Minimal Concept Identification and Reliability

Sanjay Jain
Department of Information Systems and Computer Science
National University of Singapore
Singapore 119260, Republic of Singapore
Email: sanjay@iscs.nus.edu.sg

Abstract

Identification, by algorithmic devices, of grammars for languages from positive data is a well studied problem. In this paper we are mainly concerned about the learnability of indexed families of uniformly recursive languages. Mukouchi introduced the notion of minimal and reliable minimal concept inference from positive data. He left open a question about whether every indexed family of uniformly recursive languages that is minimally inferable is also reliably minimally inferable. We show that this is not the case.

1 Introduction

Consider the identification of formal languages from positive data. A machine is fed all the strings and no nonstrings of a language L, in any order, one string at a time. The machine, as it is receiving strings of L, outputs a sequence of grammars. The machine is said to identify L just in case the sequence of grammars converges to a grammar for L. A class of languages is said to be identifiable if some machine identifies each language in the class. This is essentially the paradigm of identification in the limit (called **TxtEx**-identification) introduced by Gold [Gol67]. Several important classes of languages such as pattern languages [Ang80a] and length bounded elementary formal systems [Shi94] have been shown to be **TxtEx**-identifiable. On the other hand, any class of languages which contains all the finite languages, and at least one infinite language cannot be **TxtEx**-identified ([Gol67]).

Recently several researchers have focused on the identification of indexed families of uniformly recursive languages (see for example [ZL95]) (see Section 2 below for definitions). The reason for this is mainly due to the fact that several important class of languages, such as pattern languages and length bounded elementary formal systems, are indexed families of uniformaly recursive langauges. Mukouchi [Muk94] considered a variation of **TxtEx**-identification criterion for indexed families of uniformly recursive languages, which we informally describe below (see formal definitions in Section 2 below).

Suppose $\mathcal{L} = L_0, L_1, \ldots$ is an indexed family of uniformly recursive languages. Suppose L is a non-empty language. Then $L_j \in \text{range}(\mathcal{L})$ is said to be a *minimal concept* of L within $\mathcal{L}$, iff (1) $L \subseteq L_j$, and (2) $(\forall L_i \in \text{range}(\mathcal{L}))[L \subseteq L_i \subseteq L_j \Rightarrow L_i = L_j]$. Intuitively, minimal concept of L within $\mathcal{L}$ can be considered as a "good" approximation (from above) of L within $\mathcal{L}$.

A machine $\mathbf{M}$ *minimally* $\mathbf{TxtEx}$-*identifies* $\mathcal{L} = L_0, L_1, \ldots$ if it satisfies the following condition: For any nonempty language L such that there exists a minimal concept of L within $\mathcal{L}$, $\mathbf{M}$ on any positive presentation of L converges to an index j such that L_j is a minimal concept for L within $\mathcal{L}$.

Reliable identification introduced by Minicozzi [Min76] requires that a machine should not converge on texts it does not identify. Based on this Mukouchi considered reliable minimal **TxtEx**-identification. In *reliable minimal* **TxtEx**-*identification* in addition to minimal **TxtEx**-identification, as described above, it is required that: for any nonempty language L, if there exists no minimal concept for L within $\mathcal{L}$, then $\mathbf{M}$ diverges on all positive presentation of L.

Mukouchi showed that the class of length bounded elementary formal systems (and their bounded union) can be reliably, minimally **TxtEx**-identified. On the other hand, since minimal **TxtEx**-identification implies **TxtEx**-identification, not every class of indexed family can be minimally **TxtEx**-identified. Mukouchi also showed that there exist **TxtEx**-identifiable classes which cannot be minimally **TxtEx**-identified. However, Mukouchi left open the question whether reliable minimal inference restricts minimal inference criteria. We show that this is indeed the case.

We now proceed formally.

2 Preliminaries

N denotes the set of natural numbers. $\emptyset$, $\subseteq$, $\subset$, $\supseteq$, $\supset$, respectively denote empty set, subset, proper subset, superset, and proper superset. A language is a non-empty subset of N. We let L, with or without subscripts, range over languages. $\mathcal{L}$ ranges over sets of languages. $\langle \cdot, \cdot \rangle$ denotes a computable, bijection from $N \times N$ to N. Intuitively, $\langle \cdot, \cdot \rangle$ is a pairing function. Similarly one can define $\langle \cdot, \cdot, \cdot \rangle$ as a computable bijection from $N \times N \times N$ to N. Quantifier $\overset{\infty}{\forall}$ denotes for all but finitely many.

A *text* is an infinite sequence of elements from N (in other words, text can be considered as a mapping from N to N). We let T (with or without subscripts), range over texts. $T[n]$ denotes the finite initial sequence of T of length n. SEQ denotes the set of all finite sequences. We let σ, (with or without subscripts and superscripts), range over finite sequences. $\sigma \diamond (w)$ denotes the concatenation of w at the end of finite sequence σ.

Suppose $T = w_0, w_1, w_2, \ldots$. Then, the content of T, denoted content(T), is the set $\{w_0, w_1, \ldots\}$. For a finite sequence σ, we define content(σ) analogously. T is said to be a *text for* L, iff content$(T) = L$.

An *inductive inference machine* is an algorithmic mapping from SEQ to N. $\mathbf{M}$ ranges over inductive inference machines. $\mathbf{M}$ is said to converge on T to i, (written $\mathbf{M}(T)\!\downarrow = i$) iff, $(\overset{\infty}{\forall} n)[\mathbf{M}(T[n]) = i]$. If there exists an i such that $\mathbf{M}(T)\!\downarrow = i$, then we say that $\mathbf{M}(T)\!\downarrow$. Otherwise, we say that $\mathbf{M}(T)\!\uparrow$.

A sequence of non-empty recursive languages, $\mathcal{L} = L_0, L_1, \ldots$ is said to be an indexed family of uniformly recursive languages (often denoted by just indexed family), just in case, there exists a recursive function f, such that

$$f(i,x) = \begin{cases} 1, & \text{if } x \in L_i; \\ 0, & \text{otherwise.} \end{cases}$$

range($\mathcal{L}$) $= \{L_i \mid i \in N\}$, denotes the set of languages in the indexed family $\mathcal{L}$.

Definition 1 [Gol67, Ang80b] Suppose $\mathcal{L} = L_0, L_1, \ldots$ is an indexed family of uniformly recursive languages. $\mathbf{M}$ **TxtEx**-identifies $\mathcal{L}$ iff $(\forall L \in \text{range}(\mathcal{L}))(\forall \text{ texts } T \text{ for } L)(\exists i \mid L_i = L)[\mathbf{M}(T)\downarrow = i]$.

TxtEx $= \{\mathcal{L} \mid (\exists \mathbf{M})[\mathbf{M}$ **TxtEx**-identifies $\mathcal{L}]\}$.

Definition 2 [Muk94] Suppose $\mathcal{L} = L_0, L_1, \ldots$ is an indexed family of uniformly recursive languages. Suppose $L \subseteq N$, is non-empty. $L_j \in \text{range}(\mathcal{L})$ is said to be a *minimal concept of* L *within* $\mathcal{L}$ iff

(1) $L \subseteq L_j$, and

(2) $(\forall L_i \in \text{range}(\mathcal{L}))[L \subseteq L_i \subseteq L_j \Rightarrow L_i = L_j]$.

Intuitively, one may consider minimal concept of L within $\mathcal{L}$ as a "good" approximation (from above) of L from languages in $\mathcal{L}$, in the sense that there is no other approximation which is "strictly better".

Definition 3 [Muk94] Suppose $\mathcal{L} = L_0, L_1, \ldots$ is an indexed family of uniformly recursive languages. $\mathbf{M}$ *minimally* **TxtEx**-*identifies* $\mathcal{L}$ iff, for all L such that there exists a minimal concept for L within $\mathcal{L}$,

$(\forall$ texts T for $L)(\exists i \mid L_i$ is a minimal concept for L within $\mathcal{L})[\mathbf{M}(T)\downarrow = i]$.

Based on the definition of reliable inference by [Min76] (see also [BB75, CJNM94]), Mukouchi also considered reliable minimal identification. Intuitively, a machine is reliable if it does not converge on functions it fails to identify. In other words, the machine does not give a false signal by converging to a wrong grammar: every wrong hypothesis is eventually rejected by a mind change.

Definition 4 [Muk94] Suppose $\mathcal{L} = L_0, L_1, \ldots$ is an indexed family of uniformly recursive languages. $\mathbf{M}$ *reliably, minimally* **TxtEx**-*identifies* $\mathcal{L}$ iff, the following two conditions are satisfied.

(1) for all L such that there exists a minimal concept for L within $\mathcal{L}$,

$$(\forall \text{ texts } T \text{ for } L)(\exists i \mid L_i \text{ is a minimal concept for } L \text{ within } \mathcal{L})[\mathbf{M}(T)\downarrow = i];$$

and

(2) for all L such that there exists no minimal concept for L within $\mathcal{L}$, $(\forall$ texts T for $L)[\mathbf{M}(T)\uparrow]$.

Note that in the above definitions, we have essentially used the indexed family being learned as the hypothesis space also [ZL95]. It can be shown that, for the above criteria (**TxtEx**, minimal **TxtEx**, and reliable minimal **TxtEx**-identification) the learnable indexed families do not change if one allows class comprising hypothesis spaces [ZL95]. Thus for the sake of simplicity, we only consider the indexed family themselves as hypothesis space for this paper.

Proposition 1 *There exists a recursively enumerable sequence* $\mathbf{M}_0$*,* $\mathbf{M}_1$*, ... of total inductive inference machines such that:*

(1) If some $\mathbf{M}$ **TxtEx***-identifies* $\mathcal{L}$*, then* $(\exists i \in N)[\mathbf{M}_i$ **TxtEx***-identifies* $\mathcal{L}]$*.*

(2) If some $\mathbf{M}$ *minimally* **TxtEx***-identifies* $\mathcal{L}$*, then* $(\exists i \in N)[\mathbf{M}_i$ *minimally* **TxtEx***-identifies* $\mathcal{L}]$*.*

(3) If some $\mathbf{M}$ *reliably minimally* **TxtEx***-identifies* $\mathcal{L}$*, then* $(\exists i \in N)[\mathbf{M}_i$ *reliably minimally* **TxtEx***-identifies* $\mathcal{L}]$*.*

For proof of part (1) of the above proposition see, for example, [OSW86]. The same proof works for part (2) and (3) also. We let $\mathbf{M}_0, \mathbf{M}_1, \ldots$ be a recursively enumerable sequence of machines satisfying Proposition 1.

3 Main Result

We now show that reliability restricts minimal **TxtEx**-identification for indexed families of recursive languages.

Theorem 1 *There exists an indexed family* $\mathcal{L}$ *of languages such that*

(1) Some machine $\mathbf{M}$ *minimally* **TxtEx***-identifies* $\mathcal{L}$*.*

(2) No machine can reliably, minimally **TxtEx***-identify* $\mathcal{L}$*.*

Proof. For each $i \in N$, we define a text $T_i = \bigcup_{j\in N} \sigma_i^j$ as follows.

Let $\sigma_i^0 = \sigma_i^1$ be finite sequences containing just one element $\langle i, 0, 0\rangle$.

For $j > 1$, σ_i^j is defined as follows:

If $\mathbf{M}_i(\sigma_i^{j-2}) \neq \mathbf{M}_i(\sigma_i^{j-1})$, then $\sigma_i^j = \sigma_i^{j-1} \diamond (\langle i, 0, j\rangle)$. Otherwise, $\sigma_i^j = \sigma_i^{j-1} \diamond (\langle i, 0, 0\rangle)$.

Let $T_i = \bigcup_{j\in N} \sigma_i^j$.

The following claim is easy to verify.

Claim 1 *(1)* $\mathbf{M}_i(T_i)\downarrow$ *iff* $\text{content}(T_i)$ *is finite.*

(2) One can (effectively in i) find a decision procedure for $\text{content}(T_i)$*.*

(3) $\text{content}(T_i) \subseteq \{\langle i, 0, x\rangle \mid x \in N\}$*.*

We define languages $L_{\langle i,j,k\rangle}$ effectively in i, j, k below.

Let $L_{\langle i,0,0\rangle} = \text{content}(T_i) \cup \{\langle i, 1, x\rangle \mid x \in N\}$.

For j, k such that $j + k > 0$, let $L_{\langle i,j,k\rangle} = \{\langle i, 0, x\rangle \mid x \leq j \ \wedge \ \langle i, 0, x\rangle \in \text{content}(T_i)\} \cup \{\langle i, 1, x\rangle \mid x \geq j + k\}$.

Let $\mathcal{L} = (L_{\langle i,j,k\rangle})_{i,j,k\in N}$.

Clearly, $\mathcal{L}$ is an indexed family of recursive languages.

Claim 2 *(1) No language in* $\text{range}(\mathcal{L})$ *contains any element of the form* $\langle i, y, x\rangle$*, for* $y > 1$*. Moreover,* $\langle i, 1, x\rangle \in L_{\langle i',j',k'\rangle} \Leftrightarrow [i' = i$ *and* $j' + k' \leq x]$*.*

(2) Suppose $L \subseteq \{\langle i, 0, x\rangle \mid x \in N\}$ *is given. Then either*

(2.1) $L \subseteq \text{content}(T_i)$*,* L *is infinite, and* $L_{\langle i,0,0\rangle}$ *is the only minimal concept of* L *within* $\mathcal{L}$*, or*

(2.2) $L \not\subseteq \text{content}(T_i)$ *and there is no minimal concept of* L *within* $\mathcal{L}$*, or*

(2.3) L *is finite and there exists no minimal concept of* L *within* $\mathcal{L}$*.*

Proof. (1) Follows, easily from the construction of $L_{\langle i,j,k\rangle}$.

(2) Suppose $L \subseteq \{\langle i,0,x\rangle \mid x \in N\}$. Note that (a) $L_{\langle i,j,k\rangle} \supset L_{\langle i,j,k+1\rangle}$, and (b) for $j+k \neq 0$, $\{\langle i,0,x\rangle \mid x \in N\} \cap L_{\langle i,j,k\rangle} = \{\langle i,0,x\rangle \mid x \in N\} \cap L_{\langle i,j,k+1\rangle}$. Thus, we have that only minimal concept for L (if any) within $\mathcal{L}$ can be $L_{\langle i,0,0\rangle}$. Part (2) is now straightforward. □

Claim 3 *There exists a machine* $\mathbf{M}$ *which minimally* $\mathbf{TxtEx}$*-identifies* $\mathcal{L}$.

Proof. Consider the following machine $\mathbf{M}$. For any nonempty sequence σ,

```
$\mathbf{M}(\sigma)$
0. Let $i$ be such that content($\sigma$) $\subseteq L_{\langle i,0,0\rangle}$ (if no such $i$ exists, then let $\mathbf{M}(\sigma) = 0$).
1. If content($\sigma$) $\subseteq \{\langle i,0,x\rangle \mid x \in N\}$, then output $\langle i,0,0\rangle$.
2. Otherwise, let $j = \max(\{x \mid \langle i,0,x\rangle \in \text{content}(\sigma)\})$, and
   $k = \min(\{x \mid \langle i,1,x\rangle \in \text{content}(\sigma)\}) - j$.
3. If $j, k \geq 0$, and content($\sigma$) $\subseteq L_{\langle i,j,k\rangle}$, then output $\langle i,j,k\rangle$. Otherwise output $\langle i,0,0\rangle$.
End $\mathbf{M}(\sigma)$.
```

It is easy to verify, using Claim 2, that $\mathbf{M}$ minimally $\mathbf{TxtEx}$-identifies $\mathcal{L}$. □

Claim 4 *No machine can reliably minimally* $\mathbf{TxtEx}$*-identify* $\mathcal{L}$.

Proof. Suppose by way of contradiction that $\mathbf{M}_i$ reliably minimally $\mathbf{TxtEx}$-identifies $\mathcal{L}$. Now,

(1) If $\mathbf{M}_i(T_i)\downarrow$, then content($T_i$) is finite subset of $\{\langle i,0,x\rangle \mid x \in N\}$, and thus, there exists no minimal concept of content(T_i) within $\mathcal{L}$.

(2) If $\mathbf{M}_i(T_i)\uparrow$, then content($T_i$) is an infinite subset of $\{\langle i,0,x\rangle \mid x \in N\}$, and thus, $L_{\langle i,0,0\rangle}$ is a minimal concept of content(T_i) within $\mathcal{L}$.

Thus $\mathbf{M}_i$ cannot reliably minimally $\mathbf{TxtEx}$-identify $\mathcal{L}$. Claim follows. □

Theorem follows from Claims 3 and 4. ■

4 Acknowledgements

I would like to thank Y. Mukouchi for helpful discussions. I also thank the referees for several helpful suggestions.

References

[Ang80a] D. Angluin. Finding patterns common to a set of strings. *Journal of Computer and System Sciences*, 21:46–62, 1980.

[Ang80b] D. Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45:117–135, 1980.

[BB75] L. Blum and M. Blum. Toward a mathematical theory of inductive inference. *Information and Control*, 28:125–155, 1975.

[CJNM94] J. Case, S. Jain, and S. Ngo Manguelle. Refinements of inductive inference by Popperian and reliable machines. *Kybernetika*, 30:23–52, 1994.

[Gol67] E. M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[Min76] E. Minicozzi. Some natural properties of strong identification in inductive inference. *Theoretical Computer Science*, pages 345–360, 1976.

[Muk94] Y. Mukouchi. Inductive inference of an approximate concept from positive data. In S. Arikawa and K. Jantke, editors, *Algorithmic learning theory: Fourth International Workshop on Analogical and Inductive Inference (AII '94) and Fifth International Workshop on Algorithmic Learning Theory (ALT '94)*, volume 872 of *Lecture Notes in Artificial Intelligence*, pages 484–499. Springer-Verlag, 1994.

[OSW86] D. Osherson, M. Stob, and S. Weinstein. *Systems that Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists.* MIT Press, 1986.

[Shi94] T. Shinohara. Rich classes inferable from positive data: Length–bounded elementary formal systems. *Information and Computation*, 108:175–186, 1994.

[ZL95] T. Zeugmann and S. Lange. A guided tour across the boundaries of learning recursive languages. In K. Jantke and S. Lange, editors, *Algorithmic Learning for Knowledge-Based Systems*, volume 961 of *Lecture Notes in Artificial Intelligence*, pages 190–258. Springer-Verlag, 1995.