INFERRING PURE CONTEXT-FREE LANGUAGES FROM POSITIVE DATA

Takeshi Koshiba, Erkki Mäkinen and Yuji Takada

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF TAMPERE

REPORT A-1997-14

UNIVERSITY OF TAMPERE
DEPARTMENT OF COMPUTER SCIENCE
SERIES OF PUBLICATIONS A
A-1997-14, DECEMBER 1997

INFERRING PURE CONTEXT-FREE LANGUAGES FROM POSITIVE DATA

Takeshi Koshiba, Erkki Mäkinen and Yuji Takada

University of Tampere
Department of Computer Science
P.O.Box 607
FIN-33101 Tampere, Finland

ISBN 951-44-4269-5
ISSN 0783-6910

Inferring pure context-free languages from positive data

Takeshi Koshiba

High Performance Computing Research Center, Fujitsu Laboratories Ltd., 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki-shi, Kanagawa 211, Japan
(koshiba@flab.fujitsu.co.jp)

Erkki Mäkinen

Department of Computer Science, University of Tampere, P.O. Box 607, FIN-33101 Tampere, Finland (em@cs.uta.fi)

Yuji Takada

Personal Systems Labs., Fujitsu Laboratories Ltd., 2-2-1 Momochihama, Sawara-ku, Fukuoka 814, Japan (yuji@flab.fujitsu.co.jp)

Abstract

We study the possibilities to infer pure context-free langauges from positive data. We can show that while the whole class of pure context-free languages is not inferable from positive data, it has interesting subclasses which have the desired inference property. We study uniform pure languages, i.e., languages generated by pure grammars obeying restrictions on the length of the right hand sides of their productions, and pure languages generated by deterministic pure grammars.

Keywords: grammatical inference, identification in the limit, pure languages, deterministic pure langauges, uniform pure langauges.

1 Introduction

In pure grammars, no distinction is made between terminals and nonterminals. It follows that the generative capacity of pure grammars is much weaker than that of corresponding Chomsky type grammars. It is argued [5,14] that the custom of dividing the alphabet of a grammar originates from the linguistic background of formal language theory and in fact, it would be more natural to study rewriting systems that do not make difference between terminals and nonterminals.

In this paper we study the possibilities to infer pure languages from posivite data. The well-known negative result by Gold [9] says that regular languages cannot be inferred from positive data only. This negative result has initiated a search for language classes having the desirable inference property. The found subclasses include, among others, 1-variable pattern languages [1], paranthesis languages [6], locally testable languages [8], deterministic even linear languages [12], and k-reversible languages [3]. Even more closely related to the present paper is Yokomori's [18] result concerning the inferability of PD0L languages from positive data and especially Tanida and Yokomori's [16] results on the inferability of monogenic pure langauges.

We show here that while the whole class of pure context-free languages is not inferable from positive data, it has interesting subclasses which have the desired inference property. The subclasses are defined by restricting the length of the right hand sides in the productions (uniform pure languages) or the number of productions (detrministic pure grammars).

The fact that the whole class of pure context-free langauges is not inferable from positive data only is earlier shown by Tanida and Yokomori [16].

2 Preliminaries

We assume a familiarity with the basics of formal language theory and grammatical inference as given e.g. in [11] and [4], respectively. As inference criterion we use "identification in the limit" [9,4].

If not otherwise stated we follow the notations and definitions of [11]. The length of a string w is denoted by $lg(w)$. A production in a (Chomsky-type) context-free grammar is said to be *terminating* if the right hand side contains no nonterminals. Otherwise, a production is said to be *continuing.*

We now define pure grammars and languages. A *pure context-free grammar* is a system $G = (\Sigma, P, s)$, where Σ is a finite alphabet, P is a finite set of productions of the form $\alpha \to \beta$, where α is in Σ and β is a word over Σ. For the sake of simplicity we assume that the empty word λ is not allowed as a right hand side of any production. Contrary to most earlier articles on pure grammars (cf. e.g. [7,14]), we suppose that the *axiom* s is a single word over Σ. Relation $\Rightarrow$ (yields directly) and its reflexive transive closure $\Rightarrow^*$ are defined in Σ^* as usual. The language generated by a system $G = (\Sigma, P, s)$ is defined as

$$L(G) = \{w \mid s \Rightarrow^* w\}.$$

A language is a *pure context-free language* if it can be generated by a pure context-free grammar. The class on pure context-free languages is denoted by $\mathcal{P}$. Note that $\mathcal{P}$ and the class of regular languages are incomparable.

We consider here pure context-free grammars and languages only. We hereafter omit the phrase "context-free", and simply talk about pure grammars and pure languages.

A pure grammar G is *monogenic* if, whenever w is in $L(G)$ and $w \Rightarrow w'$, then there are unique words w_1 and w_2 such that $w = w_1 x w_2$, $w' = w_1 y w_2$, and $x \rightarrow y$ is a production.

A pure grammar G is *deterministic* if, for each symbol a, there is at most one production with a on the left hand side. A pure language is deterministic if there exists a deterministic pure languages generating it. We denote the class of deterministic pure languages by $\mathcal{D}$.

A pure grammar G is *reduced* if every symbol appear in some word of $L(G)$. If a reduced pure grammar is monogenic then it is also deterministic. On the other hand, a deterministic pure grammar is not necessarily monogenic [14].

An *indexed family of nonempty recursive languages* is an infinite sequence $L_1, L_2, L_3, \ldots$, where each L_i is a nonempty language with decidable membership problem. The following two well-known results by Angluin [2] are essential for our further discussion.

Theorem 1 *([2]) If an indexed family of nonempty recursive languages is inferable from positive data, then there exists, on any input $i, i \geq 1$, a finite set of strings T_i such that*

(i) $T_i \subseteq L_i$, and
(ii) for all $j \geq 1$, if $T_i \subseteq L_j$, then L_j is is not a proper subset of L_i.

Let $\mathcal{L}$ be an indexed family of nonempty recursive languages. We say that $\mathcal{L}$ has *finite thickness*, if for each nonempty finite set $S \subseteq \Sigma^*$, the set $C(S) = \{L \mid S \subseteq L$ and $L = L_i$ for some $i\}$ is of finite cardinality.

Theorem 2 *([2]) If an indexed family of nonempty recursive languages has finite thickness, then it is inferable from positive data only.*

Note that thickness is not defined in terms of the number of representations (generating systems), but in terms of the number of languages.

3 A negative result

As the class of languages inferable from positive data only is known to be quite restricted, it is to be expected that $\mathcal{P}$ in not inferable from positive data. To prove this we can follow Yokomori's corresponding proof [18](Thm. 3) for propagating 0L-systems. A different proof is given in [16].

Theorem 3 *$\mathcal{P}$ is not inferable from positive data only.*

Proof. We derive a contradiction with Theorem 1.

Consider the language $L = \{b\} \cup \{a^n \mid n \geq 2\}$. L is in $\mathcal{P}$, since it can generated from axiom b with productions $b \to aa$ and $a \to aa$.

Let T be any nonempty finite subset of L, and let $T' = T \setminus \{b\}$. Further, let $T' = \{a^{n_1}, \ldots, a^{n_p}\}$.

Consider a pure grammar H with axiom b and with productions

$$\{b \to a^{n_1}, \ldots, b \to a^{n_p}\}.$$

We have $T \subseteq L(H) \subset L$ contradicting Theorem 1.

Remark 4 *The proof of Theorem 3 shows why we do not allow an arbitrary set of axioms but a single axiom string. If an arbitrary set of strings were possible as an axiom, then Theorem 3 would hold also for all reasonable defined subclasses of pure grammars. Namely, we could choose T as the axiom set, and we would not even need any productions to show that the condition of Theorem 1 does not hold.*

4 k-uniform pure grammars

We say that a pure grammar $G = (\Sigma, P, s)$ is *k-uniform*, $k > 1$, if each production $\alpha \to \beta$ in P has $lg(\beta) = k$. A pure language L is k-uniform if there exists a k-uniform pure grammar generating L. The class of k-uniform pure languages is denoted by $\mathcal{P}(k)$.

The property of a pure grammar being k-uniform has its implications to the length set of the langauge generated. (The *length set* of a language L is defined by $LS(L) = \{lg(w) \mid w \in L\}$.) Namely, the length of the axiom and the constant k together uniquely defines the length set.

It also follows directly that $\mathcal{P}(i)$ and $\mathcal{P}(j)$, $i \neq j$, cannot have any infinite language in common. Moreover, the union $\bigcup_{i>1}\mathcal{P}(i)$ of k-uniform pure languages is clearly a proper subset of $\mathcal{P}$. These remarks show that the classes of k-uniform pure languages are quite restricted. On the other, each of the classes $\mathcal{P}(i)$, $i \geq 2$, contains non-regular languages. A simple example in the case $k = 3$, is

$$G_1 = (\{a, b, c\}, \{c \to acb\}, abc)$$

with $L(G_1) = \{a^n c b^n \mid n \geq 1\}$.

Hagauer [10] has shown that also $\mathcal{P}(2)$ contains non-regular languages. Namely, he has shown that

$$G_2 = (\{a, b, c\}, \{a \to ab, b \to bc, c \to ca\}, a)$$

produces a non-regular langauge.

Theorem 5 *$\mathcal{P}(k)$, $k \geq 2$, is inferable from positive data only.*

Proof. We show that $\mathcal{P}(k)$ has finite thickness, and hence, by Theorem 2 is inferable from positive data only.

Given any set S, the length of the shortest word in S gives an upper bound to the length of the axiom. Similarly, the cardinality of Σ (the alphabet considered) gives an upper bound for the number of productions having exactly k symbols in their right hand sides. Thus, $\mathcal{P}(k)$ has finite thickness.

By letting $\mathcal{Q}(n) = \mathcal{P}(2) \cup \mathcal{P}(3) \cup \ldots \cup \mathcal{P}(n)$, where n is any natural number, we can clearly prove also the following

Theorem 6 *$\mathcal{Q}(n)$ is inferable from positive data only.*

We can continue further to this direction, and define a pure grammar G to be *length-bounded* if there exists a natural number k such that the length of any right hand side in G's productions is at most k. A pure language L is length-bounded if there exist a length-bounded pure grammar G such that $L(G) = L$.

Theorem 7 *Length-bounded pure languages are inferable from positive data only.*

Proof. Analogously to the proof of Theorem 5.

The class $\mathcal{P}(2)$ is somewhat related to the class of uniquely terminating regular languages which is known to be infereable from positive data [13].

A (Chomsky type) regular grammar $G = (V, S, P, S)$ is *uniquely terminating* if the productions in P fulfil the following conditions for each nonterminal A in G:

(i) $A \to aB$ and $A \to aC$ imply $B = C$;
(ii) A has a unique terminating production; i.e. each nonterminal has exactly one terminating production. The terminals appearing in the right hand sides of terminating productions are all different.

A regular language L is uniquely terminating if there exists a uniquely terminating regular grammar generating L. Uniquely terminating languages are inferable from positive data [13].

Each uniquely terminating regular language is a member of $\mathcal{P}(2)$ provided that there are no terminals appearing both in terminating and in continuing productions. Let $G = (V, S, P, S)$ be a uniquely terminating regular grammar. The corresponding 2-uniform pure grammar H can be generated as follows. If $S \to a$ is the unique terminating production for the start symbol S of G, then a is the axiom of H. If $A \to bB$ is a continuing productions in G and the unique terminating productions for A and B are $A \to c$ and $B \to d$. Then H has the production $c \to bd$. Other productions are not needed.

The additional requirement that no terminal can appear in productions of both type characterizes well the difference between Chomsky type grammars and pure grammars. If the requirement does not hold, then the above construction ends up with a pure 2-uniform grammar which may produce words not in the original Chomsky language.

5 Inferring deterministic pure languages

Tanida and Yokomori [16] have shown that monogenic pure languages are inferable from positive data only. Their inference algorithm updates its conjectures in time $O(N^3)$ where N is the total length of the positive samples presented.

We shall now study the inferability of deterministic pure languages. Recall that reduced monogenic pure grammars are always deterministic, but deterministic pure grammars are not necessarily monogenic.

In order to prove that deterministic pure languages are inferable from positive data, we need the concept of finite elasticity from [17,15].

A class $\mathcal{C}$ of languages has *infinite elasticity* if and only if there is an infinite sequence $w_0, w_1, w_2, \ldots$ of strings and an infinite sequence $L_1, L_2, L_3, \ldots$ of languages from $\mathcal{C}$ such that for all $n \geq 1$, $\{w_0, w_1, \ldots, w_{n-1}\} \subseteq L_n$ but $w_n \notin L_n$. If a class $\mathcal{C}$ does not have infinite elasticity, then it has *finite elasticity.*

Notice that in the above definition both the languages $L_1, L_2, L_3, \ldots$ and the strings $w_0, w_1, w_2, \ldots$ are pairwise disjoint, i.e. each language (resp. string) appears at most once in the sequence $L_1, L_2, L_3, \ldots$ (resp. $w_0, w_1, w_2, \ldots$).

Theorem 8 *[17,15] If a class $\mathcal{C}$ of langauges has finite elasticity, then $\mathcal{C}$ is inferable from positive data only.*

We can now show that $\mathcal{D}$ has finite elasticity, and hence, it is inferable from positive data only.

Theorem 9 *$\mathcal{D}$ is inferable from positive data only.*

Proof. To derive a contradiction suppose that $\mathcal{D}$ has infinite elasticity. Let $w_0, w_1, w_2, \ldots$ be a sequence of strings required in the definition of finite elasticity, and let $L_1, L_2, L_3, \ldots$ be the corresponding sequence of deterministic pure languages.

Consider, for some $n > 1$, the subset $W_{n-1} = \{w_0, w_1, \ldots, w_{n-1}\}$ and the language L_n such that $W_{n-1} \subseteq L_n$ and $w_n \notin L_n$. Let $G_n = (\Sigma, P_n, s_n)$ be a deterministic pure grammar generating L_n. Since we do not allow productions of the form $a \to \lambda$, the length of s_n is bounded by the minimum length of strings in W_{n-1}. Hence, there are only a finite number of possible axioms in grammars $G_1, G_2, G_3, \ldots$.

For at least one axiom s there exist an infinite number of grammars using this axiom. These grammars have a (growing) subset of common strings. On the other, the number of productions in each G_i is bounded by the cardinality of Σ since we consider deterministic pure languages. Clearly, such an infinite sequence of deterministic pure grammars (and languages) with a bounded number of productions cannot exist. Thus, $\mathcal{D}$ cannot have infinite elasticity, and it is inferable from positive data only.

We end this section by discussing pure grammars and languages which are both deterministic and k-uniform. The class of such languages is denoted by $\mathcal{D}(k)$, $k \geq 2$.

Given k and the alphabet Σ, there are only a finite number of possible production sets for a k-uniform, deterministic pure grammar. Let $\mid \Sigma \mid$ stand for the cardinality of Σ. For each a in Σ, there is at most one production with a

in the left hand side. The number of possible right hand sides is $| \Sigma |^k$. Hence, there are only $(| \Sigma |^k + 1)^{|\Sigma|} - 1$ possible sets of productions. Here k and $| \Sigma |$ can be considered as constants. This leaves us with the problem of finding the proper axiom.

The "proper" axiom is, of course, the longest word over Σ having the property that all sample words so far received can be generated from it by using the production set in question. Since the number of possible production sets is indeed a constant, we can suppose that we know the correct production set. Repeating the procedure of searching the axiom for each possible production set naturally increases the constant coefficient of the time complexity, but it does not effect to the asymptotic growth rate.

Suppose now that the sample contains two words $a_1 a_2 \ldots a_m$ and $b_1 b_2 \ldots b_n$. Given the set of productions, what is the longest axiom from which the two words can be generated? A straightforward approach is to step backwards from the words according to the given productions until a common predecessor is found. Hence, we find out all the matches of the right hand sides of the given productions in $a_1 a_2 \ldots a_m$ and $b_1 b_2 \ldots b_n$, replace the occurrences of the right hand sides with the corresponding left hand sides, and store the words so obtained in data structures T_a and T_b, respectively. This is repeated until T_a and T_b contain a common word, the longest possible axiom.

A concise data structure for representing T_a is an automaton which accepts the possible axioms. Such an automaton A can be defined as $A = (Q, (0,0), F, \delta)$, where $Q = \{(i,j) \mid i = 0, \ldots, n, j = 0, ..., n-1\}$ is the set of states, $(0,0)$ is the initial state, $F = \{(n,j) \mid j = 0, \ldots, n-1\}$ is the set of final states, and the transition relation δ is recursively defined as follows:

(i) for each $i = 0, ..., n-1$ and for each $j = 0, ..., n-1$, $\delta((i,j), a_{j+1}) \ni (i, j+1)$
(ii) if $\delta((i,j), a) \ni (i', j')$, $\delta((i', j'), b) \ni (i'', j'')$, $j'' < n-1$ and $c \to ab$ is a production, then $\delta((i,j), c) \ni (i'', j''+1)$.

Note that the time need for constructing this automaton representation is bouded by a polynomial in n.

The longest possible axiom is not necessarily unique. When a new sample word is received and the conjecture is to be updated, we represent the old sample words by the set of all possible axioms, and repeat the above procedure for finding the new axiom.

As an example, consider $(ab)^n c$ as the input word. Let $a \to ba$, $b \to ab$, $c \to ab$ be productions. It is easy to see that each word in $\{b, c\}^n c$ is a possible axiom. Hence, the number of possible axioms can be exponential in n.

We pose it as an open problem whether or not there exists a polynomial time inference algorithm for $\mathcal{D}(k)$ using positive data only. On the spirit of the previous discussion, the polynomial time inference algorithm would need an efficient method for constructing the intersection of two languages acceptable by automata of the type defined above.

However, we have an affirmative answer in a special case. Namely, if the length of the axiom is bounded by a constant, then deterministic, k-uniform pure languages are polynomial time inferable from positive data only.

Moreover, if the length of the axiom is bounded, then we even have the following stronger result. Let d be a fixed integer and $\mathcal{D}_d(k)$ be the class of languages generated by pure deterministic k-uniform grammars whose axioms are of length at most d. We set $\mathcal{D}_d = \bigcup_{i=2}^{\infty} \mathcal{D}_d(i)$.

Theorem 10 *$\mathcal{D}_d$ is polynomial time inferable from positive data only.*

Proof. We know that $\mathcal{D}_d(k)$ is inferable from positive data only for any fixed k. We need only to infer the value of k. For each L in $\mathcal{D}_\lceil$, there exist integers c_1 and $c_2 \leq d$ such that $LS(L) = \{c_1 \cdot n + c_2 \mid n \geq 0\}$. To infer the value of k, we need only to calculate the minimum absolute value of $lg(w_1) - lg(w_2)$ over any two words of different length presented so far. Moreover, k is at most $O(\log N)$, where N is the total length of the positive samples presented.

6 Concluding remarks

Pure (context-free) languages are not inferable from positive data. However, natural subclasses of pure languages obtained by restricting the length of the right hand sides in the productions or the number of productions are inferable from positive data or the number of productions. We have shown the existence of such inference algorithms for k-uniform pure languages and for deterministic pure langauges. Moreover, we have posed open whether there exists a polynomial time inference algorithm for deterministic, k-uniform pure languages using positive data only.

References

[1] D. Angluin, Finding patterns common to a string, *J. Comput. Syst. Sci.* **21** (1980), 46–62.

[2] D. Angluin, Inductive inference of formal languages from positive data, *Inform. Contr.* **45** (1980), 117–135.

[3] D. Angluin, Inference of reversible languages, *J. ACM* **29** (1982), 741–765.

[4] D. Angluin and C.H. Smith, Inductive inference: theory and methods, *ACM Comput. Surv.* **15** (1983), 237–269.

[5] W. Bucher and J. Hagauer, It is decidable whether a regular language is pure context-free. *Theoret. Comput. Sci.* **26** (1983), 233–241.

[6] S. Crespi-Reghizzi, G. Guida, and D. Mandrioli, Noncounting context-free languages, *J. ACM* **25** (1978), 571-580.

[7] A. Gabrelian, Pure grammars and pure languages, *Intern. J. Comput. Math.* **9** (1981), 3–16.

[8] P. Garcia, E. Vidal and J. Oncina, Learning locally testable languages in the strict sense, in: *Proceedings of the First International Workshop on Algorithmic Learning Theory* (1990), 325–338.

[9] E.M. Gold, Language identification in the limit, *Inform. Contr.* **10** (1967), 447–474.

[10] J. Hagauer, A simple variable-free CF grammar generating a non regular language. *Bull. EATCS* **6** (1978), 28–33.

[11] M.A. Harrison, *Introduction to Formal Language Theory* (Addison-Wesley, 1978).

[12] T. Koshiba, E. Mäkinen, and Y. Takada, Learning deterministic even linear languages from positive data, *Theoret. Comput. Sci.* **185** (1997), 63–79.

[13] E. Mäkinen, Inferring uniquely terminating regular langauges from positive data, *Inf. Process. Lett.* **62** (1997), 57–60.

[14] H.A. Maurer, A. Salomaa, and D. Wood, Pure grammars, *Inform. Contr.* **44** (1980), 47–72.

[15] T. Motoki, T. Shinohara, and K. Wright, The correct definition of finite elasticity: Corrigendum to identification of unions, in: *Proceedings of 4th Workshop on Computational Learning Theory* (1991), 375.

[16] N. Tanida and T. Yokomori, Inductive inference of monogenic pure contex-free languages, *Lecture Notes in Computer Science* **872** (1994), 560–573.

[17] K. Wright, Identification of unions of languages drawn from an identifiable class, in: *Proceedings of 2nd Workshop on Computational Learning Theory* (1989), 328–333.

[18] T. Yokomori, Inductive inference of 0L languages, in: G. Rozenberg and A. Salomaa (eds.), *Lindenmayer Systems: Impacts on Theoretical Computer Science, Computer Graphics, and Developmental Biology*, Springer, 1992, 115–132.