Some learning of regular languages

Nina Gierasimczuk Dick de Jongh

February 22th, 2013

Definition 1 *An* alphabet Σ *is a finite set of letters.* Σ^* *is the set of all finite sequences over* Σ*.*

If the alphabet is $\{a, b\}$, we write, e.g., *aba* for the sequence $\langle a, b, a\rangle$. The set of alls sequences over Σ can be identified with the natural numbers by coding. If $X, Y \subseteq \Sigma^*$, we write XY for $\{xy \,|\, x \in X, y \in Y\}$. X^* is the set of all finite sequences over X including the empty sequence, X^+ the set of all finite sequences over X not including the empty sequence.

Definition 2 *A* regular expression *over* Σ *is defined inductively by:*

1. *If* $a \in \Sigma$*, then* a *is a regular expression over* Σ*,*
2. *If* p, q *are regular expressions over* Σ*, then* pq*,* $p \cup q$ *and* p^* *are regular expressions over* Σ*.*

Each regular expression defines a regular language.

Definition 3

1. *If* $a \in \Sigma$*, then* $L(a) = \{a\}$*,*
2. $L(pq) = \{xy \,|\, x \in p, y \in q\}$*,*
3. $L(p \cup q) = L(p) \cup L(q)$*,*
4. $L(p^*) = (L(p))^*$*.*

It is a well-known fact that the regular languages are exactly the ones accepted by *finite automata*. It follows from Gold's second theorem that not all regular languages are identifiable.

The restricted regular expressions come in two kinds, the *restricted regular* **-expressions* are obtained by deleting the operation $\cup$, the *restricted regular* $^+$*-expressions* by deleting the operation $\cup$ and replacing * by $^+$. The class of all *erasing languages of restricted regular* **-expressions* is written $RREG_*^\Sigma$. The class of all *non-erasing languages of restricted regular* $^+$*-expressions* is written $RREG_+^\Sigma$, and is of course obtained by replacing the clause $L(p^*) = (L(p))^*$ by $L(p^+) = (L(p))^+$.

Definition 4 *A class of languages $\mathcal{L}$ has* finite thickness *if for each w there are only finitely many $L \in \mathcal{L}$ such that $w \in L$.*

Theorem 1 *$RREG^{\Sigma}_{+}$ has finite thickness.*

Proof By induction on the (literal) length of w. We will use the easily checked fact that $X^{+} = X^{++}$.

1. w has length 1, $w = a \in \Sigma$. Note first that $a \notin XY$ for any X, Y, since any element of XY has at least length 2. Also, $a \in X \Leftrightarrow a \in X^{+}$. So, if $a \in L$, then L has to be $L(a)$ or $L(a^{+})$.

2. w has length >1. w can only be a member of L^{+} if w is obtained by repetition of some smaller subsequence u of w (ignoring the case that $w \in L$). There can only be finitely many of such u, and to each of those the induction hypothesis applies.

   w can only be a member of $L_1 L_2$ if $w = uv$, $u \in L_1$, $v \in L_2$. Of course, w can only be split into such u and v in finitely many ways, and to each such u, v the induction hypothesis applies.

□

We can conclude that $RREG^{\Sigma}_{+}$ is identifiable. We we will see later that the same thing does not hold for $RREG^{\Sigma}_{*}$.

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

Proof We will assume that $\mathcal{L}$ is enumerated by a 1-1 recursive function. Let t be a text for L_i. Note that, under the assumption of 1-1 enumeration, there exists n_0 such that, for all j with $t(0) \in L_j$, $j \leq n_0$.

Define $M(t[n]) =$ the standard code for $L_{j_1} \cap \cdots \cap L_{j_k}$ where $L_{j_1}, \ldots, L_{j_k}$ are the only languages with $j_i \leq n$ such that $content(t[n]) \subseteq L_{j_m}$ $(m \leq k)$.

Note that, if $n \geq n_0$ no new languages occur. Note also that, if $n \geq n_0$, $j_m \leq n_0$ for all m. Note also that, if $L_i \not\subseteq L_j$, then at a certain point $content(t[n]) \not\subseteq L_j$ so that L_j will be dropped. So, from acertain poiunt onwards $L_{j_1}, \ldots, L_{j_k} = \{L_j \mid L_i \subseteq L_j\}$. That means that $L_{j_1} \cap \cdots \cap L_{j_k} = L_i$, i.e. $M(t[n]$ is a code for L_j. L_j is identified in the limit. □

Definition 5 *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 3 *If a uniformly recursive class $\mathcal{L}$ has recursive finite thickness, then $\mathcal{L}$ is identifiable by an effective incremental learner.*

The proof uses the canonical indices for finite sets. We assume that these are given in such a way that there are recursive functions U, I such that $D_{U(k,l)} = D_k \cup D_l$, $D_{I(k,l)} = D_k \cap D_l$, a recursive function $card$ such that $card(k) =$ the cardinality of D_k, that $n \in D_k$ is recursive, that *singleton* is a recursive function such that $D_{singleton(k)}$ is the set containing only k etc., etc.