DFA Learning: Subsampling

Minimal models from Subsamples vs Heuristic models from Full Data

Author: Matei Hristodorescu (mhristodorescu@tudelft.nl)

Introduction

Learning the smallest deterministic finite automaton (DFA) consistent with labeled traces is a key challenge in grammatical inference, with applications in areas such as linguistics, bioinformatics, and verification. While optimal DFA inference is NP-complete, heuristic methods like EDSM are often used in practice. However, large datasets make even these heuristics computationally demanding. Motivated by the Myhill-Nerode theorem and the concept of a minimal characteristic sample, we introduce two heuristics that aim to reduce training data by eliminating redundant traces while preserving distinguishing information. Our experiments show that models learned from these reduced samples outperform those generated by EDSM on full datasets and slightly surpass models from random sampling.

Research Question

What is a good way to subsample a dataset such that it retains as much information as possible?

Methodology

Our sampling method is based on an idea from the Myhill-Nerode theorem: if two traces end with the same suffix and give the same result, they likely provide similar information about the DFA. This means we can safely remove some of these traces without losing important information. To do this, we group traces by their last k characters and sample from each group. We test two ways to choose k: Dynamic k Sampling, which starts with a small k and increases it while sampling until we reach the desired number of traces, and Binary Search k Sampling, which searches for a k that gives a target number of groups and then picks diverse traces from each group using Levenshtein distance. Both methods aim to reduce the dataset while keeping the most useful traces.

Conclusion

We propose two methods to reduce DFA training sets by removing similar traces, based on an idea from the Myhill-Nerode theorem. This speeds up learning while keeping high accuracy. Our results show these methods outperform both random sampling and the EDSM heuristic.

Future Work

A possible direction for future work is to make our sampling methods more adaptive to the structure of the dataset. Instead of sampling at a fixed rate, the algorithm could selectively remove traces it identifies as redundant with a certain level of confidence. This would allow the method to work more effectively on both dense and sparse datasets, improving flexibility and overall model accuracy.

Results

Table 1: Average accuracy of EDSM heuristic vs optimal learning with different sampling methods at 25% and 50%, on the testing set.

DFA Size	EDSM	Random25	Binary25	Dynamic25	Random50	Binary50	Dynamic50
9	75.4	60.4	51	54.4	99.2	98.5	100
10	74	64.4	66.8	65.4	93	97	95.6
11	97.6	70.2	57.2	61.6	89	96.8	97.8
12	79.4	75.4	73.4	72.6	90.6	94.8	89.6
13	77	72.2	51.4	72.6	97.4	98	97.4
Average	80.6	68.5	59.9	65.3	93.8	97	96.1