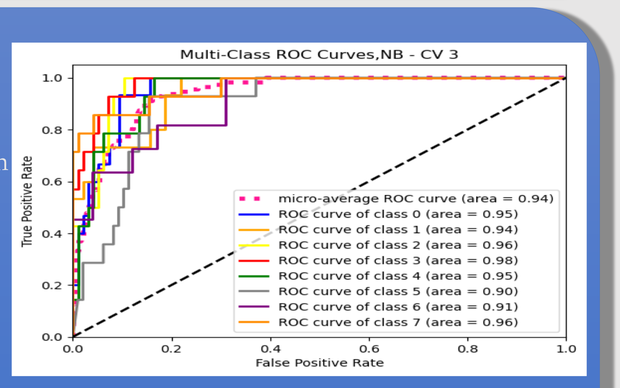

Fig. 6 – Multi-class ROC Curves Generated From One Iteration of Naive Bayes Algorithm

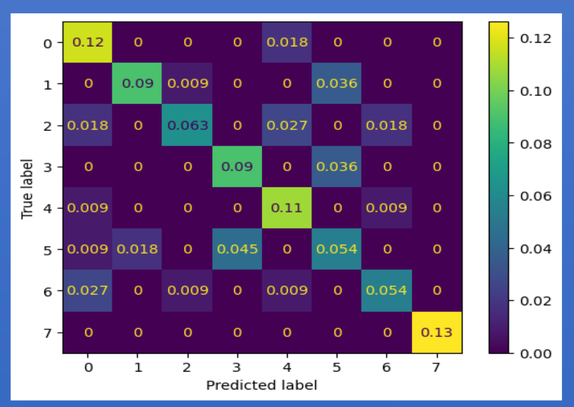

Fig. 7 – Normalized Confusion Matrix Plot Corresponding to Multi-Class ROC Curve Plot in Fig. 6

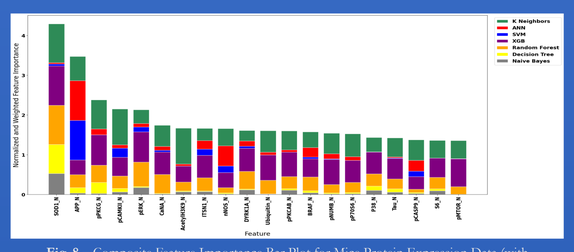

Fig. 8 – Composite Feature Importance Bar Plot for Mice Protein Expression Data (with Normalized, Weighted FI Scores)

Conclusion:

- Successful trial runs of and novel visualization/evaluation metrics from AutoMLPipe-MC on both simulated 3-class/9-class data and mice protein expression dataset has demonstrated viability of multi-class statistical analysis using automated ML pipelines
- Further research should involve using AutoMLPipe-MC to perform statistical analyses on human clinical/genetic datasets

References:

1. arXiv:2008.12829v2
2. API design for machine learning software: experiences from the scikit-learn project, Buitinck et al., 2013.
3. https://github.com/EpistasisLab/rebate-benchmark/tree/master/benchmark-data/Simulated_Benchmark_Archive
4. Ahmed MM, Dhanasekaran AR, Block A, Tong S, Costa ACS, Stasko M, et al. (2015) Protein Dynamics Associated with Failed and Rescued Learning in the Ts65Dn Mouse Model of Down Syndrome.
5. Higuera C, Gardiner KJ, Cios KJ (2015) Self-Organizing Feature Maps Identify Proteins Critical to Learning in a Mouse Model of Down Syndrome.

Lab/Research Links:

- https://github.com/UrbsLab
- https://www.med.upenn.edu/urbslab/
- www.ryanurbanowicz.com

Acknowledgements:

- This project was supported by the Penn Undergraduate Research Mentoring (PURM) Program.