Ensemble techniques for PDFA learning

Diversity-Driven Ensemble Learning with the Alergia Algorithm

Author: Błażej Łytkowski (b.lytkowski@student.tudelft.nl)
Supervisor: Simon Dieck
Responsible professor: Sicco Verwer

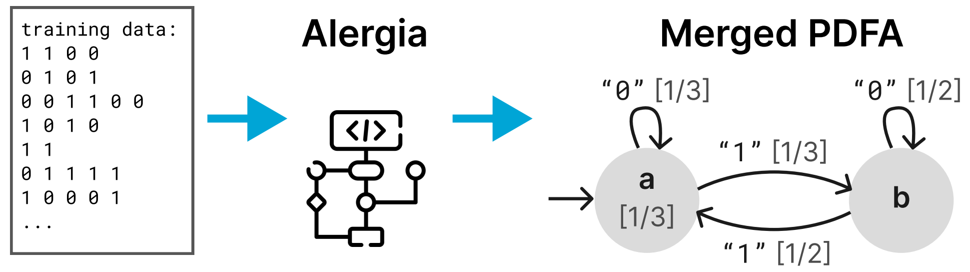
TU Delft

1. Why Ensembles for PDFA Learning?

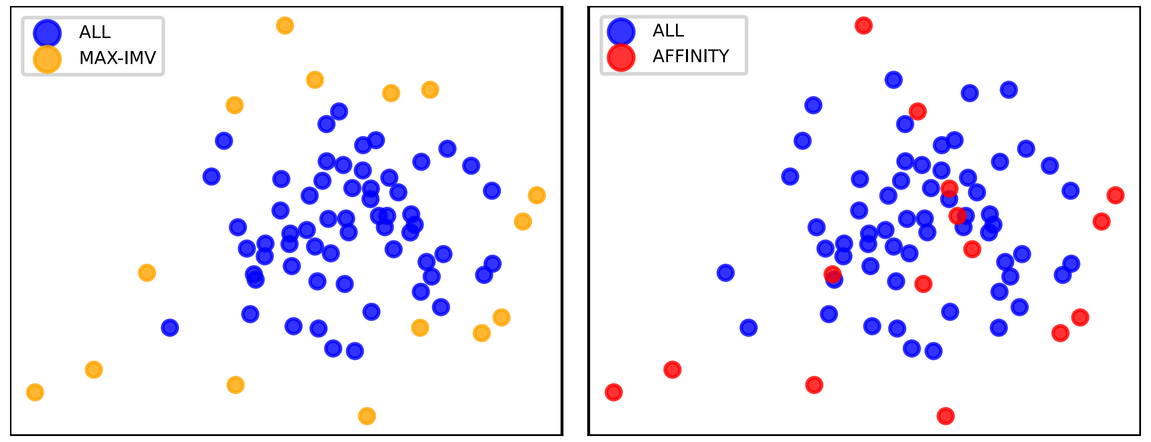
PDFA learning is used for interpretable machine learning tasks like software analysis and anomaly detection. **Alergia** is a classical algorithm for learning PDFAs from positive-only data. **Ensemble methods** are widely used for many ML methods but not explored in PDFA learning.

Research questions:

- RQ1: How to build an ensemble from Alergia?
- RQ2: How to increase ensemble diversity?
- RQ3: How does an ensemble compare to a single model?



Pruning selections visualized with MDS



2. How We Built the Ensemble

- We introduce randomness into Alergia's merge selection to produce diverse models. When selecting merges, we skew each candidate score by: $\text{score} \cdot \text{rand}(1-r, 1)$
- We use uniform voting to aggregate the probability predictions of individual models.
- The training of the random models and predicting with an ensemble is done in parallel. This allows to generate large number of models, exploring the search space.

3. Capturing Diversity: the IMV Score

We introduce **sample cross entropy** as a metric to quantify disparity between models. It is computed by generating a set of traces on model B and evaluating them on model A. The metric compares B's target probabilities and A's predictions:

$$h(A \to B) = \sum_{x} B(x) \log A(x)$$

Previous work in ensemble learning by Wood et. al. formalized ensemble diversity as a key factor in optimizing an ensemble model. The sample cross entropy metric enables us to define an **Inter-Model Variety (IMV)** score that quantifies the diversity of an ensemble E.

$$\text{IMV}(E) = \sum_{A,B \in E} h(A \to B) - h(B \to B)$$

4. Ensemble Pruning Methods

Independent generation of models enables us to parallelize the training to quickly explore a large portion of the whole search space. Simply combining all of the trained models isn't always optimal because:

- ensemble containing all produced models has a larger **computing power footprint**,
- excessive exploration leads to **overfitting**.

We address these issues by employing ensemble pruning. This works by selecting models from the generated population to create a smaller ensemble. We consider 2 types of pruning methods with different focuses:

1. Directly maximizing the diversity - achieved by Max-IMV, which heuristically maximizes the IMV score.
2. Finding representative coverage of the model space - done by employing clustering methods: K-Medoids and Affinity Propagation.

5. Experiments & Results

Single Alergia model vs an ensemble:

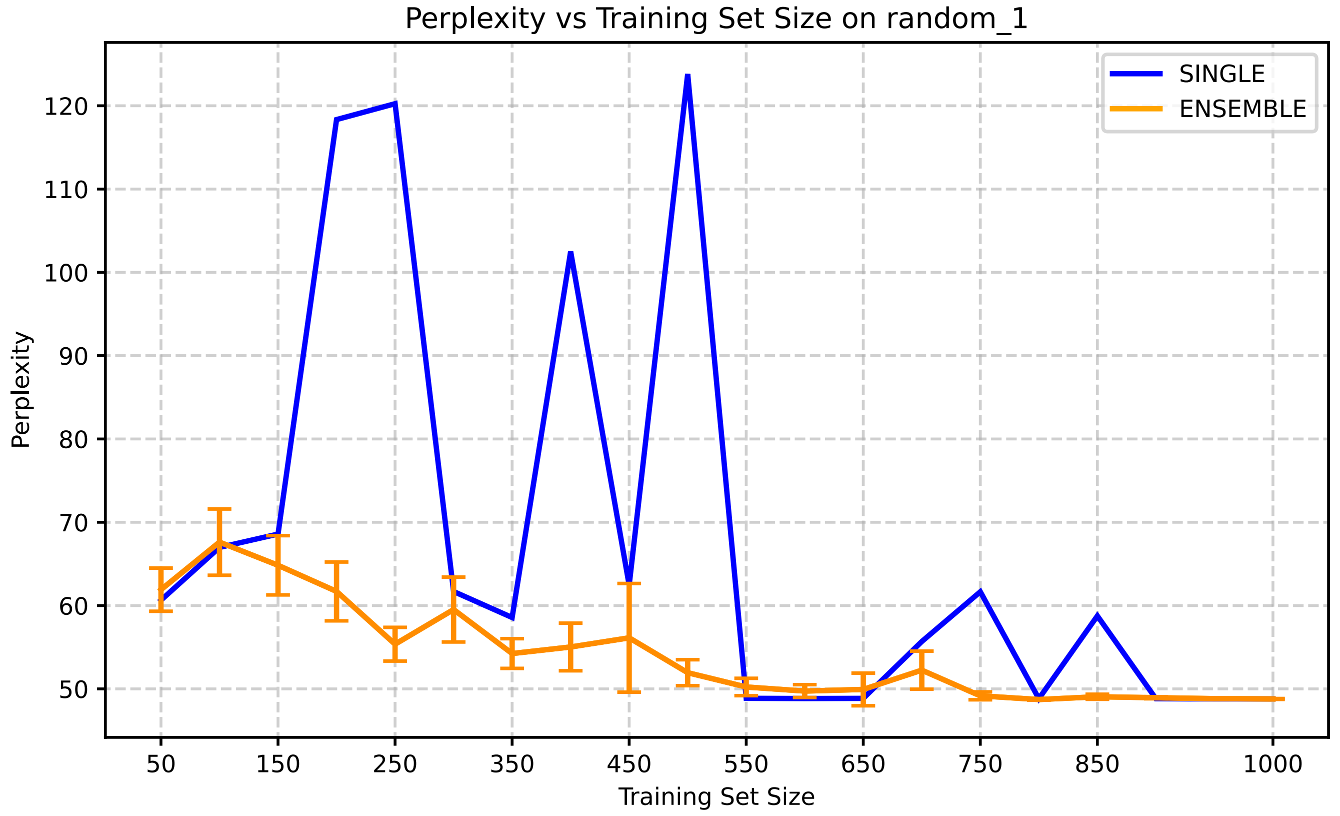
- Ensemble converges more smoothly as training set grows.
- Ensemble is more robust to skewed data.

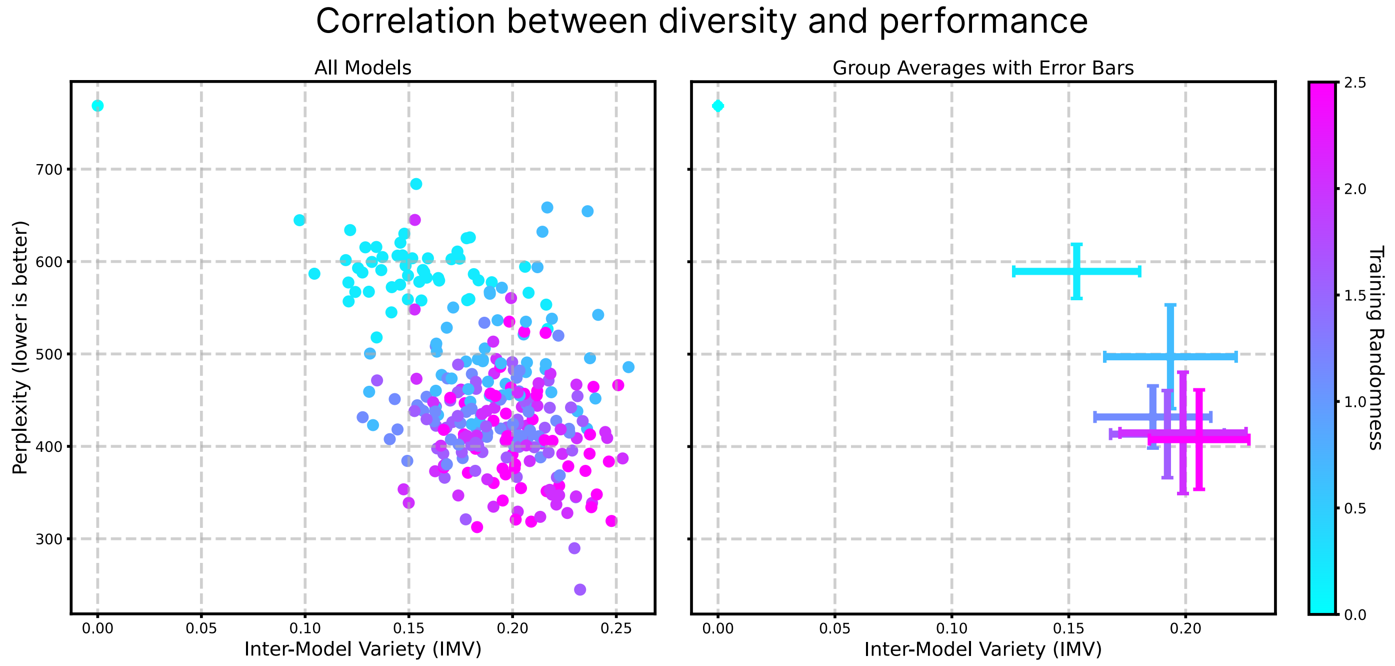
Pruning methods vs the full ensemble:

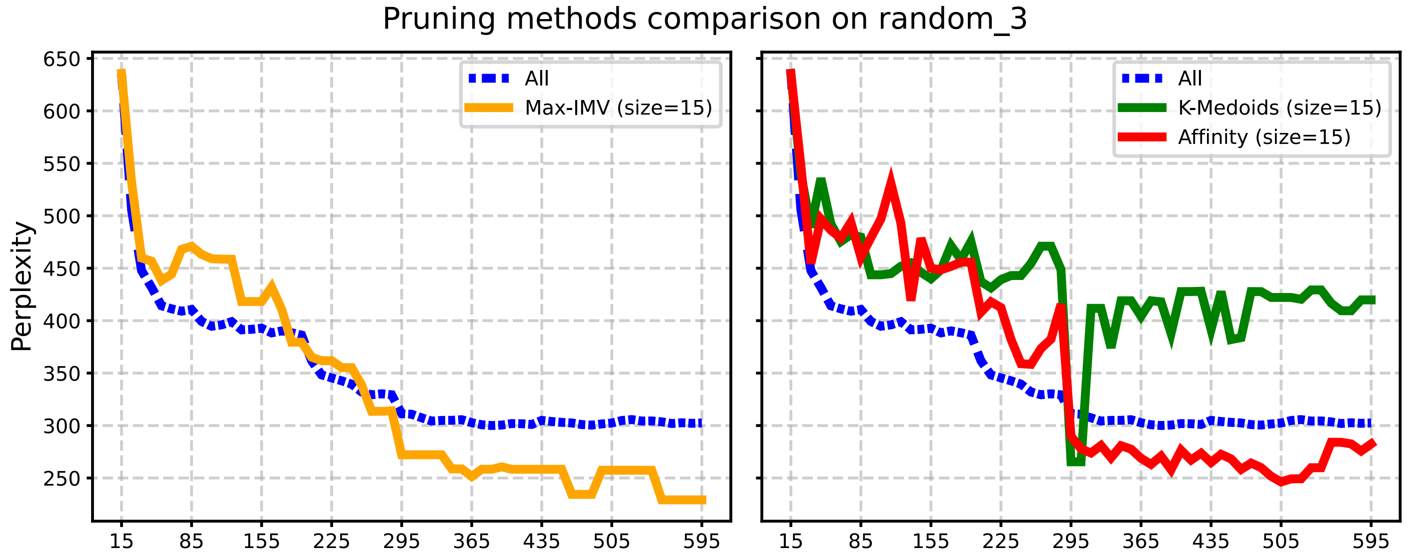
- Pruned ensembles often match or outperform full.
- Especially useful with sparse data.

Anomaly detection on the HDFS dataset:

- The pruned ensemble excels at all sizes of the training data.
- Regular ensembles outperform single model on medium sized training sets.







6. Conclusions

- This research shows that ensemble learning can significantly enhance PDFA learning, a domain that traditionally relies on single, deterministic models.
- By encouraging diversity through controlled randomization and pruning, we can improve generalization and stability, without increasing model complexity.
- In the context of anomaly detection on sparse datasets, we demonstrate that a well-pruned, diverse ensemble can outperform random ensembles trained on much larger data volumes.
- This suggests that Alergia ensembles with smart model selection can substitute for lack of data, offering a lightweight, interpretable alternative to heavier machine learning methods.