New Polynomial Bounds for the Identification in the Limit Paradigm using Generative Grammars

Rémi Eyraud[1] and Jeffrey Heinz[2]

[1] QARMA team, Laboratoire d'Informatique Fondamentale, Marseilles, France
[2] Dept. of Linguistics and Cognitive Science, University of Delaware, Newark, USA

Complexity analysis in computer science examines worst-case scenarios. If an algorithm solves a problem, are there instances of the problem space that require the algorithm to consume an extraordinary amount of resources before outputting their solutions?

This work deals with the open question about how to define learning paradigms in order to best analyze the complexity of learning algorithms, in the particular case of algorithms that aim at learning formal languages. For reasons that are mostly due to the type of models infered, the formalization of identification in the limit [3] is often used in this field. Several refinements of Gold's first formalization have been defined with the aim to introduce efficiency bounds to strengthen learning results obtained in this identification in the limit framework. A widely used one is named identification with polynomial time and data [2], where the algorithm updates its hypothesis in a time polynomial in the size of the available data, and is required to admit a characteristic sample whose size is polynomial in the size of the learning target. (A sample S is characteristic if the algorithm is ensured to converge to a representation of the learning target on any sample containing S).

However, there exist classes of languages identifiable in the limit by algorithms which return representations that are exponentially smaller than their characteristic samples. There is an intuition that these algorithms, by finding correct, compact representations of observed data are doing something right, and that a definition of the complexity of learning algorithms should view such algorithms favorably.

This work proposes a modified paradigm. We define *structurally complete sets* and instead consider the property that a characteristic sample of a language L be polynomial in the size of a minimal structurally complete set for L. While the previous refinement is not centered on the target language but on its selected representation, which is an important shift in perspective, this new formalization focuses the attention on the strings you can obtain using the target representation.

Formally, the newly designed paradigm is defined as follows:

Definition (Polynomial Structurally Complete Identification). A class $\mathbb{L}$ of languages is *identifiable in polynomial time and structurally complete data* for a class $\mathbb{R}$ of representations if and only if there exist an algorithm $\mathfrak{A}$ and two polynomials $p()$ and $q()$ such that:

1. Given a sample S for $L \in \mathbb{L}$ of size m, $\mathfrak{A}$ returns a hypothesis $H \in \mathbb{R}$ in $\mathcal{O}(p(m))$ time ;
2. For each representation R of a language $L \in \mathbb{L}$, there exists a *characteristic sample* CS whose size is in $\mathcal{O}(q(k))$, where k is the size of the smallest structurally complete set for R.

Where a characteristic sample is a set of data such that on any sample that contains it, the algorithm outputs an hypothesis that is correct. The notion of structurally complete set is the following:

Definition (Structually Complete Set). Given a generative grammar G, a structurally complete set (SCS) for G is a set of data SC such that for each production $\alpha \to \beta$, there exists an element $x \in SC$, an element $\gamma \in I$ and two elements $\eta, \tau \in (\Sigma \cup N)^*$ such that $\gamma \Rightarrow^* \eta\alpha\tau \Rightarrow \eta\beta\tau \Rightarrow^* x$.

Where a generative grammar is a device defined by a tuple $< \Sigma, N, P, I >$ where Σ is the alphabet of the language, N is a set of variables usually called non-terminals, $P \subset (N \cup \Sigma)^* \times (N \cup \Sigma)^*$ is the set of generative rules, I is the finite set of axioms, which are elements of $(\Sigma \cup N)^*$. The language it represents is the strings over Σ that can be derived from an element of I using the rules of P ($\Rightarrow$ is the derivation relation, and $\Rightarrow^*$ its symetric and transitive closure).

We conjecture that most algorithms whose characteristic sample is known to be polynomial in the size of the target representation will also admit a characteristic sample polynomial in the size of the smallest structurally complete set. This is the case for instance for the well-known $RPNI$ algorithm [4] that learns regular languages from positive and negative examples. Moreover, some algorithms whose characteristic sample are not polynomial in the size of the grammar can admit a characteristic sample that is polynomial in the size of the smallest structurally complete set. This is the case for instance of the algorithm SGL that learns the substitutable context-free languages [1].

References

1. A. Clark and R. Eyraud. Polynomial identification in the limit of substitutable context-free languages. *Journal of Machine Learning Research*, 8:1725–1745, 2007.
2. C. de la Higuera. Characteristic sets for polynomial grammatical inference. *Machine Learning Journal*, 27:125–138, 1997.
3. E. M. Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967.
4. J. Oncina and P. García. Identifying regular languages in polynomial time. In *Advances in Structural and Syntactic Pattern Recognition*, volume 5 of *Series in Machine Perception and Artificial Intelligence*, pages 99–108. 1992.