
Machine Learning, 44, 5–7, 2001

Introduction

Grammatical Inference, variously refered to as automata induction, grammar induction, and automatic language acquisition, refers to the process of learning grammars and languages from data. Machine learning of grammars finds a variety of applications in syntactic pattern recognition, adaptive intelligent agents, diagnosis, computational biology, systems modelling, prediction, natural language acquisition, data mining and knowledge discovery.

Some factors that have contributed to increased activity in grammatical inference techniques over recent years include the following:

Grammatical inference rests upon strong mathematical foundations from formal language theory and inductive inference and hence researchers in grammatical inference are able to draw on a large body of theoretical results, formal models, and algorithms from these areas to develop grammatical inference algorithms with provable properties.

Grammatical inference is one of the few techniques in Machine Learning that provides for learning of functions defined over inputs of variable size: the length of the input strings, the size of the target grammar or automaton are not typically provided as parameters to a grammatical inference algorithm.

Grammatical inference, over the past six or seven years, has been able to extend its most natural techniques to learning from different types of data including: numerical data, trees or terms and under a variety of assumptions about the learner and the environment (e.g., when only positive examples are available).

Although much remains to be done in order to cope with noisy data, partially labelled data, and more generally, heterogeneous, unstructured data, grammatical inference algorithms are beginning to be successfully applied in a number of domains including: design of natural language interfaces to databases, information extraction from text, data-driven knowledge discovery in bioinformatics, and related areas.

This has helped foster a thriving community of grammatical inference researchers around the world. The interested reader can find pointers to people, research, publications, and events in this area on the grammatical inference web page at

http://www.cs.iastate.edu/∼honavar/gi/gi.html

which is also mirrored at

http://www.univ-st-etienne.fr/eurise/gi/gi.html.

Traditionally, grammatical inference has been studied by researchers in several research communities including: Information Theory, Formal Languages, Automata Theory, Language Acquisition, Computational Linguistics, Machine Learning, Pattern Recognition,

Computational Learning Theory, Neural Networks, etc. Over the past few years, several conferences (e.g., the International Colloquium on Grammatical Inference) and workshops have sought to bring together researchers working on grammatical inference in these areas.

Against this background, Tom Dietterich, the former editor-in-chief of the Machine Learning journal, invited us to guest edit a special issue of the journal to highlight some of the recent advances in this area. Approximately 28 papers were submitted in response to the call for papers that was issued in late 1998. Of these, 8 papers were selected for inclusion in the special issue on the basis of a rigorous peer review by at least 3 independent reviewers.

The papers included in this special issue cover a range of topics and approaches in grammatical inference. The papers by Parekh and Honavar and by Denis present new positive results on learnability of *simple* DFA and DFA from *simple* examples. These results suggest that variations of the PAC learning model, especially those that restrict the examples to be drawn from a *simple* distribution or those that impose a bias on the hypothesis space so as to favor *simple* hypotheses might lead to practical learning algorithms for function classes that have been proven to be hard to learn within the PAC model. Rossmanith and Zeugman describe a new learning model called *stochastic finite learning* and identify the entire class of pattern languages that are learnable within this model. Oliveira and Silva describe and algorithms and heuristics for the inference of minimum size DFA that is consistent with a labeled training set—a well-known NP-complete problem.

Pico and Casacuberta present statistical estimation techniques for infering stochastic finite-state transducers from examples. Amengual, Sanchis, Vidal, and Benedi present an approach to learning finite state language models from data that contains sentences that often do not strictly adhere to the formal rules of syntax (e.g., due to vocabulary variations, missing words, superfluous words, etc.). Giles, Lawrence, and Tsoi demonstrate an approach to time series prediction using recurrent neural networks for grammar inference from noisy financial time series data. Carrasco and Oncina present an algorithm for stochastic inference of regular tree languages. This result has implications for learning of context-free grammars.

While these results are indicative of some of the recent advances in grammatical inference, they are by no means exhaustive. Given the recent activity in this area, new results are appearing on a regular basis on a number of problems including:

Different models of grammar induction: e.g., learning from examples, learning using examples and queries, incremental versus non-incremental learning, distribution-free models of learning, learning under various distributional assumptions (e.g., simple distributions), impossibility results, complexity results, characterizations of representational and search biases of grammar induction algorithms.

Algorithms for induction of different classes of languages and automata: e.g., regular, context-free, and context-sensitive languages, interesting subsets of the above under additional syntactic constraints, tree and graph grammars, picture grammars, multi-dimensional grammars, attributed grammars, parameterized models, etc.

Theoretical and experimental analysis of different approaches to grammar induction including artificial neural networks, statistical methods, symbolic methods, information-theoretic approaches, minimum description length, and complexity-theoretic approaches, heuristic methods, among others.

Language acquisition e.g., acquisition of grammar in conjunction with language semantics, semantic constraints on grammars, language acquisition by situated agents and robots, neural models of language, acquisition of language constructs that describe objects and events in space and time, developmental and evolutionary constraints on language acquisition, etc.

Applications in bioinformatics (e.g., RNA and protein structure prediction, gene recognition) structural pattern recognition, information retrieval, information extraction, logic programming, database query processing and translation, text processing, adaptive intelligent agents, systems modelling and control, and other domains.

The interested reader is referred to proceedings of International Conference on Machine Learning, European Conference on Machine Learning, International Colloquium on Grammatical Inference, Pacific Symposium on Biocomputing, among others for a sampling of recent papers in grammatical inference.

We are grateful to all of the anonymous reviewers for their work. We would like to thank Tom Dietterich for inviting us to highlight some of the best current work in grammatical inference in the Machine Learning journal. We would like to thank Robert Holte, the current editor-in-chief of the Machine Learning journal for his patience and support. We are grateful to Ms. Karen Cullen, Ms. Melissa Sullivan, Melissa Fearon and other members of the editorial and production staff for their support.

It is our hope that this special issue will help foster more interaction among not only the various subfields of grammatical inference, but also between grammatical inference, machine learning, computational learning theory, and various application areas such as bioinformatics, natural language processing, information extraction, speech processing, among others.

Vasant Honavar
Iowa State University
http://www.cs.iastate.edu/∼honavar/

Colin de la Higuera
University Jean Monnet at St. Etienne
http://www.univ-st-etienne.fr/eurise/cdlh/cdlh.html