Artificial data and language theory

Franck Thollard THOLLARD@UNIV-ST-ETIENNE.FR
Laboratoire Hubert Curien (ex EURISE) - UMR CNRS 5516
18 rue du Professeur Lauras - 42000 Saint-Étienne Cedex 2 - France

Antoine Cornuéjols ANTOINE@LRI.FR
Mathématiques et Informatique Appliquées (MIA), UMR 518 (INA-PG, ENGREF, INRA)
16, rue Claude Bernard - 75005 Paris - France

Abstract

Tests on artificial data sets are important in order to systematically assess the merits and limits of learning algorithms. This requires that distributions over both the hypothesis space and the instance space be controlled, which, in turn, implies the definition of control parameters. This paper surveys the control parameters that have been used in the context of grammatical inference, and provides directions for further identification of relevant parameters.

1. Introduction

The assessment of the merits and limits of learning algorithms generally requires systematic and thorough empirical testing on representative data sets. In this context, the use of well-chosen artificial data sets allows one, in principle, to systematically control the relevant characteristics of the learning problem. In particular, this means to control the characteristics of the probability distributions over both the hypothesis and the instance spaces. While this usually presents little difficulty, the case of grammatical inference is specific in that, first, the input domain: the space of sequences, is made of objects of different sizes, and, second, and more importantly, the output domain: the space of finite state automata, is made of structured objects of varying complexity.

Special care is therefore required in designing the random generation of objects in the input or output domains. In particular, the random generation of grammars requires that key parameters defining the structure of interest be provided. So far, in the literature on grammatical inference, these parameters have mostly been supplied on a case by case basis. This paper first provides an overview of the various "solutions" that have been adopted and of some principled approaches described in the combinatoric algorithms field. It then points out that recent findings about a phase transition phenomenon in grammatical inference points to the need of novel parameters to describe grammars.

2. Control parameters over automata

Uniform, or non uniform, distribution over the space of structures of interest requires that a set of parameters be chosen, over which the generation process is controlled. Let θ be such a set of parameters upon which $\mathcal{L}_\theta$, the class of finite state machines, is defined. Two questions naturally arise: (i) what set of parameters θ is relevant for describing – some subclass of – the finite state machines? and (ii) how to design uniform (or not) random generation of elements of $\mathcal{L}_\theta$? We now review some answers to these questions.

2.1. Benchmarking in grammatical inference

One key parameter that appears in all studies is the *number of states* of the automata (see for instance, [4]). In addition, in the Abbadingo competition, the automata were designed with a small *depth* in order to prevent the data-sparseness problem. The number of transitions is also usually considered when the automata are non-deterministic.

Most experiments on artificial data use small automata (namely around 20 states). This is either because demonstrating some features of the algorithm is the main goal [6, 3] or because the proposed approaches do not scale up. Moreover, [1] conclude that "automata reduction makes sense only if one is manipulating automata with a high level of redundancy". This shows that redundancy should also be considered as a descriptive feature, or as a control parameter.

2.2. Principled approaches

In recent years, the general problem of studying and simulating random process has particularly benefitted from progresses in the area of random generation of combinatorial structures. The seminal works of Wilf and Nijenhuis in the late 70's [5] have led to efficient algorithms for generating uniformly at random a variety of combinatorial structures. In 1994, Flajolet, Zimmermann and Van Cutsem [2] have widely generalized and systematized this work. Briefly, their approach is based on a recursive decomposition of the combinatorial structures to be generated. This constitutes the basis for powerful tools for random generation of complex entities, such as graphs, trees, words, paths, etc. The application of these tools to the random generation of grammars is natural.

However, as noted before, having a uniform random sampling of n nodes graphs does not guarantees uniform random sampling of the automata since the problem of the useless states remains.

3. A phase transition and its meaning

In [7], the variation of the coverage rate of the automata generated during state merging operations has been systematically investigated. Learning and test sequences were drawn according to a uniform distribution with respect to their length, meaning that there was an equal probability of drawing strings of any length ℓ up to a maximum value ℓ_{Max}. The parameters governing the random generation of target automata (deterministic or not) were: the number Q of states in the DFA, the number B of output edges on each state, the number L of letters on each edge, and the fraction a of accepting states, taken in [0,1].

It was then observed that the coverage rate of the automata generated by state merging algorithms undergoes a spectacular and sudden transition from low coverage hypotheses (automata) to high coverage ones, particularly in the DFA case.

One immediate conclusion is that induction algorithms based on state merging operators are up to a fundamental problem. But another conclusion emerges viz. that the current set of parameters chosen to describe automata is inadequate to account for the generalization capacity of the automata, and that *new parameters, more apt at describing their structural properties, need to be identified*. If such was the case, there is hope that this would point out new learning operators that would permit a better sampling of the hypothesis space and, therefore, to escape the phase transition phenomenon that cripple current learning algorithms.

4. Summary and further work

The following parameters have been found useful in designing artificial learning tasks.

- # states, # alphabet [all]
- automata density (or redundancy) : $\bar{\Gamma}^+$, $\frac{\bar{\Gamma}^+}{|\Sigma|}$ [1]
- automata depth [4]
- Symmetric difference with a given language (e.g. $L = \{x \in \Sigma^\star : |x| = n\}$).

Specifically in the case of probabilistic grammars, the following features might be useful control parameters:

- $\epsilon-$dispersion (minimal number of string s.t. their sum equals $1 - \epsilon$),
- average and/or variance of the probabilities.
- average length of the strings.
- Distance / Similarity w.r.t. another (probabilistic) language. Similarity measure could be the Kullback-Leibler or the $L_d, d \in [1..\infty]$.
- Weight of the language with respect to a non probabilistic language $L : \sum_{x \in L} P(x)$. On the contrary to the non probabilistic case, L cannot be $\Sigma^\star$ as by consistency the weight of the automaton in $\Sigma^\star$ is 1.

Overall, there is a need for a better understanding of the various parameters that could usefully describe learning tasks in grammatical inference.

References

[1] J.-M Champarnaud and F. Coulon. Nfa reduction algorithms by means of regular inequalities. *Theoret. Comput. Sci.*, 327(3):241 – 253, 2004.

[2] Philippe Flajolet, Paul Zimmermann, and Bernard Van Cutsem. A calculus for the random generation of labelled combinatorial structures. *Theor. Comput. Sci.*, 132(2):1–35, 1994.

[3] A. Habrard, F. Denis, and Y. Esposito. Using pseudo-stochastic rational languages in probabilistic grammatical inference. In *ICGI*, 2006.

[4] Kevin J. Lang, Barak A. Pearlmutter, and Rodney A. Price. Results of the abbadingo one DFA learning competition and a new evidence-driven state merging algorithm. *LNCS*, 1433, 1998.

[5] A. Nijenhuis and H. Wilf. *Combinatorial algorithms*. Academic Press, 1979.

[6] A. Oliveira and J. Silva. Efficient search techniques for the inference of minimum size finite automata. In ICML Workshop, editor, *ICML*, Nashville, 1997.

[7] Nicolas Pernot, Antoine Cornuéjols, and Michèle Sebag. Phase transitions within grammatical inference. In *IJCAI*, pages 811–816, 2005.