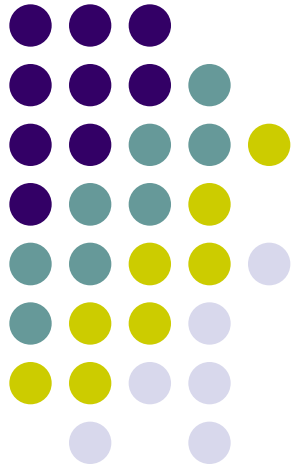
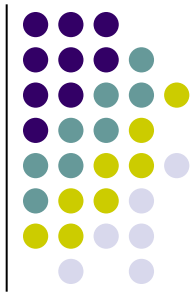
The Stamina Competition

April 2011



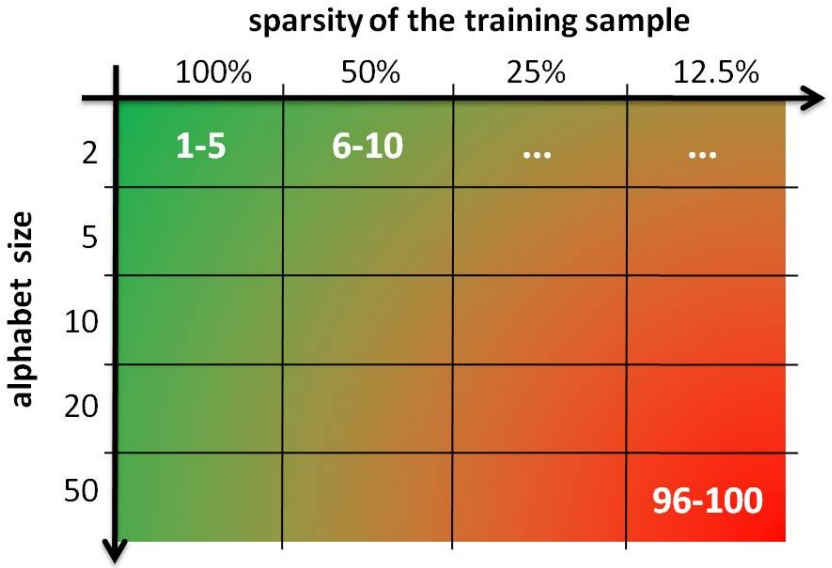
Overview



- Online Regular Induction Contest
  - Extends former competitions, especially Abbadingo
  - Cross-fertilization between the machine learning and software engineering communities
- Key points
  - Focus on the complexity of the learning with respect to the alphabet size
  - Adapted generation protocol for state machines and samples to mimic features of behavior models
- Not an evaluation of the thesis techniques *per se*
  - Unsupervised learning (i.e. no oracle, no queries)
  - No pruning with fluents, goals, control information

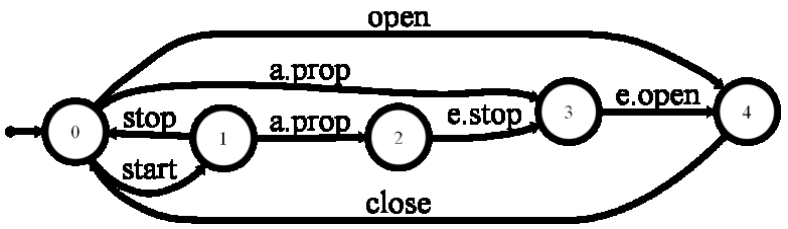
Competition overview

- 100 induction problems (20 cells of 5 problems)
- Two difficulty dimensions
  - alphabet size vs. sparsity of learning sample



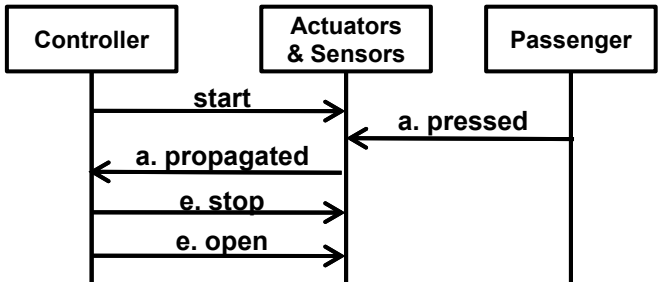
- Solving a problem
  - Download learning (labeled) and test (unlabeled) samples
  - Learn a model (typically a DFA)
  - Label the test sample using learned model
  - Submit labeling on the competition server

Scientific setup
State Machines



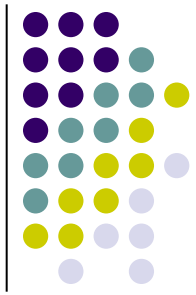
- Approach
  - Review of SE litterature to identify representative features of behavior models
  - Tuning of the Forest-fire algorithm to mimic these features
- Main features
  - Approximately 50 states (to avoid adding a third difficulty dimension to the competition)
  - Alphabet sizes ranging from 2 to 50 letters
  - Equal proportion of accepting vs. rejecting states
  - Large variance of degree distribution, to mimic behavior models

Scientific setup
Samples



- Approach
  - Generated by the target machine: random walk algorithm
  - Negative strings by randomly perturbing positive ones
    - three kinds of edit: substitution, insertion and deletion
  - Tuned to ensure good induction results using Blue-Fringe on the simplest problems
- Main features
  - Learning and test samples do not overlap
  - Learning samples may contain duplicates, as a consequence of the random walk generation
  - String length distribution: centered on *5 + depth(automaton)*

Scientific setup
Submission & Scoring



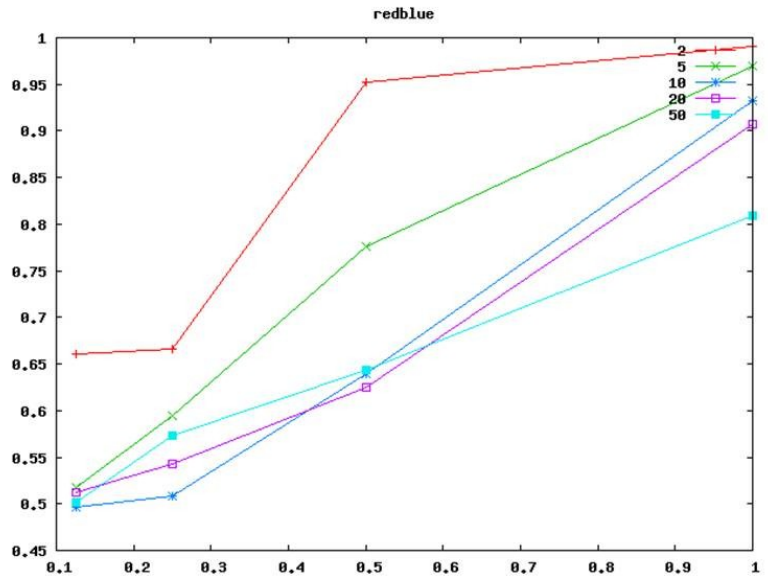
- Submission
  - Solutions submitted as binary strings labelling the test sample
  - Binary feedback (problem broken or not broken), to avoid hill-climbing
- Scoring
  - Balanced Classification Rate to place equal emphasis on accuracy in terms of positive and negative strings
  - Problem broken if BCR score >= 0.99
  - A cell is broken if all problems it contains are broken

Scientific setup
Baseline

- Problem grid empirically ajusted
  - To ensure good induction results using Blue-Fringe on the simplest problems
  - Without breaking the cell

sparsity of the training sample

alphabet size	100%	50%	25%	12.5%
2	0.989	0.95	0.67	0.66
5	0.96	0.77	0.60	0.51
10	0.93	0.64	0.50	0.50
20	0.90	0.63	0.54	0.50
50	0.80	0.64	0.57	0.50

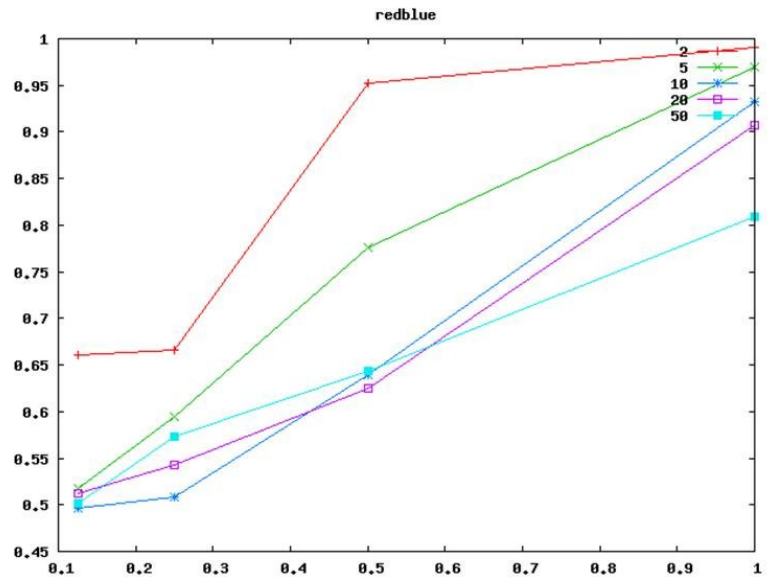


Scientific setup
Baseline: lessons learned

- RPNI and BlueFringe converge on largest alphabets
  - Theoretically expected, big samples needed in practice
- Size of the alphabet "hurts" convergence in practice
  - Confirms experimentally what we expected theoretically
  - Supports the interest of launching Stamina

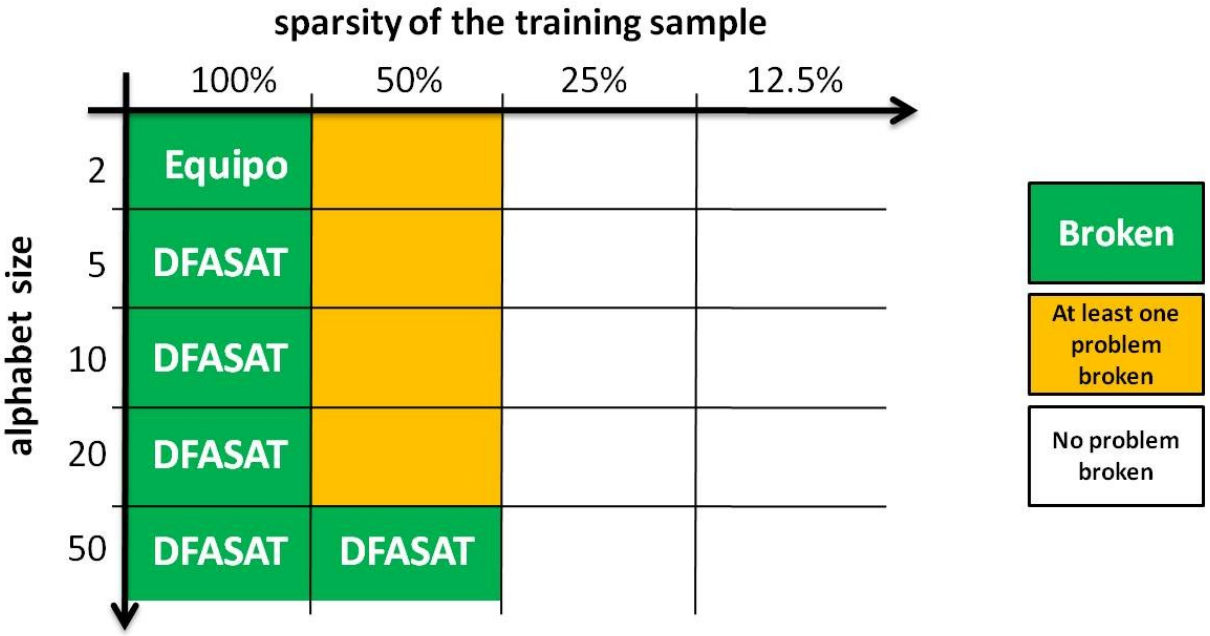
sparsity of the training sample

alphabet size	100%	50%	25%	12.5%
2	0.989	0.95	0.67	0.66
5	0.96	0.77	0.60	0.51
10	0.93	0.64	0.50	0.50
20	0.90	0.63	0.54	0.50
50	0.80	0.64	0.57	0.50



Participation overview

- Between march and december 2010 (official)
  - 1856 submissions made by 11 challengers
  - 65 winning submissions broke 42 problems
  - 6 cells broken, by 2 challengers (Equipo & **DFASAT**)



A big winner - DFASAT

Marijn Heule & Sicco Verwer

Blue-Fringe

alphabet size	100%	50%	25%	12.5%
2	0.989	0.95	0.67	0.66
5	0.96	0.77	0.60	0.51
10	0.93	0.64	0.50	0.50
20	0.90	0.63	0.54	0.50
50	0.80	0.64	0.57	0.50

DFASAT

alphabet size	100%	50%	25%	12.5%
2	0.99	0.98	0.78	0.66
5	0.99	0.96		
10	0.99	0.97		
20	0.99	0.98		
50	0.99	0.99	0.96	