An Evolutionary Approach to Interpretable Learning

Jake Robertson
Queen's University
Kingston, Ontario
jake.robertson@queensu.ca

Ting Hu
Queen's University
Kingston, Ontario
ting.hu@queensu.ca

ABSTRACT

Machine Learning (ML) interpretability is a growing field of computational research, of which the goal is to shine a light on black-box predictive models. We present an evolutionary framework to improve upon existing post-hoc interpretability metrics, by quantifying feature synergy, or the strength of feature interactions in high-dimensional prediction problems. In two problem instances from bioinformatics and climate science, we validate our results with existing domain research, to show that feature synergy is a valuable metric for post-hoc interpretability.

CCS CONCEPTS

• **Computing methodologies → Genetic algorithms**.

KEYWORDS

Machine learning, multi-objective genetic algorithms, post-hoc interpretability, feature synergy

ACM Reference Format:
Jake Robertson and Ting Hu. 2021. An Evolutionary Approach to Interpretable Learning. In *Proceedings of the Genetic and Evolutionary Computation Conference 2021 (GECCO '21).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3449726.3459460

1 INTRODUCTION

Machine learning (ML) prediction problems are increasingly characterized by high-dimensionality, or having hundreds or thousands of features and variables [5]. In such dimensions, the resulting complex ML models perform poorly in terms of interpretability, or the degree to which their underlying predictive processes can be extracted and understood [6]. However, the underlying predictive process is an important artifact in many ML applications, especially when an interpretation of the learned relationships can be used to derive scientific insight [7].

To address the demand for ML interpretability, several approaches have appeared in the literature. For ML algorithms that are not inherently interpretable, post-hoc methods aim to approximate the model's input-output relation [6]. While feature importance, a popular post-hoc metric, effectively limits the input space, feature interaction provides a detailed view of the input-output relation itself [1] [4]. Although current approaches to feature interaction improve upon feature importance, they fail to address the issue of redundancy in their interactions, and often yield unnecessarily complex results. In this study, we present a novel evolutionary framework to quantify and isolate *feature synergy*, or the strength of feature interactions in high-dimensional prediction problems.

We explore the efficacy of this framework in two problem instances from bioinformatics and climate science. Due to the absence of synthetic benchmark data and other approaches to feature synergy, we validate our results using existing domain research. Ultimately, we show that feature synergy is a valuable metric for post-hoc interpretability, especially in ML applications where scientific insight is the goal.

2 METHODOLOGY

In addressing the problem of feature synergy, the multi-objective genetic algorithm (MOGA) for feature selection provides a convenient solution. In this algorithm, the objective is to evolve a population of feature subsets with minimal size and testing error in a trained model [8]. In this selective environment, redundancy in a feature subset is counterproductive, as it corresponds with increased size. At termination, the resulting feature subsets are not only compact and accurate, but also highly synergistic.

In our implementation[1], we encode feature subsets with a binary string: a one in the i^{th} position indicates that the i^{th} feature has been selected, while a zero indicates it has not. At initialization, a random population of 200 binary strings are generated from the uniform distribution, and each corresponding feature subset's fitness is evaluated. To evaluate testing error, we employ k Fold Cross-Validation to define independent training and testing sets, and either K Nearest Neighbors (KNN) or Support Vector Machine (SVM) to train models. Based on the results of evaluation, we employ the fast-non-dominated sort algorithm [3] to select parents. After 100 parents have been selected, one-point mutation and crossover are employed to create 100 offspring and the combined population of parents and offspring automatically proceed into the next generation. The MOGA is run 100 times for 1000 generations and the combined 20,000 feature subsets are collected and analyzed for feature importance and synergy.

Given a feature f and the collection of feature subsets A where f is selected, *importance* is measured by the degree to which f occurs in compact and accurate subsets.

$$importance(f) = \sum_{a \in A} \frac{1 - error(a)}{|a|} \qquad (1)$$

Because Equation 1 is dependent on the number of solutions in the collection, we recommend normalizing the resulting feature importance distribution between 0 and 1.

GECCO '21, July 10–14, 2021, Lille, France
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3459460

[1] https://github.com/jr2021/GA_feature_synergy.git