MTA SZTAKI Közlemények 37/1987 pp.97-116

THE CLASSES OF LANGUAGES CAN BE IDENTIFIED IN THE LIMIT FROM TEXT PRESENTATIONS

LE THANH HAI-NGUYEN QUANG MINH

Department of Computer Science
Polytechnical Institute of Hanoi, Vietnam

1. INTRODUCTION

In recent years the grammatical inference problem has been given increasing attention, not primarily in order to find specific answers to specific questions, but to study various means by which large, but related, classes of problems can be approached [1]-[9], [11], [12]. Briefly stated the problem is this: given a set of finite samples from a language, select, from a set of grammars, a grammar for the given language. Although this problem is formulated in the terminology of formal language theory, it has significant implications for the more general problem of inductive inference since it is deeply concerned with larger questions of methodology [3], [4], [6], [11], [12].

In [1] a formal model for grammatical inference was proposed and the concept of "identification in the limit" was introduced . A sufficient condition under which a class of languages can be identified in the limit, was shown in [2].

The aim of this paper is to extend the results which were presented in [2] about identification in the limit from text presentations.

Section 2 introduces a formal model for grammatical inference which is proposed by Gold [1].

In section 3 the concept of "weighted metrics" on set of languages and a special class of "weighted metrics", which are called ".effective metrics", are defined.

In section 4 the concept of "convergence" of language sequence is defined and the special convergent forms of language sequences according to weighted metrics are investigated. A necessary and sufficient condition for the convergence of a language sequence is showed and some properties of the convergent language sequences are presented among which the invariable property of the convergence of language sequences according to different weighted metrics is remarkable.

In section 5 we propose two grammar inference algorithms and show the conditions under which a class of languages can be identified in the limit from text presentations by these algorithms.

2. A MODEL FOR LANGUAGE IDENTIFICATION

Definition 2.1: An *information sequence*, I(L), of a language L, is an infinite sequence of strings from the set:

$$\{+y \mid y \in L\} \cup \{-y \mid y \in T^{+} - L\}$$

where T is the *technical* alphabet.

Definition 2.2: A *positive* information sequence, $I_p(L)$, is an information sequence of L, containing only strings of the form +y.

Negative information sequence, $I_n(L)$, is similarly defined.

Definition 2.3: An information sequence is *complete* if each string in T^{+} occurs in the sequence.

Similarly, a positive information sequence is complete if each string in L occurs in the sequence.

Definition 2.4: An *arbitrary* information sequence is one in which the strings may appear in arbitrary order.

Definition 2.5: A *sample* $S_t(I)$ of an information sequence I(L) is the unordered set:

$$S_t(I) = \{\pm y_1, \pm y_2, \ldots, \pm y_t\}$$

We distinguish a positive sample:

$$S_{p,t}(I) = \{+y_1, +y_2, \ldots, +y_t\}$$

and a negative sample:

$$S_{n,t}(I) = \{-y_1, -y_2, \ldots, -y_t\}$$

Definition 2.6: A class Γ of grammars over a finite terminal alphabet T is called *acceptable* if:

(i) Grammars in the class can be effectively enumerated
(ii) Each grammar in the class is decidable.

Definition 2.7: Let Γ be any class of grammars over a terminal alphabet T. Then a *grammar inference algorithm* is a function:

$$M_\Gamma : \{\text{finite subsets of } T^+\} \rightarrow \Gamma$$

from sets in T^+ to grammars in the class Γ.

Definition 2.8: The algorithm M_Γ *identifies* the grammar G, or the language L(G), *in the limit* if, for any arbitrary complete information sequence I(L(G)), there is a time t' such that $t > t'$ implies:

$$\text{(i)} \quad A_t = A_{t'}$$

where $A_t = M_\Gamma(S_t)$

$A_{t'} = M_\Gamma(S_{t'})$

and (ii) $L(A_{t'}) = L(G)$.

Definition 2.9: A class Γ of grammars or class $L(\Gamma)$ of languages, is called *identified in the limit* if there is an algorithm that identifies in the limit any grammar G in Γ.

Definition 2.10: A grammar A_t is called *compatible* with a sample S_t if:

$$S_{p,t} \subseteq L(A_t)$$

and

$$S_{n,t} \subseteq T^+ - L(A_t)$$

We distinguish two fundamental methods of information presentation: *text* and *information*. The *text presentation* of a language L is an arbitrary complete positive information sequence of this language and the *information presentation* is an arbitrary complete information sequence.

3. A CLASS OF METRICS ON LANGUAGES

The metrics measuring the "distances" between two languages are defined on the set of all languages Λ_U over an arbitrary finite terminal vocabulary T. That is:

$$\Lambda_U = \{L \mid L \subseteq T^+\}$$

Λ_U is called the *universal class* of languages.

Definition 3.1: Let the terminal vocabulary T be in certain fixed order. Then a unique *lexicographic order* for T^+ can be defined by the following rules:

(i) For any strings $x,y \in T^+$, if $l(x) < l(y)$ then x precedes y in T^+, where $l(x)$ is the length of the string x.

(ii) For any strings $x = a_1a_2\ldots a_n$ and $y = b_1b_2\ldots b_n$, where $x,y \in T^+$ and $l(x) = l(y) = n$, let $a_i = b_i$ for $i = 1,2,\ldots,k$ and $a_{k+1} \neq b_{k+1}$ where $0 \leq k < n$. Then x precedes y in T^+ if a_{k+1} precedes b_{k+1} in T.

Definition 3.2: Let the terminal vocabulary T be in certain fixed order. Then any language $L \in \Lambda_U$ can be uniquely expressed as a binary *membership sequence* $F_L = \langle f_1, f_2, \ldots \rangle$ where $f_i = 1$ if the i'th string in the lexicographic ordering of T^+ is in L and $f_i = 0$ otherwise.

Definition 3.3: $W = \langle w_1, w_2, \ldots \rangle$ is a sequence of *weights* if, for all $i \geq 1$, w_i is positive and rational, and $\sum_{i=1}^{\infty} w_i = 1$.

The following lemmas can be easily proved.

Lemma 3.1: Consider the universal set Λ_U on which an associative binary operation is symmetric difference $\oplus$. Let $W = \langle w_1, w_2, \ldots \rangle$ be a sequence of weights. Then the function:

$$\|L\|_w = \sum_{i=1}^{\infty} f_i w_i$$

defines a norm on Λ_U.

Lemma 3.2: Let w be a sequence of weights. Then the function

$$d_w(L_1, L_2) = \| L_1 \oplus L_2 \|_w$$

defines a metric on Λ_U.

Definition 3.4: Let T^+ have the lexicographic order $T^+ = \langle x_1, x_2, \ldots \rangle$. Then a metric d with a sequence of associated weights $W = \langle w_1, w_2, \ldots \rangle$ is called *effective* if:

(i) $w_i = w_j$ for any i,j satisfying $l(x_i) = l(x_j)$

(ii) $w_i > \sum_{k=j}^{\infty}$ for any i,j satisfying $l(x_i) < l(x_j)$

4. THE CONVERGENCE OF A SEQUENCE OF LANGUAGES CORRESPONDING TO WEIGHTED METRICS

Definition 4.1: Let d be any weighted metric on Λ_U. Then a sequence of languages in Λ_U, $\langle L_1, L_2, \ldots \rangle$, is called *convergent* to a language $L \in \Lambda_U$, *corresponding to metric d*, if: for any $\varepsilon > 0$ there exists n such that for all $k > n$:

$$d(L_k, L) < \varepsilon$$

Lemma 4.1: A sequence of languages, $\langle L_1, L_2, \ldots \rangle$, converges to L if and only if for any positive integer l there exists n such that for all $k > n$, $l(L_k \oplus L) > l$, where $l(L_k \oplus L)$ is the length of the shortest string of $L_k \oplus L$.

Proof: Let T be in certain fixed order. Then T^+ has the following lexicographic ordering:

$$T^+ = \langle x_1, x_2, \ldots, x_{n_1}, x_{n_1+1}, \ldots, x_{n_2}, \ldots, x_{n_k}, \ldots, x_{n_{k+1}}, \ldots \rangle$$

where

$$l(x_{n_k+1}) = l(x_{n_k+2}) = \ldots = l(x_{n_{k+1}}) = k+1.$$

Let l be a positive integer and: $\varepsilon = \min_{1 \leq i \leq n_{l+1}-1} w_i$.

Since $w_i > 0$ then $\varepsilon > 0$. By the convergence of the sequence $\langle L_1, L_2, \ldots \rangle$, there exists n such that for all $k > n$

$$\sum_{i=1}^{\infty} f_i^k w_i < \varepsilon = \min_{1 \leq i \leq n_{l+1}-1} w_i$$

where

$$F_{L_k \oplus L} = \langle f_1^k, f_2^k, \ldots \rangle.$$

This implies:

$$f_i^k = 0 \quad \text{for all} \quad k > n \quad \text{and}$$

$$1 \leq i \leq n_{l+1}-1$$

Conversely, for any $\varepsilon > 0$ let p be the smallest index such that: $\sum_{i=p}^{\infty} w_i < \varepsilon$. For $l = l(x_p)$, there exists n such that for all $k > n$: $l(L_k \oplus L) > l = l(x_p)$.

Then $d(L_k, L) = \| L_k \oplus L \| < \sum_{i=p}^{\infty} w_i < \varepsilon$.

Definition 4.2: A sequence of languages, $\langle L_1, L_2, \ldots \rangle$, converging to L, is called convergent from *inside* (from *outside*) if $L_i \subseteq L$ ($L_o \supseteq L$) for all $i \geq 1$.

Definition 4.3: A sequence of languages, $\langle L_1, L_2, \ldots \rangle$, converging to L, is called convergent *from upside* (*from underside*) if $\|L_i\| \geq \|L\|$ ($\|L_i\| \leq \|L\|$) for all $i \geq 1$.

It is clear that any sequence of languages, converging from inside (from outside) is also a convergent from underside (from upside) sequence. However, the converse is not true.

Lemma 4.2: If sequence of languages, $\langle L_1, L_2, \ldots \rangle$, converges to L corresponding to a weighted metric defined by an order of T, it will converge to L corresponding to metric defined by any order of T.

Proof: Let metric d have a sequence of associated weights: $W = \langle w_1, w_2, \ldots \rangle$ and T have a certain fixed order. Then T^+ has the lexicographic order $T^+ = \langle x_1, x_2, \ldots \rangle$. Let $\langle L_1, L_2, \ldots \rangle$ be a sequence of languages convering to L corresponding to d. Now let T have another order and thus, T^+ has the lexicographic order: $T^+ = \langle x'_1, x'_2, \ldots \rangle$. Let's prove that $\langle L_1, L_2, \ldots \rangle$ also converges to L corresponding to metric d' defined by this order of T.

Since $\sum_{i=1}^{\infty} w_i = 1$ then for any $\varepsilon > 0$ there exists k such that $\sum_{i=k}^{\infty} w_i < \varepsilon$. Let $l = l(x'_k)$. By the lemma 4.1, there exists n such that for all $m > n$:

$$l(L_m \oplus L) > l.$$

Then

$$d'(L_m, L) < \sum_{i=k}^{\infty} w_i < \varepsilon.$$

Lemma 4.3: If a sequence of languages, $\langle L_1, L_2, \ldots \rangle$, converges to L corresponding to metric d with the sequence of associated weights $W = \langle w_1, w_2, \ldots \rangle$, it also converges to L corresponding to any other metric d' with the sequence of associated weights $W' = \langle w'_1, w'_2, \ldots \rangle$.

Proof: Let both metric d and d' be defined by the same order of T. With this order of T, let T^+ have the lexicographic order $T^+ = \langle x_1, x_2, \ldots \rangle$.

Since $\sum_{i=1}^{\infty} w'_i = 1$, for any $\varepsilon > 0$ there exists k such that $\sum_{i=k}^{\infty} w'_i < \varepsilon$. Let $l = l(x_k)$. By the lemma 4.1, there exists n such that for all $m > n$, $l(L_m \oplus L) > l$. Then $d'(L_m, L) < \sum_{i=k}^{\infty} w'_i < \varepsilon$.

From the above two lemmas we have the following theorem:

Theorem 4.1: If a sequence of languages $\langle L_1, L_2, \ldots \rangle$ converges to L corresponidng to a weighted metric, it also converges to L corresponding to any other weighted metric.

The following theorem, similar to the theorem 4.1, reserves the convergence from upside (from underside) corresponding to the effective metric.

Theorem 4.2: If a sequence of languages $<L_1,L_2,...>$ converges from upside (from underside) to L corresponding to an effective metric it also converges from upside (from underside) to L corresponding to any other effective metric.

Proof: By the theorem 4.1, if $<L_1,L_2,...>$ converges to L corresponding to an effective metric, it also converges to L corresponding to any effective metric.

Clearly, the definition of the effective metric does not depend on the order of T. Therefore, assume that T^+ has the lexicographic order $T^+ = <x_1,x_2,...>$, and d_w, $d_{w'}$ are two distinct effective metrics corresponding to the sequence of associated weights:

$$w = <w_1,w_2,...>, \quad w' = <w'_1,w'_2,...>$$

Moreover, assume that:

$$L_i = <x_{i_1},x_{i_2},...,x_{i_k},...>$$

$$L = <x_{j_1},x_{j_2},...,x_{j_k},...>$$

and

$$||L_i||_w > ||L||_w$$

We shall show that $||L_i||_{w'} > ||L||_{w'}$

Indeed, from the definition of effective metric:

If $l(x_{i_1}) < l(x_{j_1})$ then $||L_i||_{w'} > ||L||_{w'}$

If $l(x_{i_1}) = l(x_{j_1})$, $l(x_{i_2}) = l(x_{j_2}),...,l(x_{i_k}) = l(x_{j_k})$

and $l(x_{i_{k+1}}) < l(x_{j_{k+1}})$

$(k \geq 1)$ then $w'_{i_1} = w'_{j_1}$, $w'_{i_2} = w'_{j_2}$,...,$w'_{i_k} = w'_{j_k}$ and

$$w'_{i_{k+1}} > \sum_{p=k+1}^{\infty} w'_{j_p} .$$

Therefore $\| L_i \|_{w'} > \| L' \|_{w'}$

Lemma 4.4: Let $\langle L_1, L_2, \ldots \rangle, \langle L'_1, L'_2, \ldots \rangle$ and $\langle H_1, H_2, \ldots \rangle$ be sequences of languages such that $L_n \subseteq H_n \subseteq L'_n$ and $\langle L_1, L_2, \ldots \rangle$, $\langle L'_1, L'_2, \ldots \rangle$ converge to L. Then the sequence $\langle H_1, H_2, \ldots \rangle$ also coverges to L.

Proof: For all n:

$$d(H_n, L) \leq d(H_n, L'_n) + d(L'_n, L)$$

since $L_n \subseteq H_n \subseteq L'_n$, $d(H_n, L'_n) \leq d(L_n, L'_n)$

Therefore: $d(H_n, L'_n) \leq d(L_n, L) + d(L, L'_n)$

$$d(H_n, L) \leq d(L_n, L) + 2d(L, L_n).$$

By the convergence to L of the sequences $\langle L_1, L_2, \ldots \rangle$ and $\langle L'_1, L'_2, \ldots \rangle$, $d(L_n, L) \to 0$, $d(L'_n, L) \to 0$ as $n \to \infty$.

Therefore $d(H_n, L) \to 0$ as $n \to \infty$ and the sequence $\langle H_1, H_2, \ldots \rangle$ converges to L.

5. THE IDENTIFICATION IN THE LIMIT FROM TEXT PRESENTATION

Wharton [2] has shown that a class of languages which can be identified in the limit from text presentations is a

acceptable class of languages which does not contain a convergent sequence of distinct languages or in which any convergent sequence of distinct languages converges from outside.

In this section we shall propose two grammar inference algorithms and show the conditions under which a class of languages can be identified in the limit from text presentations by these algorithms.

These conditions are weaker than Wharton's, and therefore, our results may be considered to be more general.

Algorithm 5.1: Let Γ be enumerated in the order $\Gamma = \langle G_1, G_2, \ldots \rangle$ and sample at time t is $S_t = \langle y_1, \ldots, y_t \rangle$

(i) Determine a sequence of *possible solutions* at time t:

$$\Sigma_t = \langle H_1, H_2, \ldots, H_{r(t)} \rangle$$

Σ_t is formed by the following process:

Let $\Sigma_{t-1} = \langle H'_1, H'_2, \ldots, H'_{r(t-1)} \rangle$ and the last grammar in Γ that has been examined at time $t-1$ be $G_{s(t-1)}$.

Let $\Sigma'_t = \langle H'_1, \ldots, H'_{r(t-1)}, G_{s(t-1)+1} \rangle$ then Σ_t consists of that subsequence of Σ'_t whose grammars are compatible with S_t.

If $\Sigma_t \neq \phi$, the process is complete.

If $\Sigma_t = \phi$, the next grammar in Γ, $G_{s(t-1)+2}$ is tested for compatibility with S_t. This process continues untill $\Sigma_t \neq \phi$.

At $t = 0$, $\Sigma_t = \phi$

(ii) Select a *tentative solution* A_t from the sequence of possible solutions:

Determine K_t:

Let l_1 be the length of the longest string of S_t and l_2 be the length of the longest string of K_{t-1}. Then, for $t > 1$:

$$K_t = \{x \in T^+ \mid l(x) \leq \max\{l_1, l_2\}+1\}$$

and

$$K_1 = \{x \in T^+ \mid l(x) \leq l_1\}$$

A_t is selected to be the last grammar in the sequence Σ_t which has the following property:

(1) If H_i is the grammar precedes A_t in Σ_t then $L(A_t) \cap K_t \subset L(H_i) \cap K_t$ and the inclusion is strict.

Thus, at each time step of the inference process, one or more grammars in Γ is examined and a finite subsequence Σ_t of Γ is selected as a sequence of possible solutions. This sequence consists of the grammars already examined in Γ which are compatible with S_t. Since each grammar in Γ is decidable, it is also decidable whether any grammar is compatible with the sample and so, at each time, the sequence Σ_t is effectively formed from Γ. The selection of tentative solution A_t at time t is carried out in the sequence of possible solutions at this time and it is the last grammar of the sequence, which has the property (1). Since each grammar is decidable and K_t is finite, set $L(H_i) \cap K_t$ can be effectively determined for each $H_i \in \Sigma_t$. Therefore, the selection of any grammar of such property can be absolutely carried out. Note that at each time, there is at least one grammar which has this property.

Theroem 5.1: Let Γ be an acceptable class of grammars and d be an effectively weighted metric on $L(\Gamma)$. Then the grammar inference algorithm 5.1 identifies in the limit any grammar G

in Γ from any text presentation $I(L(G))$ if $L(\Gamma)$ does not contain a sequence of distinct languages which is convergent from underside.

Proof: First, we shall prove that if $I(L)$ is a text presentation for languages L then the sequence of samples $\langle S_1, S_2, \ldots \rangle$ converges from inside to L. Indeed, for any $\varepsilon > 0$ there exists k such that $\sum_{i=k+1}^{\infty} w_i < \varepsilon$. Since the information sequence is complete, there exists t' such that for all $t \geq t'$, $L \cap \{x_1, \ldots, x_k\} \subseteq S_t \subseteq L$ with the assumption that T^+ has the lexicographic order: $\langle x_1, x_2, \ldots, x_k, \ldots \rangle$. Then for all $t \geq t'$, $d(S_t, L) \leq \sum_{i=k+1}^{\infty} w_i < \varepsilon$. That is, the sequence $\langle S_1, S_2, \ldots \rangle$ converges t. L. Moreover, the sequence converges from inside since $S_t \subseteq L$.

Now, let Γ be enumerated in the order $\Gamma = \langle G_1, G_2, \ldots \rangle$ and G_k be the first grammar in Γ such that $L(G_K) = L(G)$.

Let t_1' be the first time such that $G_k \in \Sigma_t$. Obviously, t_1' is finite and $t_1' \leq k$ since at each time step, one or more grammars in Γ are examined. Thus, for all $t \geq t_1', G_k \in \Sigma_t$.

Consider the sequence of grammars $G_1, G_2, \ldots, G_{k-1}$. Each grammar in this sequence whose language does not include L(G) will be eliminated from the sequence of possible solutions Σ_t at some certain time and never be considered again. Indeed, let G_i be one such grammar and x_p be the first string in the lexicographic order of T^+ which belongs to $L(G_k) - L(G_i)$. Let t_i be the first time at which $x_p \in S_{t_i}$. Since the information sequence is complete one, such time exists.
Obviously, at this time, G_i is not compatible with the sample and hence it will never be selected to be possible solution.

Let t_2' be the maximum of the times t_i corresponding to all grammars $G_i (i < k)$ as above. Let $t_3' = \max\{t_1', t_2'\}$.

For all $t \geq t'_3$, Σ_t has the following form:

$$\Sigma_t = \langle H_1, H_2, \ldots, H_j, G_k, \ldots, H_{r(t)} \rangle$$

where $0 \leq j < r(t)$, j does not depend on t and $L(G_k) \subset L(H_i)$. Moreover, the inclusion is strict since G_k is the first in Γ such that $L(G_k) = L(G)$.

From step (ii) of algorithm 5.1, we have $K_{t-1} \subset K_t$ and $K_{t-1} \neq K_t$. The sequence $\langle K_1, K_2, \ldots \rangle$ is a sequence of distinct languages which converges from inside to T^+. Now, let's prove that there exists a time t'_4 such that for all $t \geq t'_4$. G_k has the property (1).

If $j = 0$, G_k has the property (1).

If $j \geq 1$. For all $t \geq t'_3$, consider the grammar $H_i \in \Sigma_t$ with $1 \leq i \leq j$. We have $L(G_k) \subset L(H_i)$ and $L(G_k) \neq L(H_i)$.

Let y be the first element of T^+ which belongs to $L(H_i) - L(G_k)$ and t''_i be the first time such that $y \in K_t$. Since $\langle K_1, K_2, \ldots \rangle$ converges to T^+, one such time exists. Then for all $t \geq \max\{t_3, t''_i\}$, $L(G_k) \cap K_t \subset L(H_i) \cap K_t$ and the inclusion is strict.

Let $t'_4 = \max\{t_3, t''_1, \ldots, t''_j\}$. Then for all $t \geq t''_4$, G_k has the property (1). Continuely, let's prove that there exists the time t' such that for all $t \geq t'$, G_k is the last grammar in Σ_t which has the property (1). Indeed, assume the contrary, that such time does not exist. Then for any arbitrary great t, there is a grammar $H_t \in \Sigma_t$ such that H_t has the property (1) and H_t follows G_k in Σ_t. We must have $S_t \subseteq L(H_t) \cap K_t \subset L(G_k) \cap K_t \subseteq L(G_k) = L(G)$ and $L(H_t) \cap K_t \neq L(G_k) \cap K_t$.

By the lemma 4.4 and the convergence to L(G) of the sequence of samples, $\langle S_1, S_2, \ldots \rangle$, the sequence $\{L(H_t) \cap K_t\}$ also converges to L(G).

We have:

$$d(L(H_t),L(G)) \leq d(L(H_t),L(H_t)\cap K_t)+d(L(H_t)\cap K_t,L(G))$$

$$\leq d(K_t,T^+)+d(L(H_t)\cap K_t,L(G))$$

Since the sequence $\langle K_1,K_2,\ldots\rangle$ converges to T^+, the sequence $\{L(H_t)\}$ converges to $L(G)$. Moreover, the sequence $\{L(H_t)\}$ is a sequence of distinct languages, because any grammar, following G_k in Γ, will be eliminated from Σ_t at some finite time.

Since $L(H_t)\cap K_t \subset L(G)\cap K_t$ and from the definition of the effective metric we have $\|L(H_t)\| < \|L(G)\|$. Thus $\{L(H_t)\}$ is a sequence of distinct languages which converges from underside to $L(G)$. This contradicts to the hyposthesis.

Thus, there exists t' such that for all $t \geq t'$, G_k is the last grammar in Σ_t which has the property (1). According to the step (ii) of the algorithm 5.1:

For all $t \geq t'$ $\quad A_t = G_k \quad$ and $\quad L(G_k) = L(G)$.

We receive a stronger result in the case when it is decidable whether or not any two grammars of Γ are equivalent.

Algorithm 5.2: Let Γ be enumerated in the order $\langle G_1,G_2,\ldots\rangle$ and the sample at time t is $S_t = \langle y_1,\ldots,y_t\rangle$

(i) Determine a sequence of possible solutions at time t: $\Sigma_t = \langle H_1,\ldots,H_{r(t)}\rangle$ as step (i) of algorithm 5.1.

(ii) Select a tentative solution A_t from the sequence of possible solutions:

A_t is the last grammar in the sequence having the following property:

(2) If H_i is a grammar preceding A_t in Σ_t then

$$L(A_t) \subset L(H_i)$$

and the inclusion is strict.

Theorem 5.2: Let Γ be an acceptable class of grammars and d be an weighted metric on $L(\Gamma)$. Moreover, suppose that it is decidable whether or not any two grammars of Γ are equivalent. Then $L(\Gamma)$ can be identified in the limit from text presentations if and only if $L(\Gamma)$ does not contain a sequence of distinct languages which converges from inside.

Proof: Necessarity: See [2], theorem 2.2.3.

Sufficiency: Consider the grammar inference algorithm 5.2. We shall show that it identifies in the limit any language L in $L(\Gamma)$.

Assume that Γ enumerated in the order $\langle G_1, G_2, \ldots \rangle$ and G_k is the first grammar in Γ such that $L(G_k) = L$.

According to the proof of theorem 5.1, there exists a time t_3' such that for all $t \geq t_3'$, Σ_t has the following form: $\Sigma_t = \langle H_1, \ldots, H_j, G_k, \ldots, H_{r(t)} \rangle$, where $0 \leq j < r(t)$, j does not depend on t and $L(G_k) \subset L(H_i)$. Moreover, the inclusion is strict.

Thus, for all $t \geq t_3'$, G_k is a grammar of Σ_t which has the property (2). Now we shall show that there is a time t' such that for all $t \geq t'$, G_k is the last grammar in Σ_t which has the property (2).

Indeed, assume the contrary that such time does not exist. Then for any arbitrary great t, there is a grammar $H_t \in \Sigma_t$ such that H_t has the property (2) and H_t follow G_k in Σ_t. We have:

$$S_t \subseteq L(H_t) \subset L(G_k) = L.$$

The sequence of samples $\langle S_1, S_2, \ldots \rangle$ converges to L and by lemma 4.4 the sequence $\{L(H_t)\}$ is a sequence of distinct languages which converges from inside to L, a contradiction.

Thus, there exists t' such that for all $t \geq t'$, G_k is the last grammar in Σ_t which has the property (2).

According to the step (ii) of the algorithm 5.2 for all $t \geq t'$ $A_t = G_k$ and $L(G_k) = L$.

6. CONCLUSION

Thus, by using the class of effective weighted metrics on the set of languages and the concept of convergence from underside we have achieved results more general than Wharton's for the identification in the limit from text presentations. Simultaneously, we have developed the necessary and sufficient condition for a sequence of languages be convergent corresponding to a weighted metric. This condition express the essence of the convergence of a sequence of languages and directly leads to the invariable property of the convergence corresponding to different weighted metrics. This property, essentially show the accuracy of our approach.

REFERENCES

[1] Gold, E.M., Language identification in the limit. Information and Control 10(1967), pp. 447-474.

[2] Wharton, R.M., Grammatical inference and approximation. Technical Report, No. 51, Department of Computer Science, University of Toronto, 1973.

[3] Crespi-Reghizzi, S., The mechanical acquisition of precedence grammars, Report UCLA-ENG-7054, School of Engineering and Applied Science, University of California, 1970.

[4] Fu, K.S., Syntactic methods in pattern recognition. School of electrical engineering, Purdue University, Academic Press, New York and London, 1974.

[5] Lu, S.Y. and Fu, K.S., Stochastic Error-Correcting Syntax Analysis for Recognition of Noisy Patterns, IEEE Trans. on Computers, Vol. C-26, No. 12, 1977.

[6] Majumdar, A.K. and Roy, A.K., Inference of fuzzy regular pattern grammar, Pattern Recognition Letters, 2 (1983), pp. 27-32.

[7] Fu, K.S. and Booth, T.K., Grammatical Inference--Introduction and Survey, IEEE Trans. on Systems, Man and Cybernetics, Vol. SMC-5, Jan. and July 1975.

[8] Fu, K.S., Recent Progress in Syntactic Pattern Recognition, School of electrical engineering, Purdue University, W. Lafayette Indiana 47907, U.S.A, 1981.

[9] Thompson, R.A., Determination of probability grammars for functionally specified probability measure languages, IEEE Trans. on Computers, Vol. C-23, June 1974.

[1o] Aho, A.V., and Ullman, J.D., The theory of parsing, translation and compiling, Vol. I., Parsing, Prentice--Hall, Englewood Cliffs, New Jersey, 1972.

[11] You, K.C. and Fu, K.S., A syntactic approach to shape recognition using attributed grammars, IEEE Trans. on Systems, Man and Cybernetic, Vol. SCM-9, June 1979.

[12] You, K.C. and Fu, K.S., Distorted shape recognition using attributed grammars and error-correcting techniques, Computer Graphics and Image Processing, Vol. 13, 1980.

The classes of languages can be identified in the limit from text presentations

Le Thanh Hai, Nguyen Quang Minh

Summary

The problem of identification in the limit from text presentations is considered. A special class of weighted metrics on languages and the special convergent forms according to weighted metrics are investigated and some properties of the convergent language sequence are presented. Finally, two grammar inference algorithms are proposed and the condition of being identifiable in the limit from text presentations by these algorithms for language classes are discussed.

A nyelvek azon osztályai, amelyeket limeszben identifikálni lehet szöveg-mintákból

Le Thanh Hai, Nguyen Quang Minh

Összefoglaló

A cikkben a "limeszben" /határátmenetben/ való identifikálást ismertetik a szerzők. Bevezetnek a nyelvek között speciális, sulyozott metrikákat és azokban való konvergenciát vizsgálják. Végül két algoritmust is megadnak, amelyek segitségével feltételek adhatók meg annak eldöntésére, hogy az adott nyelv limeszben identifikálható-e szöveg-minták alapján vagy sem.