Current Trends in Grammatical Inference

Colin De La Higuera
EURISE, Université de Saint-Etienne
France
cdlh@univ-st-etienne.fr

Abstract. Grammatical inference has historically found it's first theoretical results in the field of inductive inference, but it's first applications in the one of Syntactic and Structural Pattern Recognition. In the mid nineties, the field emancipated and researchers from a variety of communities moved in: Computational Linguistics, Natural Language Processing, Algorithmics, Speech Recognition, Bio-Informatics, Computational Learning Theory, Machine Learning. We claim that this interaction has been fruitful and allowed in a few years the appearance of formal theoretical results establishing the quality or not of the Grammatical Inference techniques, and probably more importantly the discovery of new algorithms that can infer a variety of types of grammars and automata from heterogeneous data.

1. Grammatical Inference: The Last Few Years

Whilst it is generally accepted that the first theoretical foundations in grammatical inference were given by M.E. Gold [3], the first usable grammatical inference algorithms were proposed inside the Syntactic Pattern Recognition community [2]. The algorithms were typically based on some syntactic property from formal language theory: iteration lemma, Nerode's equivalence etc. Typical applications included classification and analysis of patterns, biological sequence classification, Character Recognition, etc. The main problem was that of dealing with positive data only (no counter-examples). The inability to cope with noisy data, or data that doesn't fit exactly into a finite state machine led to defining classes of languages that were too complex, and thus, the good formal language theories were lost.

The question of learning automata and grammars from strings was nevertheless a sufficiently general question to spread out of this community, and start being dealt with by researchers from other fields. D. Angluin, through several [1] used language learning as a central problem for Computational Learning Theory: the problem led to defining and studying different learning paradigms, as learning with the help of oracles.

In 1989 L. Pitt [10] proved the difficulty of learning Deterministic Finite Automata (DFA) under a variety of learning paradigms; these negative results were strengthened in the following years by other researchers, proving that within the common PAC-learning framework, inferring DFA was simply too hard.

F.J. Ferri et al. (Eds.): SSPR&SPR 2000, LNCS 1876, pp. 28-31, 2000.

In 1993 the first, rather informal, ICGI[1] meeting took place in Britain. It was followed by ICGIs in Spain (1994), France (1996), USA (1998) and Portugal (2000) [9, 6, 5]. These meetings provided the researchers in the area with the occasion to meet and discuss new algorithms, paradigms and problems. Significant results have thus been obtained in the last few years [4].

Another factor that has enabled recent progress has been that of the availability of benchmarks, and large scale competitions [7, 8]. Also applications and data from a variety of other fields (Bio-Informatics, Speech Recognition, Time Series) have also put emphasis on developing new methods [11].

2. Some Recent Grammatical Inference Results

Grammatical inference, otherwise referred to as grammar induction or language learning is now a specific research domain with it's community, it's conference, and some well established techniques.

The advantages of using grammatical inference may be seen as the following:

- The objects grammatical inference deals with are formal grammars or automata, with a well-studied theory that can be used to design new algorithms or to prove the convergence of these algorithms.
- Grammatical inference is one of the few (reasonably) unbiased methods allowing to learn recursive concepts. Other techniques that could claim to do the same would be:

– Inductive Logic Programming: yet in existing systems, the user has to declare which predicates he wants to be seen called recursively, and even then the recursion has to be fairly simple.
– Neural Nets: a convincing comparison between these and formal grammar learning is still needed, but unlike automata learning the architecture of the neural network has to be given in advance (in the cases of the use of some long term memory, the length of this memory must also be given).
– Hidden Markov Models are models closely related to stochastic finite automata. Again, the number of states of the desired HMM is a usual parameter of the learning/training algorithm.

During the last few years one can point out the following new research directions and results[2]:

- Using the Kolmogorov Complexity framework: after the failure to obtain positive results in the PAC-framework, research under the simple PAC framework has been pursued: the idea is that the learning examples are drawn according to a simple distribution, *i.e.* one where the simplest examples have higher probability than the most complex ones. After proving that well known algorithms allow for the DFA class to be simple PAC learnable, the framework has permitted the introduction of learning algorithms for new classes such as the linear grammars.
- Adapting string learning algorithms to tree grammar learning.

---

[1] International Colloquium on Grammatical Inference

[2] These can be found in the ICGI proceedings, or in the Machine Learning Journal special issue on Grammatical Inference, to appear, or on the Grammatical Inference homepage [4].

- Finding new classes for which polynomial learning is possible: these can be either super classes of the regular languages when learning with counterexamples or subclasses when only positive examples are given.
- Building incremental Grammatical Inference algorithms.
- Using Genetic Programming, Taboo search, or Constraint Satisfaction techniques to find a solution. Once the combinatorics of the problem show that a greedy approach can only be successful in some cases, Artificial Intelligence heuristics can be of help.

3. Open Problems and Key Issues

Although in the last few (6 or 7) years a lot of novel results have been driven in the Grammatical Inference community, some crucial problems have either not been dealt with, or only in a yet very unsatisfactory way. Between these:

- The need of algorithms that can deal with noisy data: the usual benchmarks the community has been using in the late 90s were concerned with learning large (500 states) automata from positive and negative data. But in all cases this data has to be noise free: if one introduces even one incorrectly labeled string in the learning set for the top algorithms today, there is no chance of obtaining a correct solution, and it is plausible that the returned solution (if any) would be too large to be used. Several techniques have been used to deal with this problem:
  - Learning stochastic finite automata: the assumption here is not that the language is regular but that the distribution is. Several algorithms have been proposed since [9].
  - Learning non-deterministic finite automata. The problem is difficult and little has been done [11].
  - Reducing the problem to a graph coloring problem/constraint satisfaction problem. The idea is to use NP-problem solvers on the combinatorial task involved (one is really trying to find the smallest consistent grammar).
  - Using top down algorithms: nearly all the algorithms generalize a most specific grammar by state merging. It is well known that for a technique to be noise tolerant, working your way from a most general concept to a more specialized one is a better idea.
- The need to build algorithms that learn context free grammars. It can be proved that DFA are more or less the largest class that can be efficiently learned by provably converging algorithms [6]. Yet of course, at least for practical reasons the class is insufficient. A lot of work has been made, with many different ideas (genetic algorithms, taboo search, maximum likelihood approaches) into learning context-free grammars. Obviously the negative theoretical results imply that no formal proof of the validity of these ideas can be given. Nevertheless it is clear that learning context free grammars is a real challenge for the community. One of the first steps should be the acceptance of some common benchmark or task.
- Learn grammars and something else… Formal grammars have a nice generalization capacity, for only certain sorts of objects (in theory they can formalize a lot of things, but the size of a DFA representing a decision list, or some logical formula might be extravagant). Association of Grammatical Inference

techniques with other Machine Learning/ Syntactic Pattern Recognition procedures would be of mutual benefit. One could for instance:

- Use grammatical inference algorithms in Inductive Logic Programming tasks[3]
- Combine symbolic grammatical inference techniques with neural nets
- Combine Grammar Learning with other Machine Learning techniques, like learning decision trees and lists [6].

4. Acknowledgements

The author would like to thank Laurent Miclet for helping him understand some of the different links between Grammatical Inference and Structural Pattern Recognition.

References

1. Angluin, D: On the Complexity of Minimum Inference of Regular Sets. Information and Control 39 (1978) 337–350.
2. Bunke, H., Sanfeliu, A. (eds): Syntactic and Structural Pattern Recognition, Theory and Applications. Series in Computer Science 7, World Scientific, Singapore New Jersey London Hong Kong (1990).
3. Gold, M.E.: Language Identification in the Limit. Information and Control 10-5 (1967) 447-474.
4. de la Higuera, C., Parekh, R. The grammatical inference homepage: http://www.univ-st-etienne.fr/eurise/gi/gi.html and http://www.cs.iastate.edu/~honavar/gi/gi.html.
5. Honavar, V., Slutzki, G. (eds.): Grammatical Inference, *Proceedings of ICGI '98*. Lecture Notes in Artificial Intelligence Vol. 1433, Springer Verlag, Berlin Heidelberg New York (1998).
6. Miclet, L, de la Higuera, C. (eds.): Grammatical Inference: Learning Syntax from Sentences, *Proceedings of ICGI '96*. Lecture Notes in Artificial Intelligence Vol. 1147, Springer Verlag, Berlin Heidelberg New York (1996).
7. Lang, K., Pearlmutter, B.: the Abbadingo competition. http://abbadingo.cs.unm.edu/ (1997).
8. Lang, K., Pearlmutter, B., Coste, F.: the Gowachin Learning Competition http://www.irisa.fr/Gowachin/ (1998).
9. Oncina J., Carrasco, R. (eds): Grammatical Inference and Applications, *Proceedings of ICGI '94*. Lecture Notes in Artificial Intelligence Vol. 862, Springer Verlag, Berlin Heidelberg New York (1994).
10. Pitt, L.: Inductive Inference, DFA's, and Computational Complexity. In: Jantke, K. (ed): Analogical and Inductive Inference. Lecture Notes in Artificial Intelligence Vol. 397, Springer-Verlag, Berlin Heidelberg New York (1989) 18-44.
11. Sakakibara, Y.: Recent Advances of Grammatical Inference. Theoretical Computer Science 185 (1997) 15-45.

[3] Systems like Merlin and Gift use grammatical inference as the inference engine of logic programs: they do not combine GI with existing ILP systems.