Expanding FIBERS Evolutionary Feature Binning to Automatically Optimize Number of Risk Groups in Biomedical Survival Analysis

Praneel Varshney[1], Ryan J. Urbanowicz[2], Malek Kamoun[3]

[1]SEAS 2027, University of Pennsylvania, [2]Cedars Sinai Medical Center, [3]Perelman School of Medicine

Background

- Survival analyses are essential in biomedical research for understanding the time until an event of interest, such as organ failure or patient death.
- Our research aims to identify risk factors for kidney allograft failure (GF).
- The Human Leukocyte Antigen (HLA) system, encoded by genes within the Major Histocompatibility Complex (MHC) on chromosome 6, drive the human immune response[1]
  - HLA genes code for proteins that help the immune system distinguish between self and non-self[1].
- HLA matching is a process that compares the HLA types of a patient and potential donor to determine if they are a match for a transplant.
- Traditional methods of kidney matching relied on antigen-level mismatches (MMs) in HLA
  - This fails to account for the variability at the amino acid (AA) level[2,3].

FIBERS

FIBERS (Feature Inclusion Bin Evolver for Risk Stratification):

- Evolutionary binning algorithm for modeling and feature learning[2,3].
- Optimizes bins (groups) of AA-MMs based on their ability to stratify donor-recipient pairs into risk groups.
- Limited assumptions as FIBERS explores the solution space stochastically.

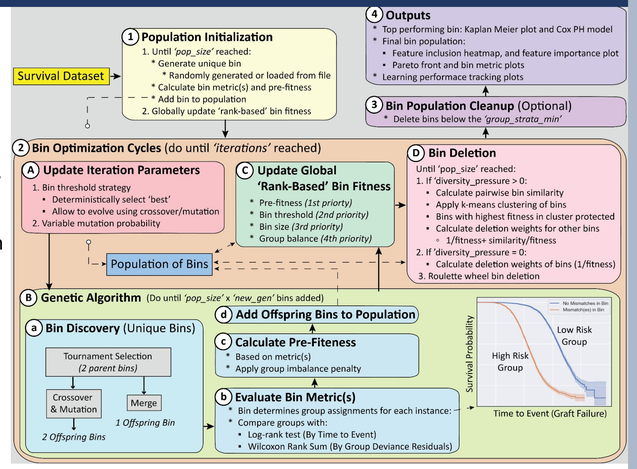

Fig. 1: schematic of FIBERS algorithm

Bin Thresholds

- Previous versions of FIBERS optimized bins of AA-MM features based on their ability to stratify donor-recipient pairs into two kidney GF risk groups (**low** and **high**) through **one** threshold value[2,3].
- Adaptive thresholding[3]: optimizes thresholds of bins when toggled on.

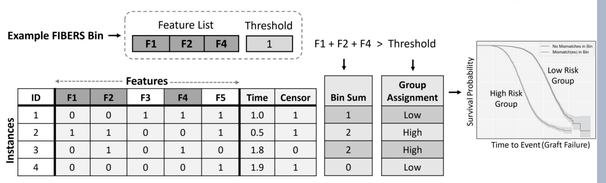

Fig. 2: FIBERS bin details

FIBERS can be run in **two** ways in regards to thresholds:

1. A threshold is inputted. All bins have this same threshold. Features will be optimized
2. No threshold is inputted. Thresholds vary. Features and thresholds optimized

Code

https://github.com/UrbsLab/scikit-FIBERS
https://github.com/UrbsLab/scikit-FIBERS/tree/varshney

Contact

pvarsh@seas.upenn.edu
Ryan.Urbanowicz@cshs.org

Methodology

- **OBJECTIVE:** Flexibly expand FIBERS to be able to discover and consider bins with multiple (3) risk groups
  - Practical use: FIBERS bins can offer more flexible recommendations for potential kidney donor/recipient pairs
  - Reduce assumptions about the ideal threshold(s) for optimal bins
  - Enable FIBERS to explore the threshold space
    - Allow FIBERS to evaluate 2-group and 3-group bins simultaneously
    - Retain FIBERS current function with only 2-group bins
  - Redefine bin fitness evaluations:
    - Fitness should be applicable to multi-group bins
    - Fitness should be comparable across 2-group and 3-group bins
- **METHODS:**
  1. Implement new fitness metrics: Multivariate Logrank test and Kruskal-Wallis
  2. Expand FIBERS to only consider bins with 3 groups
     a. Bins will include a **list** of thresholds rather than a single threshold value
  3. Update genetic operators to modify thresholds
     a. Uniform crossover, mutation, merge
  4. Reincorporate 2 group bins with 3 group bins
     a. Introduce new toggleable parameter: multi_thresholding
        i. True → 2 and 3 group bins will be explored
        ii. False → only 2 group bins will be explored (original function)
  5. Test FIBERS 3.0 functionality with 2-group and 3-group **simulated data**

Analysis & Preliminary Results

Simulated data parameters	
Instances	10000
Total features	100
Predictive features	10
Low risk proportion	0.5
Ground-truth threshold (2-group dataset)	3
Ground-truth thresholds (3-group dataset)	1, 3

FIBERS run parameters	
Iterations	100
Population size	50
Threshold evolving probability	0.5

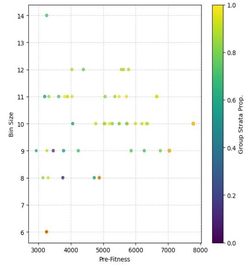

Fig. 3: 2-group population Pareto Front

	0
Features in Bin:	[P_1, P_10, P_2, P_3, P_4, P_5, P_6, P_7, P_8,...
Threshold(s)	[3]
Fitness	1.0
Pre-Fitness:	7777.433011
Log-Rank Score:	7777.433011
Log-Rank p-value:	0.0
Bin Size:	10
Group Ratio:	0.5
Count At/Below Low Threshold:	5000
Count Between Tresholds	0
Count Above High Threshold:	5000
Birth Iteration:	37

Fig. 4: 2-group top bin

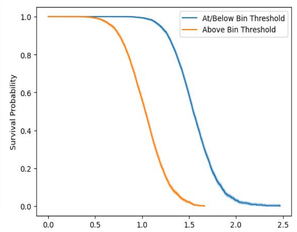

Fig. 5: 2-group Kaplan Meier survival curves

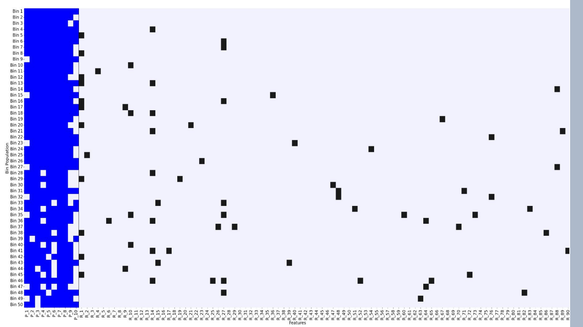

Fig. 6: 2-group bin population heatmap

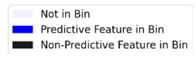

Fig. 7: key for bin population heatmaps (right)

Top bins

- Achieve the correct ground-truth thresholds that were optimal by the simulation data
- Include all predictive features and no non-predictive features

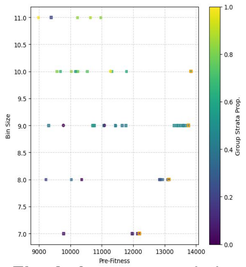

Fig. 8: 3-group population Pareto Front

	0
Features in Bin:	[P_1, P_10, P_2, P_3, P_4, P_5, P_6, P_7, P_8,...
Threshold(s)	[1, 3]
Fitness	1.0
Pre-Fitness:	13835.794617
Log-Rank Score:	13835.794617
Log-Rank p-value:	0.0
Bin Size:	10
Group Ratio:	0.2524
Count At/Below Low Threshold:	4846
Count Between Tresholds	2630
Count Above High Threshold:	2524
Birth Iteration:	68

Fig. 9: 3-group top bin

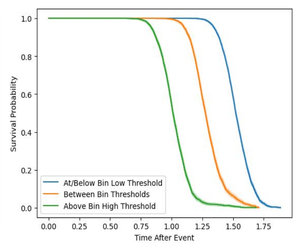

Fig. 10: 3-group Kaplan Meier survival curves

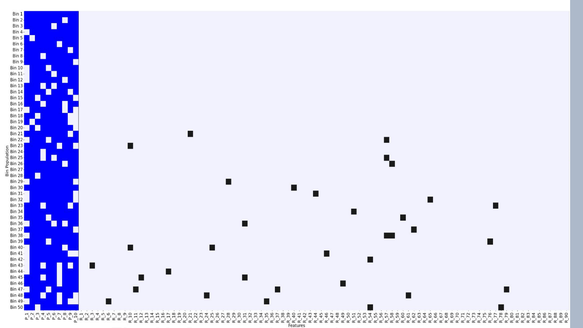

Fig. 11: 3-group bin population heatmap

Conclusions & Future Directions

CONCLUSIONS:

- FIBERS 3.0 **adds** 3-group functionality while **maintaining** 2-group functionality
- 3-group runs require greater iterations or population to maintain accuracy
- Significant runtime increases (4x) observed with multi_thresholding on
  - Occurs because FIBERS must explore a larger threshold space
- Bins with larger thresholds are discovered in later iterations
  - Larger threshold bins have more features which take longer to discover

FUTURE DIRECTIONS:

- Grid search for hyperparameter tuning
- Experimental analysis to show flexibility of the algorithm
- Full simulation study with with 30 replicates (same data/FIBERS configs but different random seeds), threshold combinations, and different dimensions
  - Determine FIBERS 3.0 strengths and weaknesses
- Apply to real-world Scientific Registry of Transplant Recipients (SRTR) data

References

1. Pedro A. Reche, Ellis L. Reinherz, Sequence Variability Analysis of Human Class I and Class II MHC Molecules: Functional and Structural Correlates of Amino Acid Polymorphisms, Journal of Molecular Biology, Volume 331, Issue 3, 2003, Pages 623-641, ISSN 0022-2836, https://doi.org/10.1016/S0022-2836(03)00750-2.
2. Dasariraju, S., Gragert, L., Wager, G.L., McCullough, K., Brown, N.K., Kamoun, M. and Urbanowicz, R.J., 2023. HLA amino acid Mismatch-Based risk stratification of kidney allograft failure using a novel Machine learning algorithm. Journal of Biomedical Informatics, 142, p.104374.
3. Urbanowicz, R., Bandhey, H., McCullough, K., Chang, A., Gragert, L., Brown, N., Kamoun, M., 2024, April. FIBERS 2.0: Evolutionary Feature Binning For Biomedical Risk Stratification in Right-Censored Survival Analyses With Covariates.