Towards an Interpretable Data-driven Trigger System for High-throughput Physics Facilities

Chinmaya Mahesh† Kristin Dona⋆ David W. Miller⋆ Yuxin Chen⋆

†University of Illinois Urbana-Champaign ⋆University of Chicago

Motivation: LHC Trigger System

- Data filtering algorithms (trigger algorithms) targeted at discovery sciences must operate at the level of 1 part in 10^5 due to resource constraints.
- Design relies heavily on prior knowledge of the feature space being probed.
- *redundant* labeling schemes and *cost-ineffective* algorithm execution.

Data Driven, Explainable Triggers

- Refine the trigger and data filtering algorithms at future physics facilities.
- Each trigger algorithm incurs a latency at runtime. Thus, finding the *most efficient* set of trigger algorithms *at runtime* is crucial for a real-time trigger system.

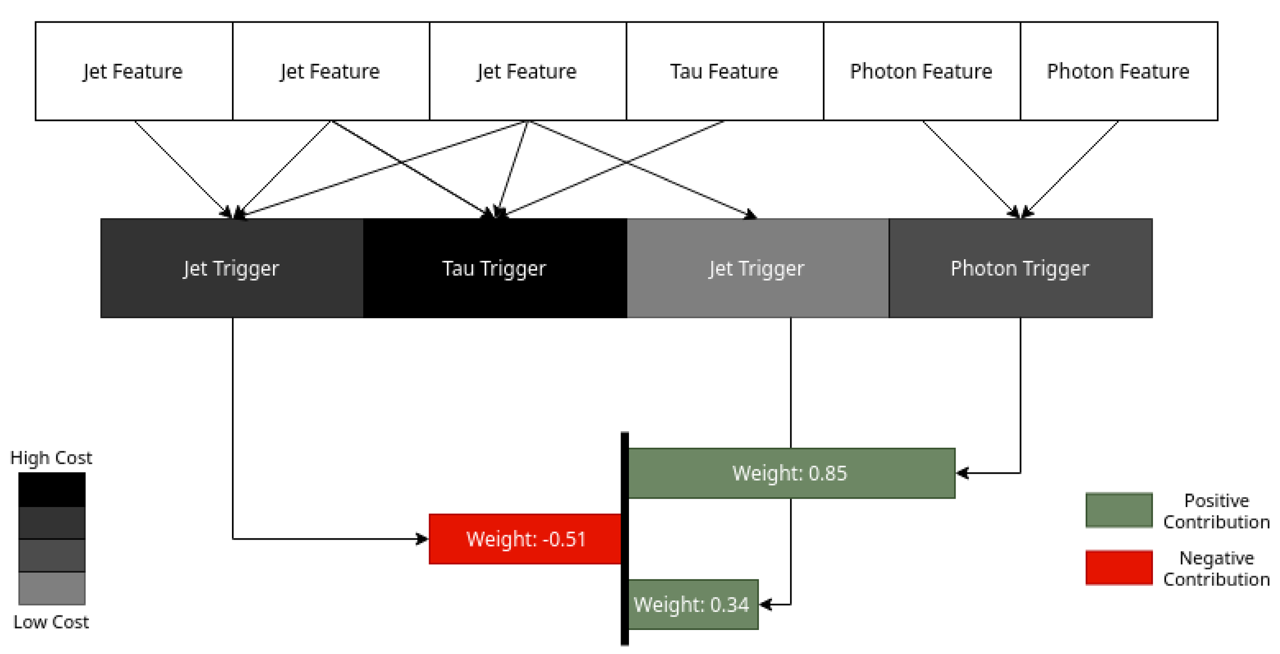


Figure 1: An example cost-effective explanation of an event.

Example of Non-interpretable LHC Trigger Recommendation

> Only applying the *b-jet trigger* to an event such as $H \to bb$, rather than also applying a *threshold di-jet trigger.*

With an interpretible algorithm we hope to gain information that this decision was made because the most important physics feature for this event is the *b-jet tagging value.*

Local Interpretable Explanations (LIME)

- Uses local interpretable surrogate models to explain individual predictions of black box models.
- Does *not* take into account *cost* of each feature.

Problem Statement

> Our work extends LIME and can be viewed as a sparsity-based locally interpretable model, where we seek a *minimal-cost explanation* for the LHC trigger outputs.

- Given a dataset $X \in \mathbb{R}^{n \times p}$ (*n collision events*; each event is described by p numerical features), a set of labels T (known as *triggers*), and an outcome matrix $y = \{0, 1\}^{n \times |T|}$ (i.e. triggers each event satisfies).
- *cost function* $c(f_i)$: the cost of using feature f_i to predict the outcome of an event.
- **Goal:** Identify the *most cost-efficient subset of features* that enables us to *maximize coverage* of X in the trained model while using selected features to make predictions.

Our Approach: Cost-effective (CE) LIME

LIME with Elastic Net

- **Recap**: LIME trains a sparse model with a *dataset of perturbations of x*. The trained weight vector of this model describes the importance of each feature.
- We adopt *elastic net* as a general formulation (with the LASSO and ridge regressions being special cases), which trades off model interpretability (sparsity) and accuracy:

$$\hat{\beta} = \arg\min_{\beta} \left(|y - X\beta|^2 + (1-\alpha)\lambda \sum_{i=1}^{p} |\beta_i| + \alpha\lambda \sum_{i=1}^{p} |\beta_i|^2 \right).$$

Cost Effective Elastic Net

To obtain a $\hat{\beta}$ which is both sparse and cost efficient, we propose adding a coefficient of $c(f_i)$, which is the cost of feature i, to each respective term $|\beta_i|$ and $|\beta_i|^2$ in the elastic net penalty:

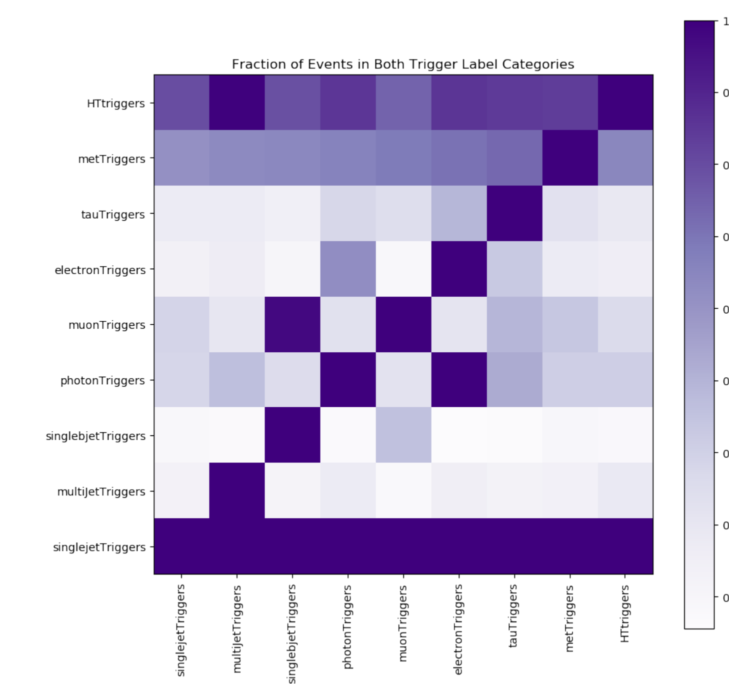
$$\boxed{\hat{\beta} = \arg\min_{\beta} \left(|y - X\beta|^2 + (1-\alpha)\lambda \sum_{i=1}^{p} |\beta_i| \cdot c(f_i) + \alpha\lambda \sum_{i=1}^{p} |\beta_i|^2 \cdot c(f_i) \right)}$$

Submodular Pick

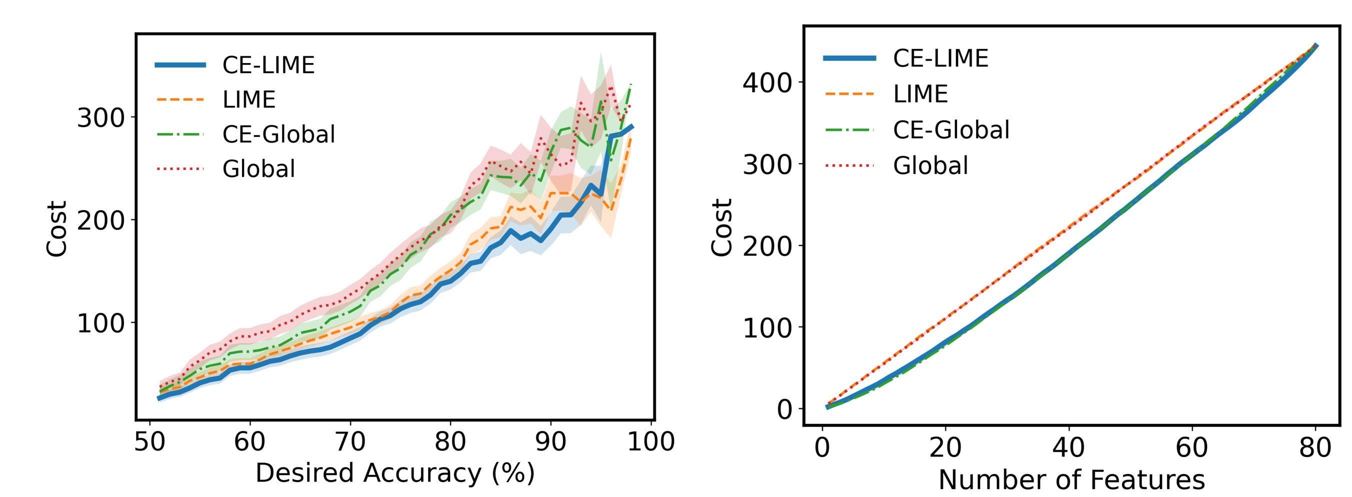
- A model-wide, *global explanation* similar to the event specific explanation is desired.
- LIME with Submodular Pick (SP-LIME) creates an importance vector I, which gives us a total ordering of all features F that enables us to select an optimal subset of F.
- We call this method of using a modified SP-LIME with a cost-effective elastic net *CE-LIME*.

Experimental Setup

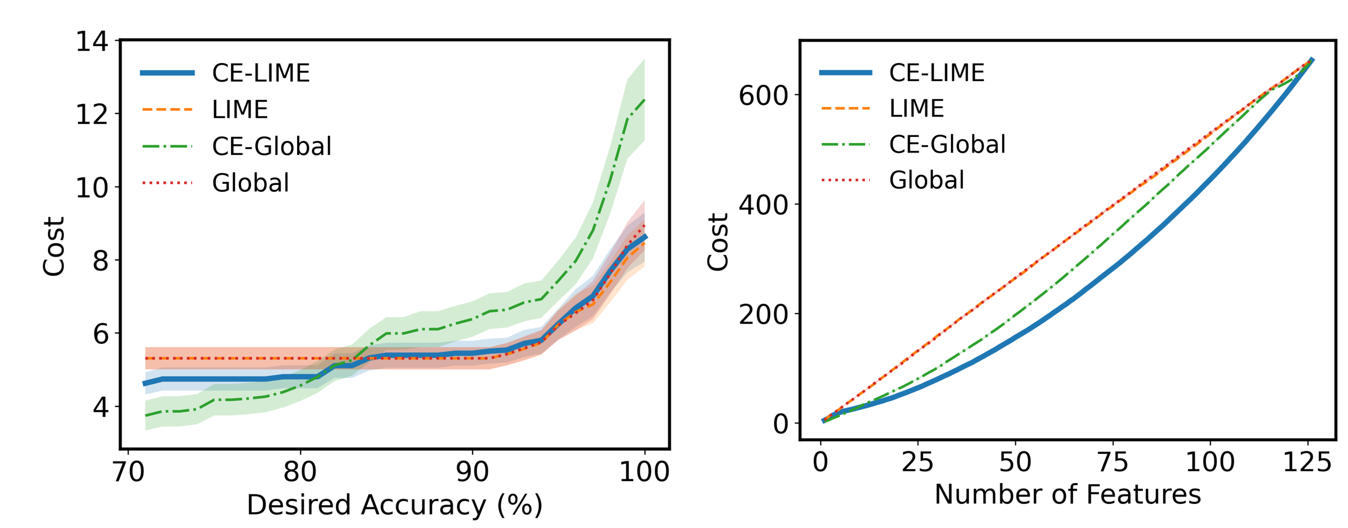
- Toy dataset
  - randomly generated by `make_classification` of scikit-learn.
  - cost function c created from a uniform distribution in the interval $[0, 10]$.
- CMS Open Data
  - publicly available; cf. CERN Open Data Portal, 2017.
  - 9 different triggers with randomized cost of features in every trial, with the costs being uniformly distributed in $[0, 10]$.
  - the figure shows the fractional overlap between features which share trigger labels and trigger label categories. The large fractional overlap emphasises the potential for these algorithms to be optimized.



Experimental Results



Dataset (100 trials)



Open Data (10 trials)

* *This work was supported by the Center for Data and Computing at the University of Chicago via a Discovery Grant.*