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.

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

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.

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

GECCO '21, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3459460

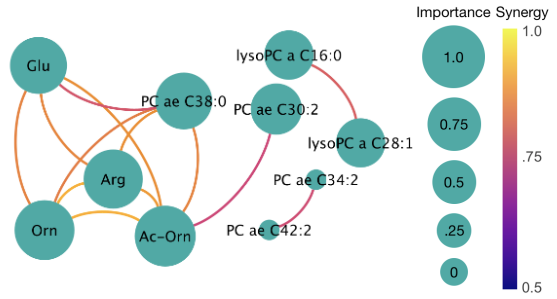

Figure 1: Metabolic synergy (with accuracy-factor) learned by the MOGA configured with KNN to create models.

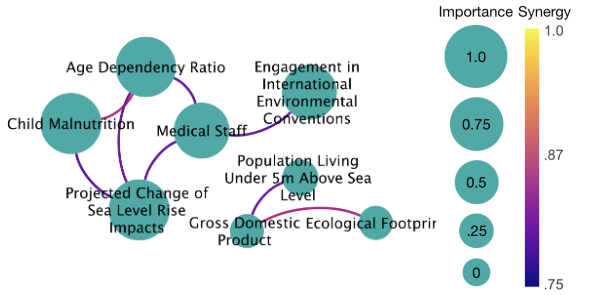

Figure 2: Sociopolitical factor synergy (with accuracy-factor) learned by the MOGA configured with SVM to create models.

Given a pair of features (f_i, f_j), a collection of feature subsets A where f_i is selected, and a collection of subsets B where f_j is selected, *synergy* is measured by the *Jaccard similarity coefficient*, or the degree to which the pair co-occurs.

$$synergy(f_i, f_j) = J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Due to the random nature of the MOGA, it is possible that a feature pair (f_i, f_j) emerges as synergistic due to a late generation co-occurrence. In order to isolate direct feature synergies, we recommend scaling Equation 2 by the testing accuracy of the feature subset $\{f_i, f_j\}$ in a trained model. However, this factor is counterproductive when searching for higher-order synergistic interactions.

3 RESULTS AND DISCUSSION

We first apply our framework to identify metabolic synergy in knee-osteoarthritis (OA) patients, using the NFOAS data set from the Memorial University of Newfoundland. The resulting synergies (Fig. 1) between the most important metabolites are either inputs or outputs to the biological synthesis of arginine (Arg). This amino acid has been previously identified as a bio-marker for OA, due to its inhibition of cathepsin, an enzyme that degrades cartilage [9]. Further analysis of this interaction reveals that these features provide a clear decision boundary between classes: a high concentration of Glutamine and Ornathine, paired with a low concentration of Arg and N-Acetylornithine corresponds to 97% of the case group.

Next, we apply our framework to identify sociopolitical factors in climate change vulnerability prediction, using the ND-GAIN Index from the University of Notre Dame. In the results (Fig. 2), an interaction appears between several health-related factors. This interaction indicates that the ratio between medical infrastructure and population health is indicative of a country's overall vulnerability to climate-change-related sea-level-rise, warming, and natural disaster. Further analysis reveals that several Central and Northern African, as well as South and Southeast Asian countries are especially vulnerable in terms of these factors. In these countries, low international engagement indicates an incapacity to enact a policy-driven response [2].

In the above problem instances, the resulting feature synergies align either with years of extensive bioinformatics research, or simple and intuitive sociopolitical mechanisms. In the application of interpretable ML to derive scientific insight, this outcome not only confirms that models can learn and leverage scientifically valid and intuitive relationships, but suggests that this combination of post-hoc metrics provides the necessary information to extract and understand them. In the context of interpretable ML as a whole, this suggests that feature synergy is a valuable post-hoc metric in general, and should be tested in other application areas, such as identifying and removing machine bias with respect to legally protected attributes (race, gender, etc.). Feature synergy could also be applied as a general pre-processing step in order to define more compact and interpretable models. However, the degree to which feature synergy is general, or not "over-fit" to the training data, is presently unknown.

ACKNOWLEDGEMENTS

We would like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Queen's University School of Computing for funding and supporting this research.

REFERENCES

[1] Sumanta Basu, Karl Kumbier, James B. Brown, and Bin Yu. 2018. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences* 115, 8 (2018), 1943–1948. https://doi.org/10.1073/pnas.1711236115

[2] Chen Chen, Ian Noble, Jessica Hellmann, Joyce Coffee, Murillo M, and Nitesh Chawla. [n.d.]. University of Notre Dame Global Adaptation Index Country Index Technical Report. ([n. d.]).

[3] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (2002), 182–197. https://doi.org/10.1109/4235.996017

[4] Ting Hu, Karoliina Oksanen, Weidong Zhang, Ed Randall, Andrew Furey, Guang Sun, and Guangju Zhai. 2018. An evolutionary learning and network approach to identifying key metabolites for osteoarthritis. *PLoS Computational Biology* 14, 3 (2018). https://doi.org/10.1371/journal.pcbi.1005986

[5] Mario Köppen. 2000. The Curse of Dimensionality. *5th Online World Conference on Soft Computing in Industrial Applications* 1 (2000).

[6] James W. Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080. https://doi.org/10.1073/pnas.1900654116

[7] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. 2020. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access* 8, 2 (2020), 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199

[8] Bing Xue, Mengjie Zhang, Will N. Browne, and Xin Yao. 2016. A Survey on Evolutionary Computation Approaches to Feature Selection. *IEEE Transactions on Evolutionary Computation* 20, 4 (2016), 606–626. https://doi.org/10.1109/TEVC.2015.2504420

[9] Weidong Zhang, Guang Sun, Sergei Likhodii, Ming Liu, Erfan Aref-Eshghi, Patricia Harper, Glynn Martin, Andrew Furey, Roger Green, Ed Randell, Proton Rahman, and Guangju Zhai. 2016. Metabolomic analysis of human plasma reveals that arginine is depleted in knee osteoarthritis patients. *Osteoarthritis and Cartilage* 24, 5 (2016), 827–834. https://doi.org/10.1016/j.joca.2015.12.004