Ensemble Methods for (P)DFA Learning

Random Ensemble for DFA Learning

Georgios Tsampikos Kontos[1] Sicco Verwer[1] Simon Dieck[1]
[1] Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology

Introduction

A *Deterministic Finite Automaton (DFA)* processes strings one symbol at a time by transitioning within a finite set of states, and determines whether to accept the strings by the final state.

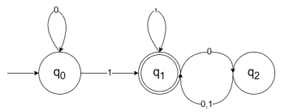

Figure: A simple DFA with three states

DFAs can be used for classification and prediction tasks. First, a prefix tree structure, the *Augmented Prefix Tree Acceptor (APTA)* [3] is built from the data. After the APTA is constructed, its states are merged according to heuristics to identify minimal DFAs.

Research Questions

- How can we create an ensemble approach to DFA learning that does not use suitability heuristics and encourages inter-model variety?
- How can we measure inter-model variety for an ensemble of DFAs?
- How do ensembles that emphasize inter-model variety perform in comparison to heuristic learned models?

Merge Tree: Handling Merge Sequences

We define the *merge tree*, a tree-like data structure to represent merge sequences from the original APTA. We can construct multiple automata in parallel by making shared intermediate decisions.

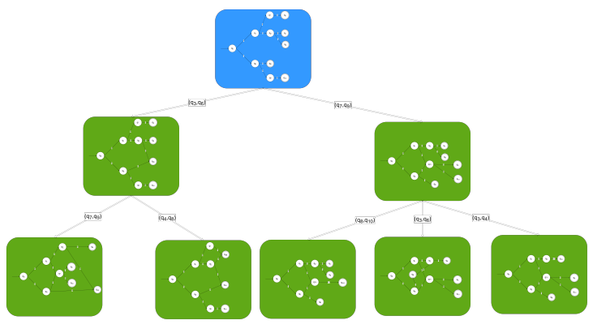

Figure: 3-level Merge Tree

TU Delft Delft University of Technology

Balanced Merge Tree Exploration and Pruning

The *Balanced Merge Tree Exploration (BMTE)* algorithm constructs diverse DFAs by distributing automata across branches of the merge tree in a balanced manner. At each node, DFAs are allocated as evenly as possible among child nodes to avoid congestion in specific branches.

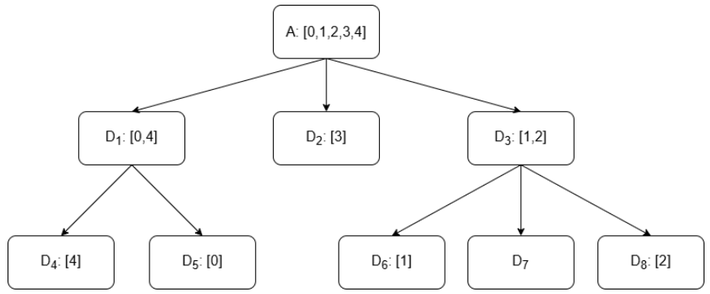

Figure: Balanced Distribution of DFAs among children

Identical merges can still occur in different levels of branches, leading to structural similarities. To address this issue, we can *prune* branches of the merge tree that have previously been visited.

Agreement Connectivity: A Global Metric for Inter-Model Variety

We introduce *agreement connectivity*, a global metric of the ensemble's diversity. We construct a graph $\mathcal{G}$ with the models of the ensemble as nodes and edges weighted by the pairwise agreement rate of the automata they connect. The agreement connectivity is defined as the algebraic connectivity [2] of $\mathcal{G}$.

The agreement connectivity of an ensemble measures how difficult it is to isolate groups of models that behave similarly to one another. High agreement connectivity implies the ensemble agrees to a large extent. Low agreement connectivity suggests that models often deviate from the ensemble's consensus.

Experimental Setup

We evaluate BMTE and BMTE with pruning on 20 datasets of varying size and sparsity, originating from the STAMINA competition [5]. We create ensembles of 100 automata and average results over 20 iterations of the experiments. Our algorithms are compared to a single automaton learned with EDSM [4] and an ensemble of 100 independently created models. Experiments are conducted using the 12 CPU cores in the Delft Blue supercomputer [1].

Results

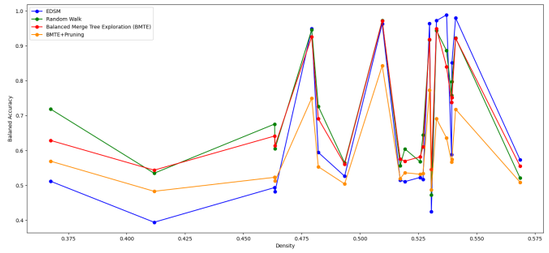

Figure: Balanced Accuracy

BMTE and random walk ensembles surpass EDSM and pruned BMTE in predictive accuracy and diversity of automata. While Random Walk slightly outperforms BMTE for sparse data, the ensembles created by the latter are more robust and consistent.

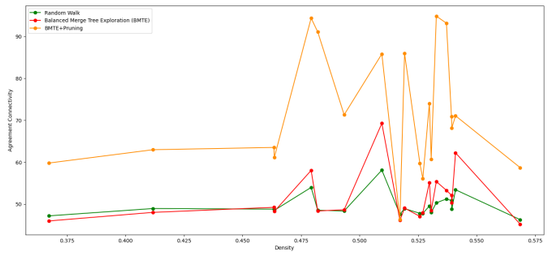

Figure: Agreement Connectivity

Conclusions

BMTE outperforms the state-of-the-art suitability heuristic in sparse inputs, while achieving comparable performance in denser datasets. Furthermore, pruning repeated merges hinders the ensemble's diversity and predictive performance.

References

[1] D. H. P. C. C. (DHPC). *DelftBlue Supercomputer (Phase 2)*. 2024.
[2] M. Fiedler. "Algebraic connectivity of graphs". In: *Czechoslovak mathematical journal* 23.2 (1973), pp. 298–305.
[3] M. J. Heule and S. Verwer. "Exact DFA identification using SAT solvers". In: *Grammatical Inference: Theoretical Results and Applications: 10th International Colloquium, ICGI 2010, Valencia, Spain, September 13-16, 2010. Proceedings 10*. Springer. 2010, pp. 66–79.
[4] K. J. Lang, B. A. Pearlmutter, and R. A. Price. "Results of the abbadingo one DFA learning competition and a new evidence-driven state merging algorithm". In: *International Colloquium on Grammatical Inference*. Springer. 1998, pp. 1–12.
[5] N. Walkinshaw, B. Lambeau, C. Damas, K. Bogdanov, and P. Dupont. "STAMINA: a competition to encourage the development and assessment of software model inference techniques". In: *Empirical software engineering* 18.4 (2013), pp. 791–824.