Performance Evaluation of Selected Variable Ordering Methods for NFA Induction

Tomasz Jastrząb
Silesian University of Technology
Gliwice, Poland
tomasz.jastrzab@polsl.pl

Extended abstract of the 14th International Conference on Grammatical Inference *ICGI'18*, September 5–7, 2018, Wrocław, Poland.

Abstract

It is known that nondeterministic finite automata (NFA) minimization is computationally hard. For instance, finding a minimal NFA given a deterministic finite automaton is PSPACE-complete [19]. Also, it is known that minimal NFAs are not identifiable in the limit from polynomial time and data [7] and they are also not efficiently approximable [11]. However, there are some successful induction algorithms presented in the literature, including *DeLeTe2* [9] and Nondeterministic Regular Positive Negative Inference [1], or the state merging methods discussed in [6, 10]. Moreover, there are also solutions transforming the induction problem into the constraint satisfaction problem (CSP) [14, 15, 16, 17, 18, 22], which we follow here.

Author Keywords

NFA learning algorithms; constraint satisfaction problems.

In the paper, we deal with finite automata, which are important for numerous practical applications [23, 24, 25]. The automata are *finite*, *nondeterministic*, *minimal* and *consistent* with the given sample $S = (S_+, S_-)$. This means that they accept all words from set S_+ (*examples*) and reject all words from set S_- (*counterexamples*), and no two states can be merged together without losing the consistency. The automata are defined by the quintuple $A = (Q, \Sigma, \delta, q_0, Q_F)$, where Q is the (minimal) finite set of states, Σ is the input alphabet, $\delta : Q \times \Sigma \to 2^Q$ is the transition function, $q_0 \in Q$ is the initial state and $Q_F \subseteq Q$ is the set of final states [13].

In order to find a minimal NFA we set $k = 1, 2, \ldots$ and ask whether there exists a k-state automaton consistent with the given sample S. We propose and evaluate selected variable ordering methods used in the CSP formulation of the problem [15, 18, 22]. Additionally, to assess how the sizes of sets S_+ and S_- affect the performance, we consider the samples for which $|S_+| = |S_-|$, $|S_+| \gg |S_-|$ and $|S_+| \ll |S_-|$ hold.

The proposed variable ordering method *min-max-ex* selects first the variables appearing most frequently in the shortest (in the number of product terms) constraints related to the examples. Zero is assigned before one, to each selected variable. For zeros, the constraints related to examples are checked first, for ones, we start with the other constraints. The method *min-max-cex* selects the most frequent variables appearing in the shortest constraints related to counterexamples. It also assigns ones before zeros, consequently reversing the order in which the constraints are checked against possible contradictions. In both methods we require only one type of constraints to be explicitly satisfied, with the other satisfied implicitly provided that no contradictions exist.

In the experiments we compared the two proposed methods with the well-known *deg* method [8], which is based on the decreasing number of constraints the variable is involved in (*degree*). The alternative methods include weight- and impact-based methods [4, 20], domain-size based method *dom* [12], and various combinations of *dom* and degree-based methods, which follow the *dom* heuristic but in case of ties use the variable degree [3, 5, 21].

The experiments involved 300 input samples constructed from the randomly drawn sets of amino acid sequences [2]. The number of sequences belonging to the set of examples (resp. counterexamples) was 5 (resp. 45), 25 (resp. 25), and 45 (resp. 5), for the samples, for which $|S_+| \ll |S_-|$, $|S_+| = |S_-|$, and $|S_+| \gg |S_-|$ hold. The induction algorithm was implemented in Java and was run on an Intel Xeon E5-2640 2.60GHz processor with 16 logical cores and 8 GB RAM. The induced automata had 2-4 states.

We found 296 NFAs, failing in case of 4 balanced samples (with the time limit of 3 hours). The analysis of the success rates $r = S/N \cdot 100\%$, where S is the number of solved samples and $N = 100$ is the total number of samples, shown that the proposed methods prevail mainly in case of balanced samples ($r = 82\%$ for *min-max-ex*, as compared to $r = 69\%$ for *deg*), which generally turned out to be the hardest samples. On the other hand, based on the mean and median run times, we observed that the *min-max-ex* method is better when $|S_+| \ll |S_-|$ holds, while *min-max-cex* is the fastest for the cases in which $|S_+| \gg |S_-|$ is true. This was expected, since the respective imbalanced samples favor these methods by shortening the ordering time (as they operate on examples or counterexamples only). It is also worth to note that for all samples, the mean values for the best- and worst-performing methods, differed by an order of magnitude or more (take as an example the mean times for the balanced samples, being 5 s, 57 s and 110 s, for *min-max-cex*, *deg* and *min-max-ex*, respectively).

To conclude, let us underline that the results for imbalanced samples are important, since it is not uncommon that we know just a few factors causing a disease (set S_+) and much more factors that are not responsible for this particular disease (set S_-). Hence, being able to classify these factors efficiently and correctly using the induced NFAs, can be of help in some bioinformatics tasks. To improve the induction efficiency further, we think it is worthwhile to work on some hybrid algorithms, combining different ordering methods, such as the ones proposed in [17].

Acknowledgements

The research was supported by National Science Centre Poland (NCN), project registration no. 2016/21/B/ST6/02158 and the Institute of Informatics grant no. BKM-509/RAU2/2017.

REFERENCES

1. G.I. Alvarez, J. Ruiz, A. Cano, and P. García. 2005. Nondeterministic Regular Positive Negative Inference NRPNI. In *Proc. of the XXXI Latin American Informatics Conference (CLEI'2005)*. 239—249.
2. J. Beerten, J. Van Durme, F. Rousseau, and J. Schymkowitz. 2014. WALTZ-DB. Database of amyloid forming peptides. (2014). http://waltzdb.switchlab.org/, last access: 26/05/2018.
3. Ch. Bessière and J-C. Régin. 1996. MAC and Combined Heuristics: Two Reasons to Forsake FC (and CBJ?) on Hard Problems. In *Principles and Practice of Constraint Programming—CP96 (LNCS)*, Vol. 1118. Springer-Verlag, Berlin Heidelberg, 61—75.
4. F. Boussemart, F. Hemery, Ch. Lecoutre, and L. Sais. 2004. Boosting systematic search by weighting constraints. In *Proc. of ECAI'04*. IOS Press, 146—150.
5. D. Brélaz. 1979. New methods to color the vertices of a graph. *Communications of ACM* 22, 4 (1979), 251—256.
6. F. Coste and D. Fredouille. 2003. Unambiguous automata inference by means of state merging methods. In *Proc. of European Conference on Machine Learning (ECML 2003)*. Springer, Heidelberg, 60—71.
7. C. de la Higuera. 1997. Characteristic sets for polynomial grammatical inference. *Machine Learning Journal* 27 (1997), 125–138.
8. R. Dechter and I. Meiri. 1989. Experimental Evaluation of Preprocessing Techniques in Constraint Satisfaction Problems. In *Proc. of IJCAI'89*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 271—277.
9. F. Denis, A. Lemay, and A. Terlutte. 2004. Learning regular languages using RFSAs. *Theoretical Computer Science* 313, 2 (2004), 267—294.
10. P. García, M. Vázquez de Parga, G.I. Alvarez, and J. Ruiz. 2008. Universal automata and NFA learning. *Theoretical Computer Science* 407, 1—3 (2008), 192—202.
11. H. Gruber and M. Holzer. 2007. Inapproximability of nondeterministic state and transition complexity assuming $P \neq NP$. In *Developments in Language Theory (LNCS)*, T. Harju, J. Karhumaki, and A. Lepisto (Eds.), Vol. 4588. 205–216.
12. R. M. Harallick and G. L. Elliot. 1980. Increasing Tree Search Efficiency for Constraint Satisfaction Problems. *Artificial Intelligence* 14 (1980), 263—313.
13. J.E. Hopcroft and J.D. Ullman. 1979. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley Publishing Company.
14. K. Imada and K. Nakamura. 2009. Learning Context Free Grammars by Using SAT Solvers. In *Proc. of the International Conference on Machine Learning and Applications (ICMLA '09)*. 267—272.
15. T. Jastrząb. 2016. On Parallel Induction of Nondeterministic Finite Automata. *Procedia Computer Science* 80 (2016), 257—268.

16. T. Jastrząb. 2017. Parallel induction of nondeterministic finite automata revisited. In *Proc. of the International Conference of Computational Methods in Science and Engineering (ICCMSE 2017) (AIP Conference Proceedings)*, Vol. 1906. AIP Publishing.

17. T. Jastrząb. 2018. Two Parallelization Schemes for the Induction of Nondeterministic Finite Automata on PCs. In *Proc. of the International Conference on Parallel Processing and Applied Mathematics (PPAM 2017) (LNCS)*, Vol. 10777. Springer, Cham, 279—289.

18. T. Jastrząb, Z. Czech, and W. Wieczorek. 2016. Parallel Induction of Nondeterministic Finite Automata. In *Proc. of the International Conference on Parallel Processing and Applied Mathematics (PPAM 2015) (LNCS)*, Vol. 9573. Springer, Cham, 248—257.

19. T. Jiang and B. Ravikumar. 1993. Minimal NFA problems are hard. *SIAM Journal of Computation* 22, 6 (1993), 1117–1141.

20. P. Refalo. 2004. Impact-based search strategies for constraint programming. In *Principles and Practice of Constraint Programming—CP 2004 (LNCS)*, Vol. 3258. Springer-Verlag, Berlin Heidelberg, 557—571.

21. B. M. Smith and S. A. Grant. 1997. Trying harder to fail fast. In *Proc. of ECAI'98*. John Wiley & Sons, 249—253.

22. W. Wieczorek. 2012. Induction of non-deterministic finite automata on supercomputers. In *Proc. of the International Conference on Grammatical Inference (ICGI 2012)*, J. Heinz, C. de la Higuera, and T. Oates (Eds.), Vol. 21. 237—242.

23. W. Wieczorek. 2017. *Grammatical Inference: Algorithms, Routines and Applications*. Studies in Computational Intelligence, Vol. 673. Springer International Publishing, Switzerland.

24. W. Wieczorek and O. Unold. 2014. Induction of Directed Acyclic Word Graph in a Bioinformatics Task. In *Proc. of the International Conference on Grammatical Inference (ICGI 2014) (JMLR Workshop and Conference Proceedings)*, Vol. 34. 207—217.

25. W. Wieczorek and O. Unold. 2016. Use of a Novel Grammatical Inference Approach in Classification of Amyloidogenic Hexapeptides. *Computational and Mathematical Methods in Medicine* 2016 (2016), article ID 1782732.