Sufficient conditions: finite thickness, finite elasticity and characteristic sets

Nina Gierasimczuk Dick de Jongh

February 26th, 2013

1 Finite thickness

Let us recall the characterization of effective identifiability.

Theorem 1 (Angluin's Theorem) *A uniformly recursive family $\mathcal{L}$ is effectively identifiable iff there exists a recursive function F such that for each i, n, $F(i, n)$ is a finite set D_i^n, with $m \leq n \Rightarrow D_i^m \subseteq D_i^n$, and the limit $D_i = \bigcup_{n \in \mathbb{N}} D_i^n$ is a telltale set for L_i.*

Definition 1 (Angluin 1980) *A class of languages $\mathcal{L}$ has* finite thickness *if for each $w \in \mathbb{N}$ there are only finitely many $L \in \mathcal{L}$ such that $w \in L$.*

Theorem 2 (Angluin 1980) *If uniformly recursive $\mathcal{L}$ is of finite thickness, then $\mathcal{L}$ is effectively identifiable.*

Note that the converse is generally not true, i.e., there are classes that are effectively identifiable but are not of finite thickness, e.g., the class of all finite sets.

Proof Let $\mathcal{L}$ be a uniformly recursive family of languages and let $\mathcal{L}$ be of finite thickness. We will show that $\mathcal{L}$ satisfies the condition from the characterization Theorem 1.

For each $i, n \in \mathbb{N}$, $L_i^{(n)}$ will denote the computable finite set $L_i \cap \{0, \ldots, n\}$. We define a procedure to enumerate a set T_i from i in stages, beginning with stage 0.

Stage 0. Let w_1 be the least element of L_i. Set $A_1 := \{w_1\}$, put w_1 into T_i and go to stage 1.

Stage n. ($n \geq 1$) Take the first pair (j, m) such that (j, m) has not been considered in previous stages and $A_n \subseteq L_j$ and $L_j^{(m)} \subset L_i^{(m)}$ if there is any such pair. If one is found, set $A_{n+1} := L_i^{(m)}$, put the elements of A_{n+1} into T_i, and go to stage $n+1$, otherwise set $A_{n+1} := A_n$ and go to stage $n+1$ as well.

Clearly T_i may be (uniformly) effectively enumerated from i. Also, each A_n constructed is an initial segment of L_i and properly contains A_{n-1} if $n > 1$. There are finitely many distinct languages from the given family that contain the string w_1, by finite thickness of $\mathcal{L}$. Each stage of the procedure must find and "cancel" (by enlarging A_{n+1} so that it is not contained in the language) at least one such language containing w_1. Further, no such language need be cancelled more than once (by the monotonicity of $A_1 \subseteq A_2 \subseteq \ldots$), so there are only finitely many stages in the execution of this procedure in which A_n gets to be properly extended.

That is, there exists some positive integer n such that in this procedure $A_m = A_n$ for all $m \geq n$. Thus $T_i := \bigcup_{n \in \mathbb{N}} A_n$ is of finite cardinality, and clearly $T_i \subseteq L_i$. Assume that for some $j \geq 1$, $T_i \subseteq L_j$ and $L_j \subset L_i$. Then for some m, $L_j^{(m)} \subset L_i^{(m)}$, and the procedure must eventually find in some stage n this (or some other) pair (j, m) such that $A_n \subseteq L_j$ and $L_j^{(m)} \subset L_i^{(m)}$, and go on to stage $n+1$ extending A_n to A_{n+1} with an element not in L_j, contradicting $T_i \subseteq L_j$. □

2 Exact, class comprising and class preserving learning

Definition 2 *A class $\mathcal{L}$ of non-empty recursive languages is* indexable *iff there is a uniformly recursive family of languages $(L_i)_{i \in \mathbb{N}}$ such that $\mathcal{L} = \{L_i \mid i \in \mathbb{N}\}$. Such a family is called an indexing of $\mathcal{L}$.*

Definition 3 *An indexed family $\mathcal{L}$ is* exactly learnable *if $\mathcal{L}$ is identifiable with respect to itself, i.e., the learning function uses $\mathcal{L}$ as hypothesis space.*

Definition 4 *A family $\mathcal{L}$ is* identifiable by a class preserving learning function *M, if there is a space $\mathcal{G} = (G_j)_{j \in \mathbb{N}}$ of hypotheses such that any G_j describes a language from $\mathcal{L}$ and M infers $\mathcal{L}$ with respect to $\mathcal{G}$.*

Here, any produced hypothesis is required to describe a language belonging to $\mathcal{L}$ but M is free to use a possibly different enumeration of $\mathcal{L}$ and possibly different descriptions of any $L \in \mathcal{L}$.

Definition 5 *A family $\mathcal{L}$ is* identifiable by a class comprising learning function *M, if there is a space $\mathcal{G} = (G_j)_{j \in \mathbb{N}}$ of hypotheses such that any $L \in \mathcal{L}$ has a description G_j but $\mathcal{G}$ may additionally contain elements G_k not describing any language from $\mathcal{L}$ and M infers $\mathcal{L}$ with respect to $\mathcal{G}$.*

3 Finite thickness revisited

Theorem 3 *If uniformly recursive $\mathcal{L}$ is of finite thickness, then $\mathcal{L}$ is effectively identifiable by a class comprising learning function.*

Proof Let us assume that a 1-1 recursive indexing $(L_i)_{i\in\mathbb{N}}$ of the class $\mathcal{L}$ is given. Let $(L'_k)_{k\in\mathbb{N}}$ be an indexing comprising $\mathcal{L}$, such that $L'_k = L_{j_1} \cap \ldots \cap L_{j_z}$, if k is the canonical index of the finite set $\{j_1, \ldots, j_z\}$. The proposed learning method uses the family $(L'_k)_{k\in\mathbb{N}}$ as a hypothesis space. (This is not quite correct, but we will correct the sloppyness at the end of the proof.) We define

$$M(t[n]) = k, \text{ where } k \text{ is the canonical index of the set}$$
$$D = \{j \mid j \leq n,\ \mathit{content}(t[n]) \subseteq L_j\},$$

if this D is non-empty, otherwise $\{j\}$ if L_j is the first language that contains $\mathit{content}(t[n])$.

M considers a set of possible hypotheses in each learning step and outputs a hybrid hypothesis which is constructed from this set. Let t be a text for L_i. Since $\mathcal{L}$ is of finite thickness and $(L_j)_{j\in\mathbb{N}}$ is a one-one indexing of $\mathcal{L}$, there is a number n such that (1) for each L' such that $L_j \not\subseteq L'$ there is a $w \in \mathit{content}(t[n]), w \notin L'$, (2) $t_0 \notin L_j$ for every $j > n$. From that n onwards there is a fixed set of languages in D and all of those languages contain L_j as a subset. So, the intersection of those languages is L_j, and hence method M stabilizes on a correct hypothesis for the target language.

Correction: $(L'_k)_{k\in\mathbb{N}}$ may not immediately qualify as a uniformly recursive set, because $L'_k = L_{j_1} \cap \ldots \cap L_{j_z}$ may be empty, we do not want empty languages, and emptyness is in general an undecidable property. It can be transformed into a uniformly recursive family by considering with each w all the L'_k-languages it is an element of, which is a decidable matter. This is an enumerable class. Then one takes the union of all the classes for each w. In this manner one obtains a uniformly recursive class of languages in which the empty languages which shouldn't occur are avoided. □

Definition 6 (Koshiba 1995; Lange and Zeugmann 1996) *A uniformly recursive class $\mathcal{L}$ has* recursive thickness *if there is a recursive function F such that, for each w, $F(w)$ is the canonical code k for the finite set D_k such that $w \in L_i$ iff $i \in D_k$.*

Theorem 4 *If a uniformly recursive class $\mathcal{L}$ has recursive finite thickness by the function F, then $\mathcal{L}$ is identifiable by an effective incremental learner.*

Proof An incremental method for learning a class $\mathcal{L}$ which is of recursive finite thickness, witnessed by an indexing $(L_j)_{j\in\mathbb{N}}$ of $\mathcal{L}$, uses an indexing $(L'_k)_{k\in\mathbb{N}}$ comprising $\mathcal{L}$, such that $L'_k = L_{j_1} \cap \ldots \cap L_{j_z}$, if k is the canonical index of the finite set $\{j_1, \ldots, j_z\}$ (as discussed in the previous proof).

1. On input t_0, M outputs the canonical index $F(t_0)$ of the set $D = \{j \mid j \in \mathbb{N},\ t_0 \in L_j\}$.
2. On input $t[n+1]$, let k be the hypothesis returned by M on $t[n]$. M computes the set D_k for which k is the canonical index, and then computes the set $D' = \{j \mid j \in D_k,\ t_{n+1} \in L_j\}$. M outputs the canonical index $I(k, F(t_{n+1}))$ of D'.

□

This method uses a recursive indexing comprising the target class and is iterative, i.e., in each step of the learning process, it uses only its previous hypothesis and the latest positive example presented in the text.

4 Finite elasticity

Definition 7 (Wright 1989; Motoki et al. 1991) *A class $\mathcal{L}$ is of* infinite elasticity *iff there is an infinite sequence $w_0, w_1, \ldots$ of natural numbers and an infinite sequence $L_1, L_2, \ldots$ of languages in $\mathcal{L}$ such that the following conditions are fulfilled for all $n \in N^+$.*

1. $\{w_0, ..., w_{n-1}\} \subseteq L_n$;
2. $w_n \notin L_n$.

$\mathcal{L}$ is of finite elasticity *iff $\mathcal{L}$ is not of infinite elasticity.*

Theorem 5 (Wright 1989) *Let $\mathcal{L}$ be a uniformly recursive family. If $\mathcal{L}$ is of finite elasticity, then $\mathcal{L}$ is effectively identifiable.*

Again, this criterion is sufficient, but not necessary for learnability in the limit from text. It is easily seen that the class of all finite languages is in LimTxt but is of infinite elasticity. Note that each class possessing the finite thickness property is also of finite elasticity. The converse is not valid in general; for instance, the class of all languages containing exactly two numbers is of finite elasticity but not of finite thickness.

Proof We will show that $\mathcal{L}$ satisfies the condition of Angluin's theorem. Define:

$$D_i = \{x \mid \exists j(x \simeq \mu y(y \in L_i - L_j))\}.$$

The predicate $x \in D_i$ is recursively enumerable, because $x \in D_i$ iff $\exists j\,(x \in L_i \wedge x \notin L_j \wedge \forall y < x\,(y \in L_i \rightarrow y \in L_j))$. We claim that D_i is a telltale set for L_i in $\mathcal{L}$. It is clear that if $D_i \subseteq L_j$, then $L_i \subseteq L_j$. For contradiction assume that D_i is infinite. Then for any $n \in \mathbb{N}$ there is a j such that $L_i - L_j \neq \emptyset$ and $\mu y\,(y \in L_i - L_j) > n$. Then a pair witnessing the infinite elasticity of $\mathcal{L}$ is as follows. Choose some s_0 and $L^0 \in \mathcal{L}$ be such that $s_0 \in L_i - L^0$. For each $n > 0$, let:

$$L^n = L_j$$

$$w_n = \mu y\,(y \in L_i - L_j),$$

where j is the least such that $L_i - L_j \neq \emptyset$ and $\mu y\,(y \in L_i - L_j) > w_{n-1}$. Clearly, $s_0, s_1, \ldots$ and $L^0, L^1, \ldots$ witness the infinite elasticity of $\mathcal{L}$. □

Finite elasticity and finite thickness can be exploited for learning bounded unions of languages from a uniformly recursive class $\mathcal{L}$. Below, for any $k \in \mathbb{N}^+$, we use $\mathcal{L}^k$ to denote the class of all unions of up to k languages from $\mathcal{L}$, i.e., $\mathcal{L}^k = \{L_{j_1} \cup \ldots \cup L_{j_l} \mid L_{j_1}, \ldots, L_{j_l} \in \mathcal{L}, l \leq k\}$.

Theorem 6 *Let $\mathcal{L}$ be a uniformly recursive family and $k \in \mathbb{N}^+$. If $\mathcal{L}$ is of finite elasticity, then $\mathcal{L}^k$ is of finite elasticity.*

Corollary 7 *Let $\mathcal{L}$ be a uniformly recursive family and $k \in \mathbb{N}^+$. If $\mathcal{L}$ is of finite elasticity, then $\mathcal{L}^k$ is effectively identifiable.*

In particular, since each class of finite thickness is also of finite elasticity, we obtain a stronger result.

Corollary 8 *Let $\mathcal{L}$ be a uniformly recursive family and $k \in \mathbb{N}^+$. If $\mathcal{L}$ is of finite thickness, then $\mathcal{L}^k$ is effectively identifiable.*

5 Characteristic sets

Definition 8 (Angluin 1982) *Let $\mathcal{L}$ be uniformly recursive. A family $(S_i)_{i\in\mathbb{N}}$ of non-empty finite sets is called a* family of characteristic sets for $\mathcal{L}$ *iff, for all $i, j \in \mathbb{N}$,*

1. *$S_i \subseteq L_i$,*
2. *if $S_i \subseteq L_j$, then $L_i \subseteq L_j$.*

Again, possessing a family of characteristic sets is sufficient but not necessary for effective identification.

Theorem 9 (Kobayashi) *Let $\mathcal{L}$ be a uniformly recursive family. If $\mathcal{L}$ has a family of characteristic sets, then $\mathcal{L}$ is effectively identifiable.*

Proof Let $(S_i)_{i\in\mathbb{N}}$ be a family of characteristic sets for $\mathcal{L}$. We define M as follows on a text t for a member L of $\mathcal{L}$.

$M(t[n])$ = the least $j \leq n$ such that $\mathit{content}(t[n]) \subseteq L_j$ and for no $k \leq n$, $\mathit{content}(t[n]) \subseteq L_k$ and $L_k^{(n)} \subset L_j^{(n)}$.

If such j exists, otherwise 0.[1]

Let j be the least index for the language of the text. There will exist an n such that

1. $j \leq n$ and $S_j \subseteq \mathit{content}(t[n])$,
2. for all $k < j$ such that $S_j \subseteq L_k$ a member n_k of $L_k - L_j$ exists in $\mathit{content}(t[n])$.

This is so because, for any k with $k < j$ and $S_j \subseteq L_k$, $L_j \subseteq L_k$ holds, and $L_j = L_k$ is excluded by the minimality of j. Now, $M(t[m]) = j$ for all $m \geq n$, and M learns L_j. □

[1] or use some other way to preserve consistency.

Theorem 10 *If uniformly recursive $\mathcal{L}$ has finite elasticity, then it has a family of characteristic sets.*

Proof Let w be the smallest element of an arbitrary language L_j in $\mathcal{L}$. We define a sequence of elements $w_0, w_1, \ldots$ of L_j and a sequence of languages $L'_1, L'_2, \ldots$ in $\mathcal{L}$ by the following procedure.
Stage 0: Set $w_0 = w$.
Stage n, $n > 0$: Let $w_0, \ldots, w_{n-1}$ have been defined previously. Search for the smallest pair k, v such that $\{w_0, \ldots, w_{n-1}\} \subseteq L_k$ and $v \in L_j - L_k$. If such a pair exists, set $w_n = v$, $L'_n = L_k$, and goto Stage $n + 1$.

If the procedure continues for all n, we get for all n, $\{w_0, \ldots, w_{n-1}\} \subseteq L'_n$ and $w_n \notin L'_n$, contradicting finite elasticity. So, there has to be an n such that the procedure gets stuck in Stage n. We then define $S_j = \{w_0, \ldots, w_{n-1}\}$. Assume $S_j \subseteq L_k$. If $L_j \not\subseteq L_k$, a number $v \in L_j - L_k$ exists, but then Stage n would be finished. So, that is impossible and we have $L_j \subseteq L_k$, and S_j is a characteristic set for L_j. □

An example of an uniformly recursive effectively identifiable class with infinite elasticity is the class of all finite sets.

References

Angluin, D. (1980). Inductive inference of formal languages from positive data. *Information and Control*, 45(2):117–135.

Koshiba, T. (1995). Typed pattern languages and their learnability. In Vitanyi, P., editor, *Computational Learning Theory*, volume 904 of *Lecture Notes in Computer Science*, pages 367–379. Springer Berlin Heidelberg.

Lange, S. and Zeugmann, T. (1996). Incremental learning from positive data. *Journal of Computer and System Sciences*, 53(1):88 – 103.

Motoki, T., Shinohara, T., and Wright, K. (1991). The correct definition of finite elasticity: corrigendum to identification of unions. In *Proceedings of the fourth annual workshop on Computational learning theory*, COLT '91, page 375, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Wright, K. (1989). Identification of unions of languages drawn from an identifiable class. In *Proceedings of the Second Annual Workshop on Computational Learning Theory, COLT 1989, Santa Cruz, CA, USA, July 31 - August 2, 1989*, pages 328–333. Morgan Kaufmann Publishers Inc.