Theoretical Computer Science 620 (2016) 46–72

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Probabilistic learnability of context-free grammars with basic distributional properties from positive examples

Chihiro Shibata [a,*], Ryo Yoshinaka [b,*]

[a] School of Computer Science, Tokyo University of Technology, Japan
[b] Graduate School of Informatics, Kyoto University, Japan

ARTICLE INFO

Article history:
Available online 30 October 2015

Keywords:
Grammatical inference
PAC learning
Distributional learning

ABSTRACT

In recent years different interesting subclasses of CFLs have been found to be learnable by techniques generically called *distributional learning*. The theoretical study on the exact learning of CFLs by those techniques under different learning schemes is now quite mature. On the other hand, positive results on the PAC learnability of CFLs are rather limited and quite weak. This paper shows that several subclasses of context-free languages that are known to be exactly learnable with positive data and membership queries by distributional learning techniques are PAC learnable from positive data under some assumptions on the string distribution.

1. Introduction

In the domain of grammatical inference, which studies mainly theoretical learnability of formal languages, to find an interesting subclass of context-free languages (CFLs) that can provably be learned efficiently is an important topic for several application fields including natural language processing, bioinformatics, and so on. We had few positive results on this topic and it was thought to be quite challenging for decades, before various rich subclasses of CFLs have been shown to be learnable by a series of techniques based on *distributional learning*. The idea of distributional learning dates back to Harris [12], yet it happened only in recent years that many concrete distributional learning algorithms have been proposed, analyzed and compared. The notion of "distribution" in this context does not refer to stochastic distributions, but rather the relation between substrings and contexts under the concerned language. Every string w over an alphabet Σ can be decomposed into a substring $u \in \Sigma^*$ and a context $(l, r) \in \Sigma^* \times \Sigma^*$ so that $w = lur$. The theory of distributional learning is concerned with the question which substrings and which contexts can form a grammatical member of the concerned language. The language can be learned if it fulfills certain distributional properties, which may allow a learner to generalize the observed finite information on the relation between strings and contexts. A typical and probably the simplest distributional property proposed so far, is the *substitutability*. Clark and Eyraud [7], in their seminal work of recent development of distributional learning, have shown that substitutable context-free grammars (CFGs) are identifiable in the limit from positive data in polynomial time. The substitutability is the property of a language that whenever two substrings occur in the same context, they are always substitutable for each other in any other contexts as well. Following their work, Yoshinaka [21] has generalized the notion to k, l*-substitutability* and proved the learnability. Other distributional properties allow different learning algorithms under

* Corresponding authors.
E-mail addresses: shibatachh@stf.teu.ac.jp (C. Shibata), ry@i.kyoto-u.ac.jp (R. Yoshinaka).

http://dx.doi.org/10.1016/j.tcs.2015.10.037
0304-3975/

different learning schemes. Shirakawa and Yokomori's [18] and Clark's [4] algorithms exactly learn *c-deterministic* and *congruential* CFLs under Angluin's MAT model [1], respectively. CFLs with the *k-finite kernel property* (*k*-FKP) and with the *k-finite context property* (*k*-FCP) can be learned from positive data and membership queries [8,5,23]. Those distributionally learnable classes have been comprehensively studied and classified. The approaches taken for learning congruential and *k*-FKP CFGs are called *primal*, where nonterminal symbols of a grammar are represented by a finite number of substrings, and those for c-deterministic and *k*-FCP CFGs are called *dual*, where the language of a nonterminal is characterized by a finite number of contexts. They show a neat symmetry, which allows a uniform description for those algorithms [23,24]. Furthermore, it has been revealed that those techniques can be applied to the learning of extensions of CFLs (e.g., [22,11]). The current state of the art on the distributional learning techniques for the deterministic exact learning of (extensions of) CFLs seems quite mature.

However, the learning schemes for which those results are established are often criticized because they are too abstract and impractical. Although those algorithms that identify languages in the limit will conjecture a correct grammar on a good sample set, which can be of reasonable size, one cannot know if the data set given to a learner is good enough even though a lot of data are given. That is, one cannot trust in the learner's output. In addition, a teacher who answers queries is often not available.

Probabilistic learning is in general regarded more practical. Valiant's probably approximately correct (PAC) learning [20] is another classical learning scheme used in grammatical inference. In this learning scheme, a learner receives strings from an arbitrary distribution on Σ^* with labels indicating whether respective strings belong to the learning target or not, and then it is required to output with high probability a grammar whose language is a good approximation of the learning target with respect to the distribution. While his original definition of the learning scheme still seems too strict, as Kearns and Valiant [13] showed that even regular languages are not PAC learnable under a widely believed assumption in the theory of cryptography, many variants of the PAC learning have been proposed. Among those, Clark and Thollard [9] proposed to restrict probabilistic distributions from which examples are drawn to the ones defined based on representations of target languages. Namely, they showed that regular languages are PAC learnable from positive examples which are drawn from the probabilistic distribution determined by a probabilistic finite state automaton.

Their idea was applied to CFL learning as well. Clark [3] has discussed the PAC learnability of unambiguous nonterminal separable (NTS) languages from positive examples where the distributions are determined by probabilistic CFGs generating the target languages. And then Luque and Lopez [14] generalized the arguments to the learning of unambiguous k,l-NTS languages. Still, compared to the solid and intensive study of the deterministic exact distributional learning of various CFL classes, we have few probabilistic counterparts of them. Particularly unambiguous (k,l-)NTS languages are not quite expressive: In fact, some simple finite languages and infinite languages like a^+ are not (k,l-)NTS.

The goal of this paper is to expand Clark's [3] result on unambiguous NTS CFGs to other subclasses of CFGs, including ambiguous ones, that are known to be deterministically exactly learnable based on distributional learning techniques. We translate existing distributional exact learning algorithms into PAC-type ones where a learner gets positive examples drawn from the distribution determined by a probabilistic CFG and no negative examples. We show that under some assumptions on the distribution, membership queries used in the existing deterministic exact learners can be "simulated" by statistic observation of positive examples. Consequently, our target language classes contain every regular language. Moreover, we show that additional knowledge makes those classes probably *exactly* learnable. As far as the authors know, this is the first PAC-type learning result for a class of ambiguous grammars.

This paper is organized as follows. After the preliminaries following this introduction, we will introduce three learning algorithms that PAC learn CFGs with different basic distributional properties in Sections 3 to 5. The first target is unambiguous c-deterministic linear CFGs, which are a very modest yet proper extension of deterministic finite automata (DFAs). The second one is linear CFGs with the 1-FCP, which may be ambiguous. The third one is CFGs with the 1-FKP. While the former two are linear grammars learned by dual approaches, the last one is nonlinear grammars learned by a primal approach. In Section 6 we compare Clark's [3] PAC learnability result on unambiguous NTS languages with ours. We conclude the paper in Section 7. Technical proofs on probabilistic theory and properties of examples are relegated to the appendices.

2. Preliminaries

This section prepares definitions, notation and technical lemmas which will be used in this paper. Technical lemmas in probability theory are presented in Section 2.1. We then give definitions and notation on formal languages and probabilistic languages in Section 2.2. Section 2.3 is devoted to probabilistic grammars and related concepts.

2.1. Probabilities

A *distribution* over a countable domain U is a map $D : U \to [0,1]$ such that $\sum_{x\in U} D(x) = 1$. For a subset $U' \subseteq U$, we define $D(U') = \sum_{x\in U'} D(x)$. We denote the *support* by $U_D = \{x \in U \mid D(x) > 0\}$. Clearly $D(U_D) = 1$. The *infinity norm* is $\|D\|_\infty = \sup_{x\in U} D(x)$. When comparing two distributions D_1 and D_2, we define $\|D_1 - D_2\|_\infty = \sup\{|D_1(x) - D_2(x)| \mid x \in U\}$.

We remind the reader of the following basics of statistical theory.