Inferring Descriptive Generalisations of Formal Languages

Dominik D. Freydenberger[1] Daniel Reidenbach[2]

[1]Goethe University, Frankfurt

[2]Loughborough University, Loughborough

COLT 2010

Introduction

Our goal:

Learning patterns common to a set of strings.

- **pattern**: word consisting of **terminals** ($\in \Sigma$) and **variables** ($\in X$)
- $\mathrm{Pat}_\Sigma := (\Sigma \cup X)^+$: set of all patterns over Σ
- **substitution**: terminal-preserving morphism $\sigma : \mathrm{Pat}_\Sigma \to \Sigma^*$ $(\forall a \in \Sigma : \sigma(a) = a)$
- **language of a pattern** $\alpha \in \mathrm{Pat}_\Sigma$: set of all images of α under substitutions (write: $L(\alpha)$)

Example

$$\begin{aligned} L_{\mathrm{NE},\Sigma}(x\,\mathtt{a}\,y\,x) &= \{v\,\mathtt{a}\,w\,v \mid v, w \in \Sigma^+\}, \\ L_{\mathrm{E},\Sigma}(x\,\mathtt{a}\,y\,x) &= \{v\,\mathtt{a}\,w\,v \mid v, w \in \Sigma^*\}. \end{aligned}$$

The classical model

Identification in the limit of indexed families from positive data (Gold '67)

- **indexed family (of recursive languages)**: $\mathcal{L} = (L_i)_{i\in\mathbb{N}}$, where $w \in L_i$ is uniformly decidable
- **text** of a language L: a total function $t : \mathbb{N} \to \Sigma^*$ with $\{t(i) \mid i \in \mathbb{N}\} = L$
- set of all texts of L: $\mathrm{text}(L)$
- $\mathcal{L} \in \mathrm{LIM\text{-}TEXT}$ if there exists a computable function S such that, for every i and for every $t \in \mathrm{text}(L_i)$, $S(t^n)$ converges to a j with $L_j = L_i$

- NE-patterns (yes, Angluin '80)
- E-patterns (not if $|\Sigma| \in \{2, 3, 4\}$, Reidenbach '06, '08)
- terminal-free E-patterns (only if $|\Sigma| \neq 2$, Reidenbach '06)

Descriptive patterns

Definition

- Let $\mathcal{P}_\Sigma$ be a class of pattern languages over Σ.
- A pattern δ is **$\mathcal{P}_\Sigma$-descriptive** of a language L if
  1. $L(\delta) \in \mathcal{P}_\Sigma$,
  2. $L(\delta) \supseteq L$,
  3. there is no $L(\gamma) \in \mathcal{P}_\Sigma$ with $L(\delta) \supset L(\gamma) \supseteq L$.
- We write: $\delta \in D_{\mathcal{P}_\Sigma}(L)$

In other words: $L(\delta)$ is (one of) the closest generalisation(s) of L in $\mathcal{P}_\Sigma$, and δ is (one of) the best description(s) of L.

Our approach:

Learning of such generalisations.

Inferring descriptive generalisations

Definition

- Let $\mathcal{P}_\Sigma$ be a class of pattern languages over Σ.
- Let $\mathcal{L}$ be a class of nonempty languages over Σ.
- $\mathcal{L}$ can be **$\mathcal{P}_\Sigma$-descriptively generalised ($\mathcal{L} \in \mathrm{DG}_{\mathcal{P}_\Sigma}$)** if there is a computable function S such that, for every $L \in \mathcal{L}$ and for every $t \in \mathrm{text}(L)$, $S(t^n)$ converges to a $\delta \in D_{\mathcal{P}_\Sigma}(L)$.

Main conceptual differences to $\mathrm{LIM\text{-}TEXT}$:

- Infer generalisations instead of exact descriptions of the languages.
- Choose hypothesis space separate from language class.

Interesting phenomenon:

- one language can have several descriptive patterns,
- one pattern can be descriptive of several languages.

Characterisation theorem (for indexed families)

Theorem

Let Σ be an alphabet, let $\mathcal{L} = (L_i)_{i \in \mathbb{N}}$ be an indexed family over Σ, and let $\mathcal{P}_\Sigma$ be a class of pattern languages. $\mathcal{L} = (L_i)_{i \in \mathbb{N}} \in \mathrm{DG}_{\mathcal{P}_\Sigma}$ if and only if there are effective procedures d and f satisfying the following conditions:

(i) *For every $i \in \mathbb{N}$, there exists a $\delta_{d(i)} \in D_{\mathcal{P}_\Sigma}(L_i)$ such that d enumerates a sequence of patterns $d_{i,0}, d_{i,1}, d_{i,2}, \dots$ satisfying, for all but finitely many $j \in \mathbb{N}$, $d_{i,j} = \delta_{d(i)}$.*

(ii) *For every $i \in \mathbb{N}$, f enumerates a finite set $F_i \subseteq L_i$ such that, for every $j \in \mathbb{N}$ with $F_i \subseteq L_j$, if $\delta_{d(i)} \notin D_{\mathcal{P}_\Sigma}(L_j)$, then there is a $w \in L_j$ with $w \notin L_i$.*

- d is an enumeration of an appropriate subset of the hypothesis space
- f is similar to Angluin's telltales

Remarks

- Characterisation shows significant connection to Angluin's characterisation of indexed families in $\mathrm{LIM\text{-}TEXT}$.
- Main differences:
    1. our model requires an enumeration of a subset of the hypothesis space,
    2. we do not need to distinguish all L_i, L_j with $L_i \neq L_j$,
    3. the strategy in our proof might discard a correct hypothesis.
- Our strategy does not test membership or inclusion of pattern languages, but only membership for the indexed family.

Further topics

Further directions in our paper:

1. More general: Inductive inference with hypotheses validity relation (model HYP).
2. Less general: Consider a smaller class of patterns and a fixed strategy.

Inferring $\mathrm{ePAT}_{\mathrm{tf},\Sigma}$-descriptive patterns

- $\mathrm{ePAT}_{\mathrm{tf},\Sigma}$: The class of all E-pattern languages that are generated from **terminalfree** patterns.
- inclusion for $\mathrm{ePAT}_{\mathrm{tf},\Sigma}$ is well understood and decidable.
- strategy Canon: For every finite set S, return the pattern $\delta \in D_{\mathrm{ePAT}_{\mathrm{tf},\Sigma}}(S)$ that is minimal w.r.t. the length-lexicographical order.
- **telling set of L**: A finite set $T \subseteq L$ with $D_{\mathrm{ePAT}_{\mathrm{tf},\Sigma}}(T) \cap D_{\mathrm{ePAT}_{\mathrm{tf},\Sigma}}(L) \neq \emptyset$.

Theorem

Let Σ be an alphabet with $|\Sigma| \geq 2$. For every language $L \subseteq \Sigma^*$, and every text $t \in \mathrm{text}(L)$, Canon converges correctly on t if and only if L has a telling set.

Telling set languages

- $\mathcal{TSL}_\Sigma$: the class of all languages over Σ that have a telling set
- $\mathcal{TSL}_\Sigma \in \mathrm{DG}_{\mathrm{ePAT}_{\mathrm{tf},\Sigma}}$, using Canon as strategy

Some properties of $\mathcal{TSL}_\Sigma$:

- contains every DTF0L language $\Rightarrow$ superfinite
- is not countable
- does not contain all of REG
- contains all $\mathrm{ePAT}_{\mathrm{tf},\Sigma}$-languages (if $|\Sigma| \neq 2$)
- does not contain all $\mathrm{ePAT}_{\mathrm{tf},\Sigma}$-languages (if $|\Sigma| = 2$)