Learning Regular Expressions

Jan Plhák

Contents

- Problem specification
- Finite automatons orientated approach
- Regular expression orientated approach

Finite automatons orientated approach

- Gold, E. M. (1967). Language identification in the limit

- Learning from Queries
- Evolutionary algorithms
  - Genetic algorithms
  - Active Coevolutionary Learning
- Learning Regular Languages Using RFSA

Learning from Queries

- Structurally complete set
- Mighty Oracle (the teacher)
- Candidate elimination

- Membership queries
  - Is element in the target language?
- Equivalence queries
  - Is current hypothesis correct?
  - Counterexample

Learning from Queries

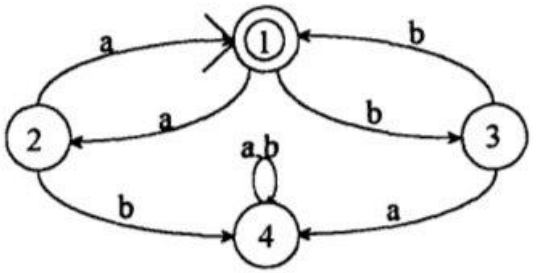
Abbadingo One - DFA learning competition - 1997 => Evidence Driven State Merging (EDSM)

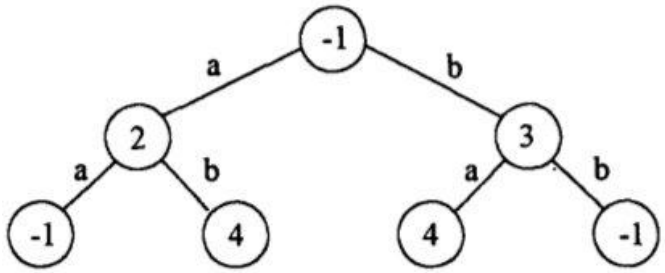
Interesting papers

- LANG, Kevin J., Barak A. PEARLMUTTER a Rodney A. PRICE. Results of the Abbadingo one DFA learning competition and a new evidence-driven state merging algorithm: 4th international colloquium, ICGI-98, Ames, Iowa, USA, July 12-14, 1998 : proceedings. *Grammatical Inference*. Berlin/Heidelberg: Springer-Verlag, 1998, s. 1. DOI: 10.1007/BFb0054059.
- Parekh R.G., and Honavar V.G. Efficient Learning of Regular Languages using teacher supplied positive samples and learner generated queries. In Proceedings of the Fifth UNB Conference on AI, Fredrickton, Canada. August 11-14, 1993.
- Florentin Ipate, Learning finite cover automata from queries, Journal of Computer and System Sciences, Volume 78, Issue 1, January 2012, Pages 221-244, ISSN 0022-0000, 10.1016/j.jcss.2011.04.002. (http://www.sciencedirect.com/science/article/pii/S002200001100047X)
- S. M. Lucas and T. J. Reynolds. Learning deterministic finite automata with a smart state labelling evolutionary algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27: 1063–1074, 2005.

Genetic algorithms

- Structurally complete set required
- S-expression
- Initial population
- Crossover, mutation
- Interesting papers:





(-1 (2 -1 4)(3 4 -1))

  - DUNAY, B.D., F.E. PETRY a B.P. BUCKLES. Regular language induction with genetic programming. *Proceedings of the First IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence*. IEEE, 1994, s. 396-400. DOI: 10.1109/ICEC.1994.349918.
  - EHRENBURG a JEROEN VAN MAANEN. A Finite Automaton Learning System using Genetic Programming. *NeuroCOLT Technical Report Series*. IEEE, 1995
  - Scott Brave. Evolving deterministic finite automata using cellular encoding. In John R. Koza, David E. Goldberg, David B. Fogel, and Rick L. Riolo, editors, Genetic Programming 1996: Proceedings of the First Annual Conference, pages 39–44, Stanford University, CA, USA, 28–31 July 1996. MIT Press.

Active coevolutionary learning

- Genetic algorithm
- Two populations
  - Model candidates of target DFA
    - Fitness - ability to correctly identify sentences from training set
  - Sentences that we are supposed to learn
    - Fitness - ability to cause disagreement
- Mutations only
- Estimation of the target DFA states needed

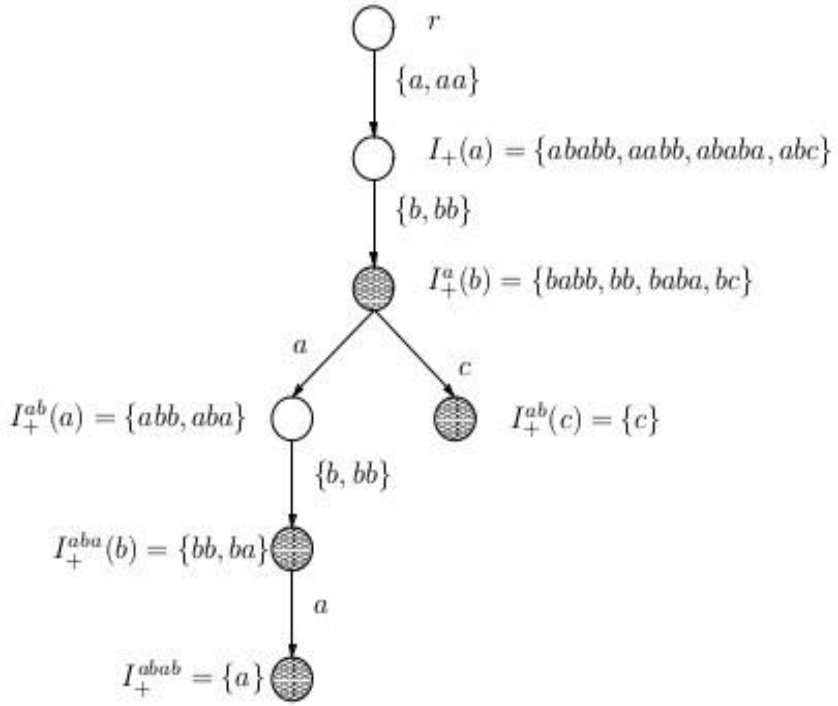
Reference: J. Bongard and H. Lipson. Active coevolutionary learning of deterministic finite automata. Journal of Machine Learning Research., 6:1651–1678, 2005.

RegExp orientated approach

- Learning from positive data
- Learning using queries
- Evolutionary algorithms

Learning from positive data

- Blockwise grouping and alignment
- Tree creation
- Introducing loops



Henning Fernau, Algorithms for learning regular expressions from positive data, Information and Computation, Volume 207, Issue 4, April 2009, Pages 521-541, ISSN 0890-5401, 10.1016/j.ic.2008.12.008.

BRĀZMA, Alvis a Kārlis ČERĀNS. Efficient learning of regular expressions from good examples. *4th International Workshop on Analogical and Inductive Inference: AII '94 5th International Workshop on Algorithmic Learning Theory, ALT '94 Reinhardsbrunn Castle, Germany*. 1994, s. 76-90. DOI: 10.1007/3-540-58520-6_55. Dostupné z: http://www.springerlink.com/index/10.1007/3-540-58520-6_55

Learning using queries

- Correction queries > edit distance
- Membership queries

Example: a+a(ab)+ab+(ab)+ab+a(bb)+

1. Empty string
2. Finding n-letter loops

KINBER, Efim. On Learning Regular Expressions and Patterns Via Membership and Correction Queries. *Grammatical Inference: Algorithms and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, s. 125. DOI: 10.1007/978-3-540-88009-7_10.

BSHOUTY, Nader a Avi OWSHANKO. Learning Regular Sets with an Incomplete Membership Oracle. *Computational Learning Theory*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001-9-13, s. 574. DOI: 10.1007/3-540-44581-1_38.