RARE: Machine Learning Approach for Binning Rare Variant Features to Detect Association with Disease

Satvik Dasariraju

Abstract:

Features with rare states, such as rare genetic variants and low frequency amino acid mismatches, pose a challenge for statistical and machine learning analyses due to limited detection power and uncertainty surrounding their role in predicting outcomes such as disease phenotype. Bins of rare variant features hold the potential to increase association detection power. However, previous binning approaches were wholly dependent on prior-knowledge assumptions, instead of data-driven techniques, and ignored multivariate interactions. Previous binning methods were also confined to genetics despite the pertinence of rare variant features across biomedicine and other scientific fields. This study develops the Relevant Association Rare-variant-bin Evolver (RARE), the first evolutionary algorithm for automatically constructing and evaluating rare variant bins with either univariate or epistatic associations, offering flexibility to initialize bins both with or without expert knowledge. RARE's ability to correctly bin simulated rare-variant associations is evaluated over a variety of algorithmic and dataset scenarios. Specifically, this study examines (1) ability to detect rare variant bins of univariate effect (with varying levels of noise), (2) using fixed vs. adaptable bins sizes, (3) employing expert knowledge to initialize bins, and (4) ability to detect rare variant bins interacting with a separate common variant. Results demonstrating the feasibility and efficacy of RARE are presented alongside an application of the algorithm in constructing bins of donor-recipient HLA amino acid positions associated with kidney graft failure in order to elucidate the role of HLA mismatches in transplantation.