Evaluating robustness of tabular models under meta-features based shifts

Source code

NeurIPS Workshop Reliable ML from Unreliable Data

Irina Deeva, Nargiza Amerkhanova, Alena Kropacheva

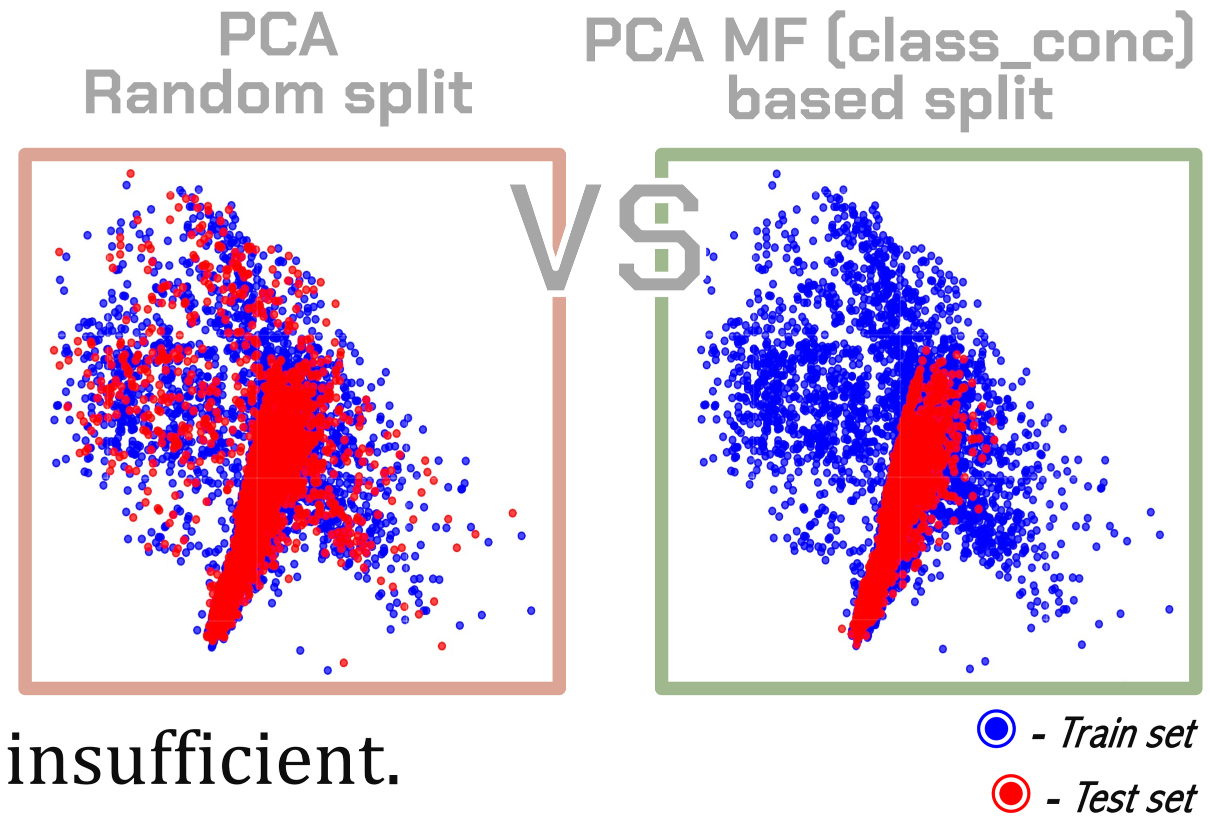
AI Institute, ITMO University Saint-Petersburg, Russia

Introduction

Problem: Standard random train-test splits fail to capture *realistic distribution shifts*, leading to overoptimistic model performance estimates.

Hypothesis: We can *create more challenging and realistic validation sets* by explicitly optimizing distributional differences *through meta-features (MFs)*.

Contribution: Evolutionary algorithm for optimizing meta-feature based data splits; Synthetic OOD generation method when splits are insufficient.



Proposed Approach

MFs are measurable properties of datasets that describe their structure and complexity (for example: IQR, mutual info, class concentration, etc)

MFs Based Split: Given source dataset the NSGA-II evo algo finds Pareto-optimal *splits that maximize the directed distributional differences between train subset S and test subset T* across meta-features

$$d_j(S,T) = \frac{m_j(\text{test})}{m_j(\text{train})}$$ while preserving class balance o_{imb}.

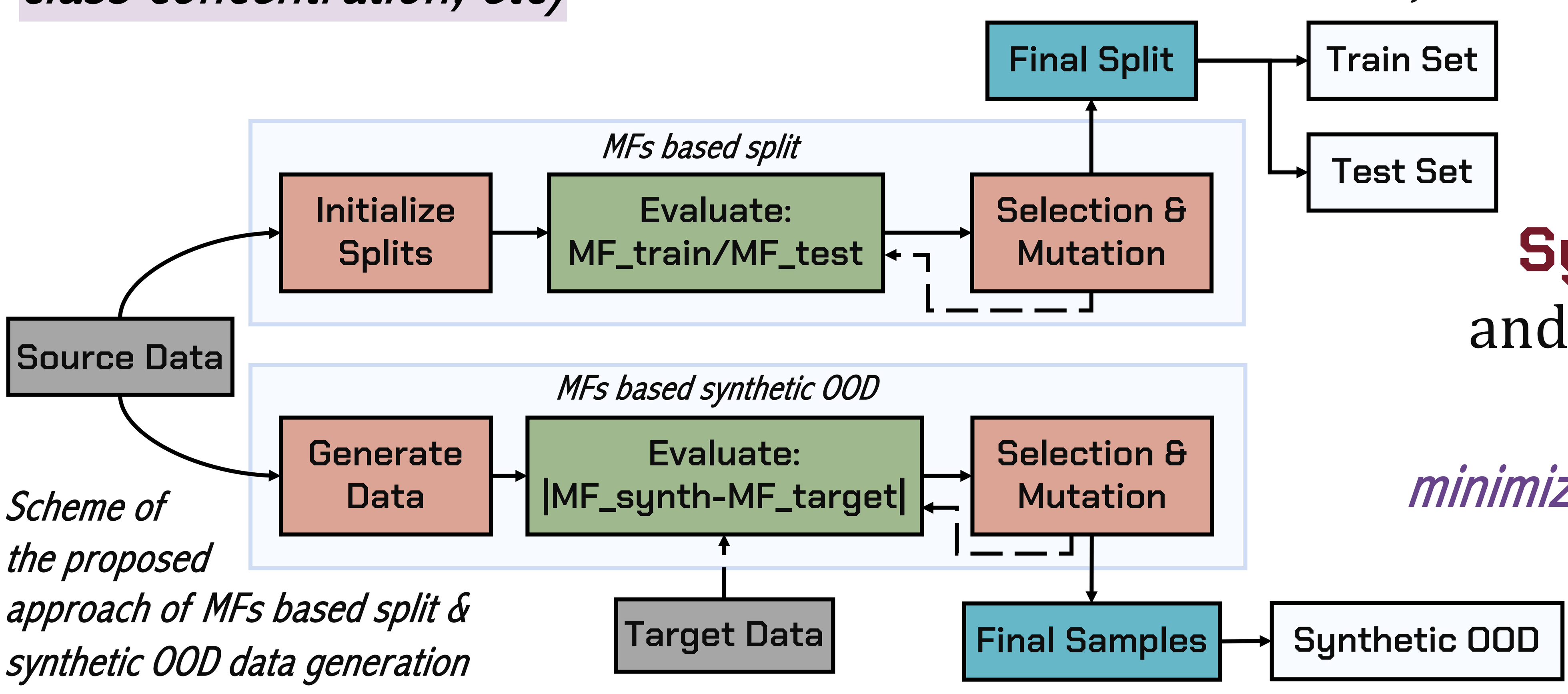
Thus, our multi-objective fitness function is:

$$f(T) = (d_1(S,T), \ldots, d_p(S,T), o_{imb}(S,T)).$$

Synthetic OOD Generation: Given source data and target MFs m_j^* the NSGA-III evo algo generates synthetic *dataset S' matching target meta-features by minimizing the L2-distances between their MFs and target MFs*.

Thus, our multi-objective fitness function is:

$$f(S') = (\|m_1(S') - m_1^*\|_2, \ldots, \|m_p(S') - m_p^*\|_2).$$



Scheme of the proposed approach of MFs based split & synthetic OOD data generation

Experiments

We tested 5 MFs: *(i)* *mut_inf*, *(ii)* *class_conc*, *(iii)* *joint_ent*, *(iv)* iq_range and *(v)* *attr_ent*. Splitting by *mut_inf, class_conc and joint_ent consistently degraded performance* across all models (LR, XGB, IRM, DRO), as these capture concept shift - changes in feature-target relationships. In contrast, *iq_range and attr_ent splits showed minimal impact*, reflecting only covariate shift without affecting model generalization.

All models - including robust architectures - *proved vulnerable to MFs-based splits*.

For datasets where MFs splits poorly matched real OOD (electricity, california), we applied *synthetic data* generation, which *strongly correlated with actual OOD performance*, validating synthetic data as a viable testing approach.

Split	Dataset	LR	XGB	IRM	DRO
Random (ID)	electricity	0.798 (20%)	0.832 (20%)	0.813 (16%)	0.814 (17%)
	taxi	0.752 (6%)	0.778 (12%)	0.790 (15%)	0.712 (6%)
	income	0.678 (38%)	0.716 (23%)	0.618 (38%)	0.514 (10%)
	california	0.823 (8%)	0.869 (13%)	**0.693** (10%)	0.821 (12%)
	acs_accidents	0.719 (16%)	0.863 (13%)	0.867 (45%)	0.702 (22%)
MFs (ID)	electricity	0.735 **(14%)**	0.749 **(12%)**	0.795 **(14%)**	0.766 **(12%)**
	taxi	**0.526** (4%)	0.592 **(6%)**	0.773 (14%)	0.505 **(4%)**
	income	**0.600 (30%)**	**0.605 (12%)**	**0.535 (30%)**	0.381 **(1%)**
	california	**0.776 (3%)**	0.831 (9%)	0.837 **(4%)**	0.786 **(8%)**
	acs_accidents	0.461 **(0%)**	0.725 **(0%)**	0.716 (30%)	0.630 (15%)
MMD (ID)	electricity	**0.419** (18%)	**0.335** (30%)	**0.435** (22%)	**0.355** (29%)
	taxi	0.703 **(2%)**	**0.564** (9%)	**0.690 (5%)**	**0.448** (21%)
	income	0.629 (33%)	0.622 (13%)	0.644 (40%)	0.650 (23%)
	california	0.828 (9%)	**0.829 (9%)**	0.890 (10%)	0.855 (15%)
	acs_accidents	**0.459** (10%)	**0.673** (6%)	**0.583 (17%)**	**0.468 (1%)**
Target (OOD)	electricity	0.596	0.633	0.655	0.646
	taxi	0.687	0.655	0.637	0.654
	income	0.297	0.488	0.240	0.418
	california	0.742	0.738	0.795	0.705
	acs_accidents	0.563	0.730	0.413	0.479

*F1-scores (% = gap to target) where ID = trained/tested on subsets from source, OOD = trained on source/tested on real OOD, **Random/MMD** = performance with % showing gaps to real OOD, **MFs** = worst-case across all tested MFs with % - closest gap to real OOD among all MFs; Bold F1 - the largest performance drops; bold % - the closest overall gaps to OOD metrics; **colors** in MFs rows indicate which specific meta-feature achieved best performance for that case.*

Dataset	Meta-features	LR	XGB	DRO	IRM
electricity	mut_inf, class_conc, iq_range	0.613 ± 0.08	0.641 ± 0.09	0.587 ± 0.08	0.613 ± 0.08
	mut_inf, class_conc	0.611 ± 0.01	0.625 ± 0.01	0.589 ± 0.01	0.632 ± 0.02
california	mut_inf, class_conc, iq_range	0.636 ± 0.05	0.692 ± 0.02	0.661 ± 0.02	0.561 ± 0.11
	mut_inf, class_conc	0.679 ± 0.07	0.713 ± 0.03	0.628 ± 0.10	0.682 ± 0.05

F1-scores for models tested on synthetic OOD data using MFs (mut_inf, class_conc, iq_range)