Project: Constraint Programming

Vasily Apasov

April 25, 2025

1 Problem Statement

Given two finite sets of binary words `Acc` (*positive* samples) and `Rej` (*negative* samples),
find a *minimal* deterministic finite automaton (DFA) $\mathcal{A} = \langle Q, \Sigma, \delta, q_0, F\rangle$ with $\Sigma = \{0, 1\}$ that accepts every word in `Acc` and rejects every word in `Rej`. The criterion of optimality is the number of states $|Q|$.

2 Model Overview

For a fixed candidate size $N = |Q|$ the model is encoded as a *Constraint Satisfaction Problem* (CSP) and solved with Gecode.
Variables encode the transition table, accepting flags and the run of the automaton on all sample prefixes. Constraints enforce deterministic behaviour, sample consistency and symmetry breaking. Search iterates $N = 1, 2, \ldots$ until the first feasible solution emerges, which is therefore minimal.

3 Variables

Transition variables

$$T_0[i] \in \{0, \ldots, N-1\} \qquad (0 \le i < N),$$
$$T_1[i] \in \{0, \ldots, N-1\} \qquad (0 \le i < N);$$

$T_b[i]$ is the target state reached from state i upon input $b \in \{0, 1\}$.
Accepting flags

$$A[i] \in \{0, 1\}, \qquad 0 \le i < N.$$

$A[i] = 1$ iff state i is accepting.
Prefix–run variables Let P be the concatenation of *all* sample prefixes[1]. Then

$$S[k] \in \{0, \ldots, N-1\}, \qquad 0 \le k < |P|.$$

$S[k]$ represents the DFA state after reading prefix P_k.

4 Constraints

Let $\langle o_w, \ell_w, \mathbf{w}\rangle$ be offset, length and content of a sample word $\mathbf{w}$. The following constraints are posted for every word.

C1. **Start** $S[o_w] = 0$.

[1] Including the empty prefix ε for every word.

C2. **Deterministic transitions**

$$\forall j \in [0, \ell_w - 1] : \quad \begin{array}{ll} \mathbf{w}[j] = 0 & \Rightarrow \quad T_0[S_{o_w+j}] = S_{o_w+j+1}, \\ \mathbf{w}[j] = 1 & \Rightarrow \quad T_1[S_{o_w+j}] = S_{o_w+j+1}. \end{array}$$

Implemented with Gecode's `element` constraint.

C3. **Acceptance test**

$$A[S[o_w + \ell_w]] = \begin{cases} 1, & \mathbf{w} \in \mathtt{Acc}, \\ 0, & \mathbf{w} \in \mathtt{Rej}. \end{cases}$$

Symmetry breaking

DFA states are anonymous; any permutation of names yields an equivalent automaton, causing factorial symmetry. We impose the *first-occurrence order*: the first time each state id appears in S must respect $0 \prec 1 \prec \cdots \prec N - 1$. In Gecode this is a single call

```
precede(*this, S, [0,1,2,...,N-1]);
```

which eliminates all $N!$ permutations.

5 Search Strategy

1. **Branch** on S first ('minimum domain size, min value'). Fixing a prefix–state instantly propagates through the `element` constraints, shrinking domains of T_0, T_1.
2. Then branch on T_0 and T_1 ('none, min value'). At this point many transitions are already singleton.
3. Finally branch on boolean array A.

The external loop increments N from 1 up to the prefix bound $|P|$; the first solution is therefore minimal.

6 Correctness Sketch

- Any solution of the CSP defines a deterministic complete DFA. C1–C3 guarantee that every positive / negative sample is accepted / rejected respectively.
- The outer search halts no later than $N = |P|$, because the prefix tree itself is a feasible DFA of that size.
- Hence the algorithm outputs a DFA that is both consistent and minimal.