The Federation

A novel machine learning technique applied on data from the Higgs Boson Machine Learning Challenge

Maximilian Mucha, Eckhard von Törne

October 25, 2022

ACAT 2022, Bari

Data analysis in High Energy Physics

- Large datasets are typical in HEP
  - $\Rightarrow$ Because of resource constraints, often only a subset of data is used
- Background dominated data $\Rightarrow$ Imbalanced data
- Complex data
  - $\Rightarrow$ Undefined values
  - $\Rightarrow$ Categorical values

Problem: Training a model on a large dataset can take a lot of computing time and resources. How can this be mitigated?

What is the Federation?

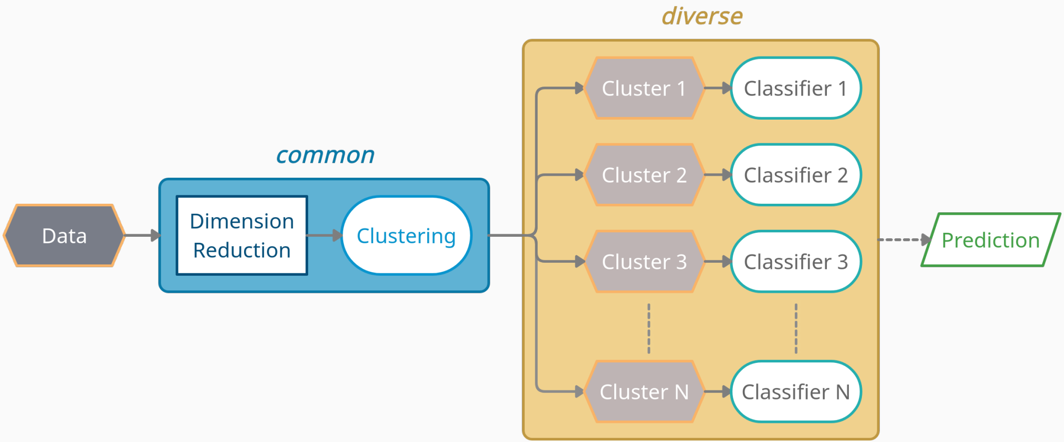
Idea:
1. Split data into smaller subsets and for each subset train a model.
2. Predict by using the ensemble of models

Issue: But how to split the data wisely and how to predict? $\Rightarrow$ Federation

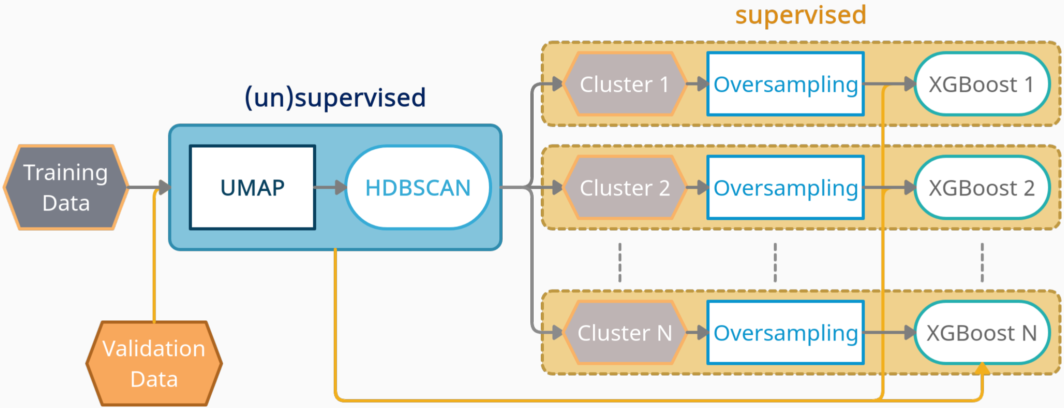
Definition

*"**federation**, the government of a federal community. In such a model there are two levels of government, one dealing with the* ***common*** *and the other with the territorially* ***diverse****."* `https://www.britannica.com/topic/federation`

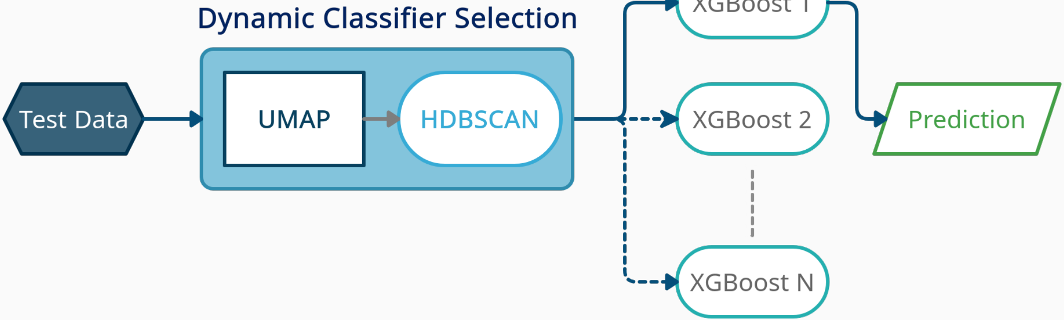
Federation – Concept



Federation – Training



Federation – Predicting



Data Sample

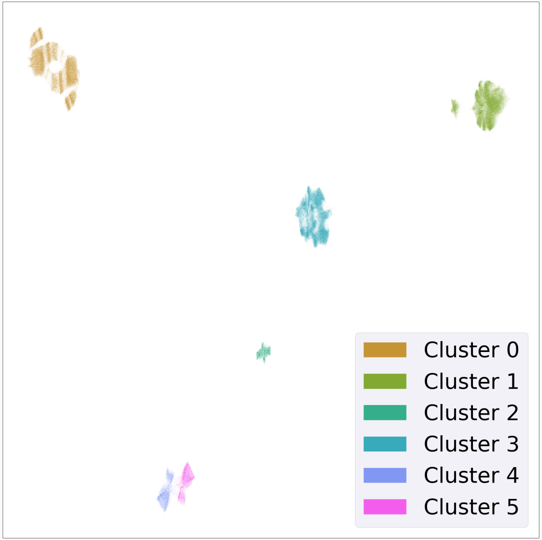
- Openly available dataset[1] from ATLAS
- Simulated $H \to \tau\tau$ signal and background events at $\sqrt{s} = 13\,\text{TeV}$
- Developed for the Kaggle Higgs Boson Challenge[2]
- Total of 30 features (some have *undefined* values)
  - → 17 kinematic features (including categorical: *PRI_jet_num*)
  - → 13 derived features
- Imbalance Ratio of IR ≈ 1.92 (IR $= \frac{N_{\text{sig}}}{N_{\text{bkg}}}$)
- 4 Subsets: training (250 000), validation (100 000), testing (450 000), unused (18 238)

[1] https://opendata.cern.ch/record/328

[2] https://www.kaggle.com/c/higgs-boson

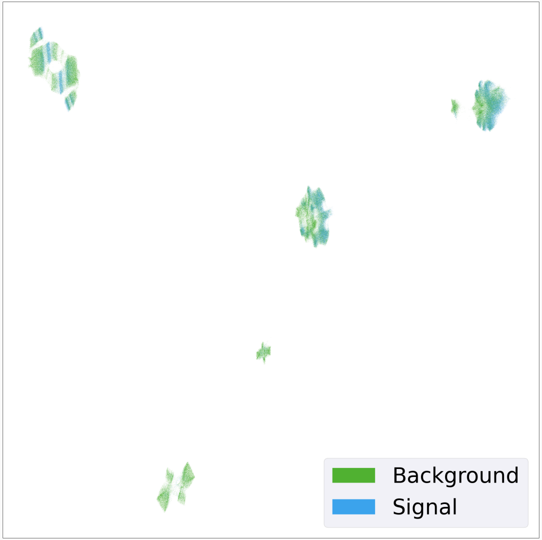
Federation – Visualization

- UMAP [1] reduces dimensions of training data from 30D to 2D
- HDBSCAN [2] finds 6 cluster in the 2D UMAP embedding
  - ⇒ 6 independent classifiers (federation members) are constructed



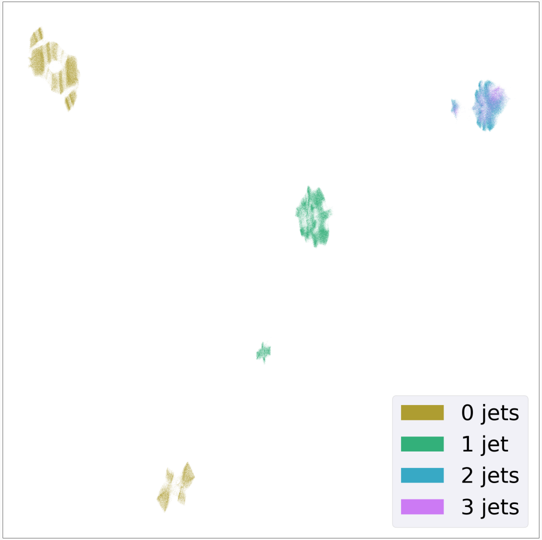
Federation – Visualization

- In some clusters, the majority of data points are background events
  - $\Rightarrow$ Oversampling is needed
- Cluster 0 has "signal bands" in local structure



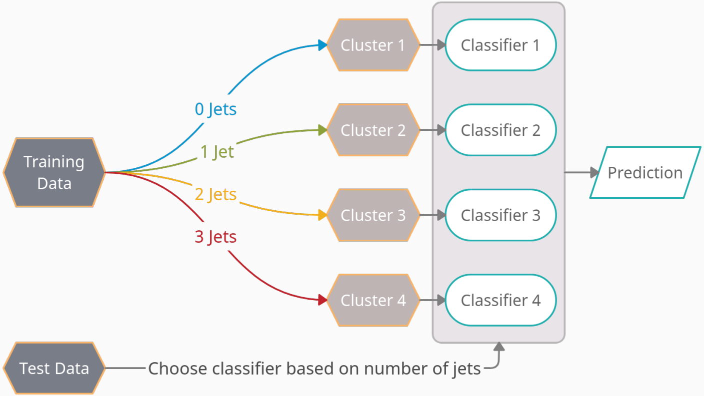
Federation – Visualization

- Global topology of the 2D-embedding is highly influenced by the number of jets feature



Baseline – Hand made clustering

- Clustering causes loss of statistics
  - ⇒ Performance of cluster based classifier degrades
- For a fair comparison, we chose as baseline a similar (feature driven) clustering
  - ⇒ Clusters based on number of jets



Performance evaluation

Figure of merit

- The evaluation metric from the Kaggle Higgs Bososn Challenge is used
  $\Rightarrow$ Approximate Median Significance (AMS)
- Predictions are sorted after the highest probability
- Only the N top predictions are marked as signal predictions

Finding the right threshold

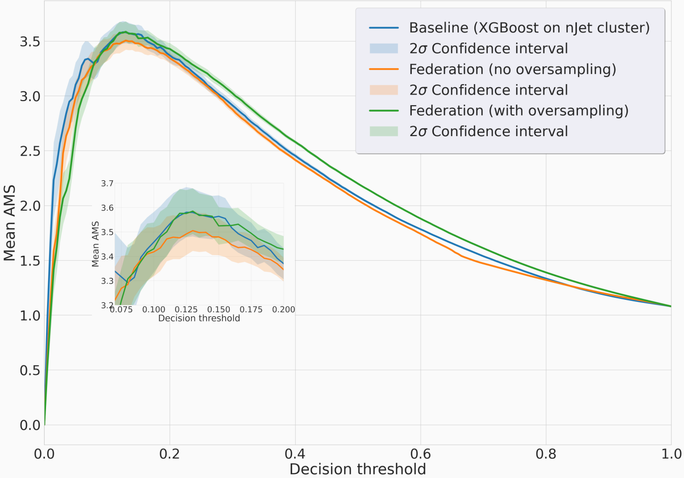
- Threshold scan on validation data
- Threshold with the highest AMS is used for the test data

Performance comparison

Method	Mean AMS $\pm$ Std	@Threshold
Single classifier	3.628 ± 0.036	0.160
Baseline (n-Jet clusters)	3.395 ± 0.067	0.090
Federation (no oversampling)	3.480 ± 0.037	0.145
Federation (with oversampling)	3.564 ± 0.041	0.145
Kaggle Challenge Submissions	**AMS**	
Winner (Gábor Melis)	3.80581	
Place 6 (Crowwork with XGBoost)	3.71885	

Bootstraped results ($N = 1000$) of test data

Federation – Performance plot



Mean bootstrapped ($N = 1000$) AMS of test data against decision threshold

Summary and Conclusion

- UMAP and HDBSCAN are the core of the Federation
  - ⇒ Creation of Federation members
  - ⇒ Used for Dynamic Classification Selection
- Oversampling the training data of the Federation members improves performance
- The training and predicting of the Federation members can be parallelized
- The Federation surpasses a comparable n-Jet based clustering approach

Thank you for listening!

References

[1] Tim Sainburg, Leland McInnes, and Timothy Q. Gentner. "Parametric UMAP: learning embeddings with deep neural networks for representation and semi-supervised learning". In: *ArXiv e-prints* (2020). arXiv: 2009.12981 [stat.ML].

[2] Leland McInnes, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering". In: *The Journal of Open Source Software* 2.11 (2017), p. 205.

[3] Leland McInnes et al. "UMAP: Uniform Manifold Approximation and Projection". In: *The Journal of Open Source Software* 3.29 (2018), p. 861.

[4] Haibo He and Edwardo A. Garcia. "Learning from Imbalanced Data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.

[5] György Kovács. "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets". In: *Applied Soft Computing* 83 (2019). (IF-2019=4.873), p. 105662. DOI: 10.1016/j.asoc.2019.105662.

[6] György Kovács. "smote-variants: a Python Implementation of 85 Minority Oversampling Techniques". In: *Neurocomputing* 366 (2019). (IF-2019=4.07), pp. 352–354. DOI: 10.1016/j.neucom.2019.06.100.

Data Sample – Classifier

- Using XGBoost as baseline to compare with previous research
- Parameters based on XGBoost Paper[3]
    - max_depth $= 6$
    - learning rate $= 0.1$
    - loss $=$ AUC of Precision-Recall Curve
    - $\gamma = 0.1$, $\lambda_{reg} = 0$
    - 30 early stopping rounds
- Using validation set for validation

[3] PMLR 42:69-80, 2015

Data Sample – Figure of merit

Approximate median significance (AMS)

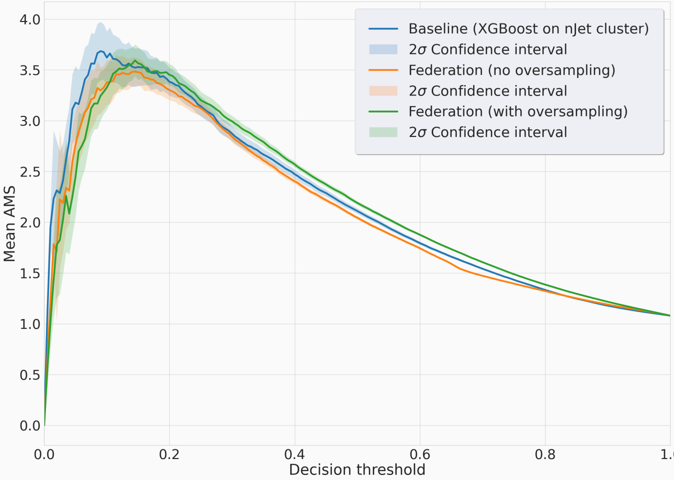
$$\text{AMS} = \sqrt{2\left((s + b + b_r)\log\left(1 + \frac{s}{b + b_r}\right) - s\right)}$$

$b_r = 10$ is a constant regularization term

$$s = \sum_{i=1}^{n} w_i \mathbb{1}\{y_i = s\}\mathbb{1}\{\hat{y}_i = s\}$$

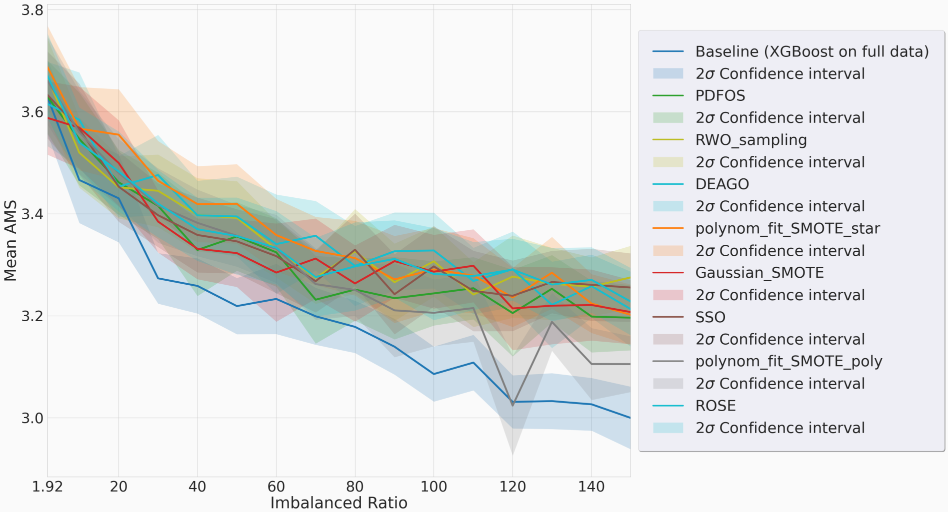
$$b = \sum_{i=1}^{n} w_i \mathbb{1}\{y_i = b\}\mathbb{1}\{\hat{y}_i = s\}$$

Federation – Performance plot



Mean bootstrapped ($N = 1000$) AMS of validation data against decision threshold

Oversampler performance comparison



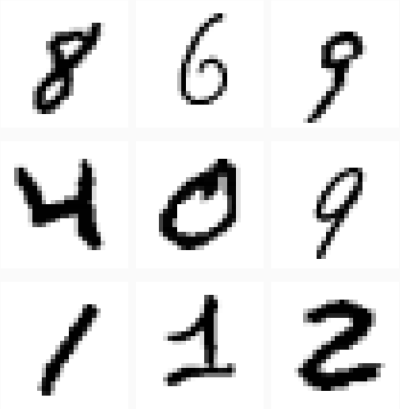
Mean bootstrapped ($N = 1000$) AMS against IR of training data for best performing oversamplers

Federation – Oversampler performance comparison

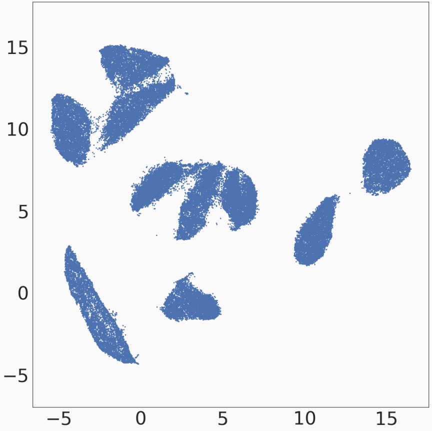
Method	Mean AMS $\pm$ Std	@Threshold
Federation (ROSE)	3.564 ± 0.041	0.145
Federation (PDFOS)	3.554 ± 0.040	0.145
Federation (polynom fit)	3.530 ± 0.034	0.160
Federation (RWO sampling)	3.529 ± 0.036	0.145
Federation (no oversampling)	3.480 ± 0.037	0.145
Federation (SMOTE)	3.451 ± 0.038	0.145

Bootstraped results ($N = 1000$) of test data

UMAP applied on MNIST



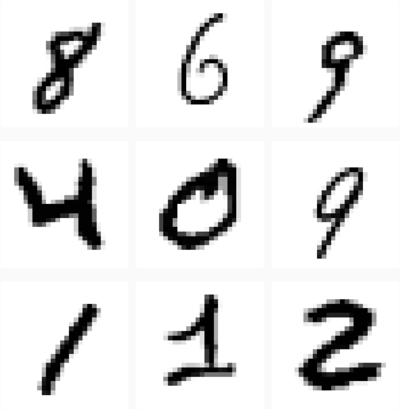
Lower dimensional UMAP embedding



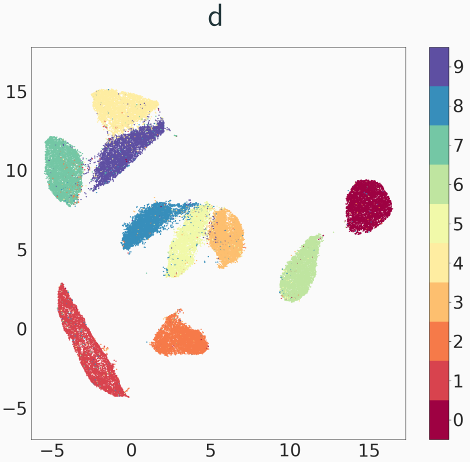
Lower dimensional UMAP embedding

d

UMAP applied on MNIST



Lower dimensional UMAP embedding



Lower dimensional UMAP embedding