Sparse Feature Interactions for Interpretable Healthcare Decision-Making

James Enouen and Yan Liu
University of Southern California – Computer Science Department

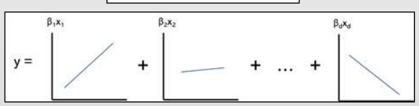
Abstract

In the modern era, machine learning and artificial intelligence are applied to new aspects of our lives with each passing day. As these fields are increasingly embedded within consequential fields like self-driving cars, recidivism prediction, credit loaning, and **healthcare advising**, it has become increasingly clear that criteria beyond ***accuracy*** are required for these multifaceted applications. It has become ever-pressing to address the needs of ***interpretability***, ***robustness***, ***safety***, and **fairness** for these decision-critical applications.

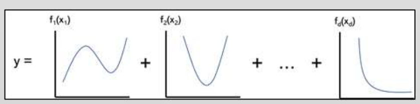
This work leverages interpretability as the cornerstone to these auxiliary requirements by blending the powerful inductive biases of deep learning with the robustness and interpretability of classic statistical methods. We demonstrate the empirical performance of our "SIAN" model on two datasets: **MIMIC-III** for 48-hour mortality prediction and causal inference for **blood donation**. We find our model can consistently ***outperform*** neural networks, kernel machines, and boosting methods while remaining much more ***interpretable***. We explore the interpretations found by our model and discuss their relevancy towards applying ML algorithms in healthcare applications, reiterating the need for ***causality-based*** methods in the ***safety-critical*** domain of healthcare.

Introduction

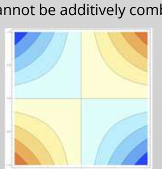
Linear Model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$



Additive Model: $y = \beta_0 + f_1(x_1) + \cdots + f_d(x_d)$



The limitation of these models is they cannot model "feature interactions" where two features are simultaneously important and cannot be additively combined e.g. the XOR function $f(x_1,x_2) = x_1 \cdot x_2$.



$\mathcal{I} = \{i_1, \ldots, i_{|\mathcal{I}|}\} \subseteq [d] := \{1, \ldots, d\}$

$f : \mathbb{R}^d \to \mathbb{R}$ $\quad y = f(x)$

$$\omega(\mathcal{I}) := \mathbb{E}_x\left[\frac{\partial^{|\mathcal{I}|} f(x)}{\partial x_{i_1} \partial x_{i_2} \ldots \partial x_{i_{|\mathcal{I}|}}}\right]^2 > 0.$$

XOR function — interaction strength definition

We use the expectation of the mixed derivative of the tuple to measure the "interaction strength" of a given tuple/ feature subset. The following model is able to handle these more general cases and can model an arbitrary function.

Generalized Additive Model with Interactions:

$$g(y) = f_\emptyset + \sum_i f_i(x_i) + \sum_{i<j} f_{i,j}(x_{\{i,j\}}) + \sum_t f_{\mathcal{I}_t}(x_{\mathcal{I}_t})$$

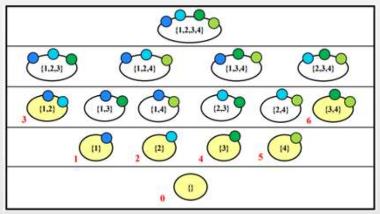
In theory, we could include all possible interaction terms just like in a typical deep neural network, however, we selectively choose a very small portion of these 2^d possible interactions where d is the number of features (in our case d=30 or d=34).

Methodology

Our algorithm focuses on only specifying a very small (sparse) subset of all the possible 2^{34} or 2^{30} feature interactions which could exist in the dataset. We argue that the very high-dimensional feature tuples are not only more difficult to model but also have shrinking effect on the true outcome as their size increases.

Hence, we train our model to focus only on a sparse subset of all feature interactions using an additive model. Because we use neural networks to fit the additive model, we refer to our method as a Sparse Interaction Additive Network (SIAN).

Below, we illustrate an example of selecting 7 of the 16=2^4 possible interactions for a small 4-dimensional input.



Our algorithm is based on the fact that: for an interaction to exist, all of its subsets need to exist as well. For example, if {1,2,3} exists, then it is required that {1,2}, {1,3}, and {2,3} exist. We use this fact to start at one-dimensional features and slowly build to higher-dimensional interactions. This has the added bonus that if all of our interactions are less than or equal to two dimensions, we can fully plot the "shape functions" of our prediction algorithm.

Algorithm 1: Feature Interaction Selection

Input: Trained prediction model f(x), and validation dataset $X^* = \{x_1, \ldots, x_v\}$
Parameter: Cutoff index K, cutoff threshold τ
Output: $\mathcal{I}$, a set of subsets of indices corresponding to approximately all feature interactions with index below K and strength below τ.

```
Set I ← {{i} : i ∈ [n]} ∪ ∅                 // The true detected interactions so far
Set J ← {{i,j} : i,j ∈ [n]; i < j}          // The next set of interactions to check
k ← 2
while k ≤ K do
    place holder
    for J in J do
        ω(J) ← 0
        for x ∈ X* do
            Compute ω_J(x) with Archipelago
            ω(J) ← ω(J) + ω_J(x)
        end for
        ω(J) ← ω(J)/|X*|
        if ω(J) > τ then
            I ← I ∪ J
        end if
    end for
    k ← k + 1
    J ← {J : J ∈ P([n]); |J| = k; (1/2^k) Σ_{I⊆J} [1_{I∈I}] > θ}
end while
return I
```

We use the "Archipelago" method to approximate the "interaction strength" as defined to the left in the Introduction section as the expectation of the square of the mixed derivative [1]. This method approximates the mixed derivative using a secant approximation or the finite differences method. We average the result over a validation set to estimate the strength across the entire dataset distribution.

Datasets

MIMIC-III:
We collect 30 covariates from the electronic health records of 32,000 ICU patients over a 48-hour period. We include features which are commonly believed to important in mortality prediction, including those required to predict the interpretable SAPS-II and SOFA scores. There is a positivity rate of around 9.2% in the dataset.

Blood Donation:
We collect 34 covariates from 60,000 potential donors. This data was collected as randomized experiment where a portion of the potential donors received a text message offering them a grocery coupon for donating blood. There was an observed increase in donations from the incentivized population. The total positivity rate was around 1.0%.

Results

Below we can see the results of multiple different models for both the MIMIC and blood donation datasets. AUROC and AUPRC are the area under receiver-operating characteristic curve and precision-recall curve, respectively. We use these metrics to analyze classification prediction for very imbalanced class frequencies. For the blood donation dataset, we run a mock experiment on held-out data to measure the percentage of donors and the economic benefit (see [3] for a detailed explanation) of a given decision algorithm.

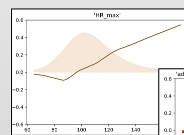
MIMIC-III Mortality

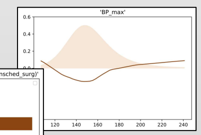
Model	AUROC	AUPRC
SAPS II	0.792	0.281
SOFA	0.703	0.225
svm	0.778	0.359
rf	0.789	0.363
xgb	0.803	**0.369**
ebm (ga^2m)	**0.816**	0.367
dnn	**0.816**	0.365
sian-1	**0.826**	0.369
sian-2	0.821	**0.384**
sian-3	0.821	0.371
sian-5	0.809	0.367

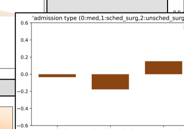
Blood Donation Likelihood

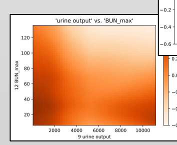
Model	donation percentage	economic benefit	AUROC T0	AUROC T1
none (0%)	0.69%	40.6±5.2 ¢	–	–
random (50%)	0.92%	43.3±7.1 ¢	–	–
all (100%)	1.15%	46.6±2.2 ¢	–	–
svm	1.03%	42.5±4.6 ¢	.541	.672
rf	1.12%	45.5±3.6 ¢	.591	.696
xgb	1.03%	45.8±5.8 ¢	.587	.716
ebm (ga^2m)	1.06%	48.2±5.1 ¢	.629	.738
dnn	1.04%	46.9±6.3 ¢	.597	.715
sian-1	1.09%	48.5±9.0 ¢	.603	.661
sian-2	1.09%	48.8±6.0 ¢	.601	.676

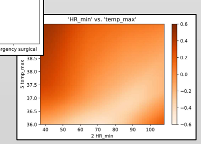
We find that our model performs consistently against support vector machines, deep neural networks, random forests, XGB boosting, and EMB [2]. Interestingly, we find that the one-dimensional and two-dimensional versions of SIAN perform the best on these two datasets rather than the higher-dimensional versions. Ongoing work hypothesizes this is a consequence of the log-ratio between the number of samples (32k, 60k) versus the dimensionality (30, 34).











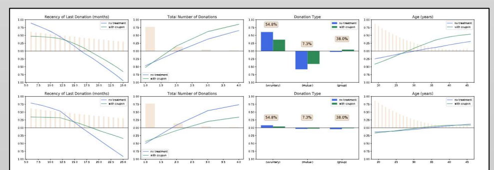


Fig. 3. Blood Donation 1D Likelihood Trends: The top row corresponds to the 1D likelihood trends of the DNN; the bottom row corresponds to the 1D trends learned by the GAM1. The blue lines correspond to the trend for participants with no coupon and the green lines correspond to participants who were offered a coupon. The orange shading in the background illustrates the approximate distribution of the given variable. These top four 1D trends capture 45.8%, 14.8%, 9.7%, and 7.0% respectively of the total variance of the DNN. We can see how they differ from the trends learned directly by the GAM1 on the auditing set.

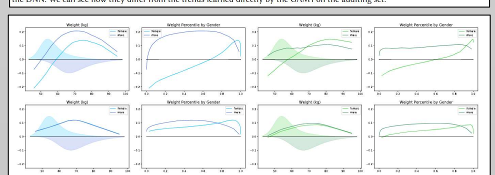
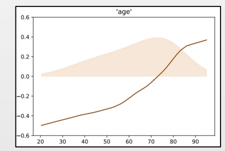


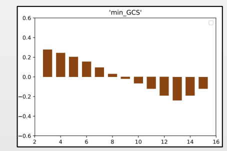
Fig. 4. Blood Donation 2D Likelihood Trends: Here we can see the interaction between gender and weight in each model. The top row corresponds to the DNN trends; the bottom row corresponds to the GAM2 trends. The left four blue figures correspond to no treatment and the right four green figures correspond to an offered coupon. The background shading is again used to illustrate the distribution of weights for each gender (female above; male below.)

Discussion

Shape Functions
We can see our model has learned many simple trends from the dataset, including that abnormally high blood pressures, abnormally high heart rates, and unscheduled emergency surgeries are indicative of higher mortality risk. Moreover, our interaction detection algorithm finds an association between urine output and blood urea nitrogen as well as an association between minimum heart rate and maximum recorded temperature.
Our causal inference dataset sees similarly interpretable trends including that more recent donors, high frequency donors, and older donors are generally more likely to donate blood.

Interpretability and Causality
Especially in the realm of healthcare, it is important to note that these models are trained to pick up on statistical correlations rather than the true causality of the process.





In the above "Glasgow Coma Scale" plot which indicates mental alertness, we can see that the risk of death gradually decreases as we score higher on the alertness scale. However, as we approach the maximum score of 15, our risk of death actually increases for scores of 14 and 15. This is a direct consequence of our correlation-based analysis and is likely a consequence of hospital staff giving extra attention to patients with below perfect GCS scores.
Similar trends have previously been observed such as lower death rates in hospital patients with ages beyond 100 and for pneumonia patients afflicted with asthma [2].

Our second dataset alleviates these correlation issues by working directly on randomized experiment data. Clinical trials are the golden standard of experimental design; however, they are extremely costly and can face ethical and legal constraints in the domain of healthcare. In this setting, we need to explicitly balance confounding variables with potentially unknown treatment distributions. We often must randomize amongst two feasible treatment paths without ensured compliance from patients. Towards these challenges, we suggest further studies on low-risk econometric domains to understand algorithmic challenges in this domain before application to the safety-critical domain of healthcare, like our low stakes experiments on blood donation likelihood.

Conclusion

We find that our sparse model performs on par with other popular and state-of-the-art machine learning methods including XGB, kernel machines, random forests, multilayer perceptrons, and EMB. Despite this level of performance, SIAN does not sacrifice interpretability using one- and two-dimensional shape functions.

We use our interpretable insights to discover discrepancies in our causal understanding of the data. We reiterate the importance of distinguishing causation from correlation in the healthcare domain and we call for further research on the robustness of causal inference in domains where purely randomized experiments are impossible from a practical standpoint.

Acknowledgements

We thank Tianshu Sun for the usage of the blood donation dataset [3].

[1] Archipelago (Tsang 2020):
https://proceedings.neurips.cc/paper/2020/hash/443dec3062d0286986e21dc0631734c9-Abstract.html
[2] Intelligable Models for Healthcare (Caruana 2015):
https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/KDD2015FinalDraftIntelligibleModels4HealthCare_igt143e-caruanaA.pdf
[3] Heterogenous Treatment Effects… (McFowland 2021):
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3718956