Distinguishing Pattern Languages With Membership Examples

Zeinab Mazadi, Ziyuan Gao, and Sandra Zilles
University of Regina, Canada

March 2014

University of Regina

Pattern languages

. . . in text mining

sample data entry:

name: Bugs Bunny; birth date: 30/04/1938; customer number: 127630041938

underlying pattern:

name: x_1; birth date: $x_2/x_3/x_4$; customer number: $x_5x_2x_3x_4$

Intuitively,

- a pattern consists of variable symbols and fixed symbols;
- a string matched by a pattern can be obtained by replacing the variables in the pattern with strings.

Note that variables may be repeated in a pattern.

Pattern languages

Definition

- $\Sigma = \{a, b, ...\}$ set of terminal symbols
- $X = \{x_1, x_2, ...\}$ set of variables such that $\Sigma \cap X = \emptyset$

Definition [Angluin 1980]

A pattern is any finite string over terminal symbols and variables. The language $L(\pi)$ of a pattern π is the set of all words that result from substituting all variables in π by *nonempty* strings of terminal symbols.

Example: $\Sigma = \{a, b, c\}$

$\pi =$	aab	x_1	x_2	bc	x_1	$babc$
$L(\pi) \ni$	aab	cbb	ba	bc	cbb	$babc$

Pattern languages

. . . and their variants in the (recent) literature

- computational learning theory:
  - learning extensions of pattern languages [Geilke, Z 2011; Geilke, Z 2012]
  - novel models of learning pattern languages [Freydenberger, Reidenbach 2013; Geilke, Z 2011]
- formal language theory:
  - embedding pattern languages into the Chomsky hierarchy [Jain, Ong, Stephan 2010; Reidenbach, Schmid 2012a]
  - complexity of the membership problem [Fernau, Schmid 2013; Geilke, Z 2011; Reidenbach, Schmid 2012b]
  - decision problems on comparing pattern languages [Freydenberger, Reidenbach 2010]
- applications:
  - bioinformatics [e.g., Arikawa et al. 1993]
  - automatic program synthesis [Nix 1985]

University of Regina

Central questions of our paper

membership example for a language L:

$$(w, \ell)$$

where $w \in \Sigma^*$ and ℓ signals whether $w \in L$ ($\ell = 1$) or $w \notin L$ ($\ell = 0$)

Question:

Given some class $\mathcal{L}$ of pattern languages, how many membership examples are needed in the worst case (over all $L \in \mathcal{L}$) to distinguish a language L from all other languages in $\mathcal{L}$?

Does this number depend on the alphabet size?

University of Regina

Teaching dimension (TD)

Definition [Goldman, Kearns 1995; Shinohara, Miyano 1991]

Let $\mathcal{L}$ be any class of languages over Σ^*. Let $L \in \mathcal{L}$.
$\mathrm{TD}(L, \mathcal{L})$, the teaching dimension of L w.r.t. $\mathcal{L}$ is the smallest number m such that there is a set of m membership examples for L that is not consistent with any other $L' \in \mathcal{L}$. Then

$$\mathrm{TD}(\mathcal{L}) = \sup_{L \in \mathcal{L}} \mathrm{TD}(L, \mathcal{L}) .$$

	ϵ	a	b	aa	ab	ba	$\cdots$	$\mathrm{TD}(L, \mathcal{L})$
L_{empty}	0	0	0	0	0	0	0	∞
L_1	1	0	0	0	0	0	0	1
L_2	0	1	0	0	0	0	0	1
L_3	0	0	1	0	0	0	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\mathrm{TD}(\mathcal{L}) = \infty$

University of Regina

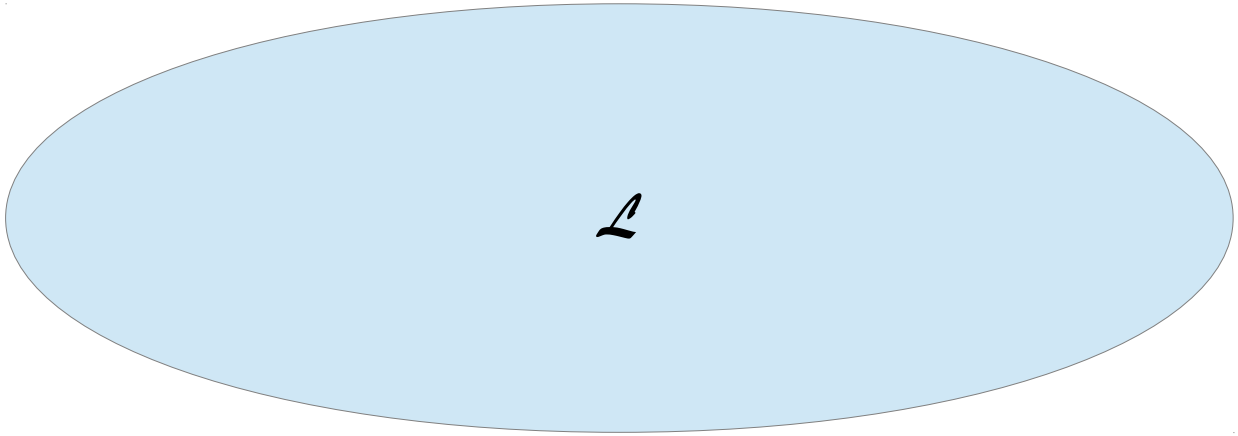
Teaching dimension (TD)

	ϵ	a	b	aa	TD$(L, \mathcal{L})$
L_{empty}	0	0	0	0	4
L_1	1	0	0	0	4
L_2	0	1	0	0	4
L_3	0	0	1	0	4
L_4	0	0	0	1	4
L_{12}	1	1	0	0	2
L_{13}	1	0	0	1	2
L_{14}	1	0	0	1	2
L_{23}	0	1	1	0	2
L_{24}	0	1	0	1	2
L_{34}	0	0	1	1	2
					TD$(\mathcal{L}) = 4$

Recursive teaching dimension (RTD)

[Z, Lange, Holte, Zinkevich 2011]

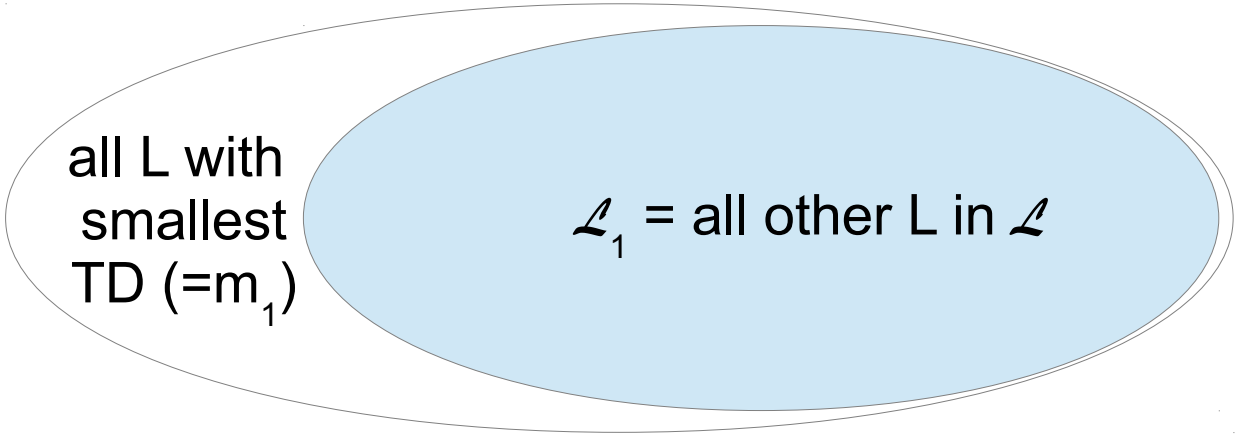
the recursive teaching dimension of $\mathcal{L}$:



Recursive teaching dimension (RTD)

[Z, Lange, Holte, Zinkevich 2011]

the recursive teaching dimension of $\mathcal{L}$:

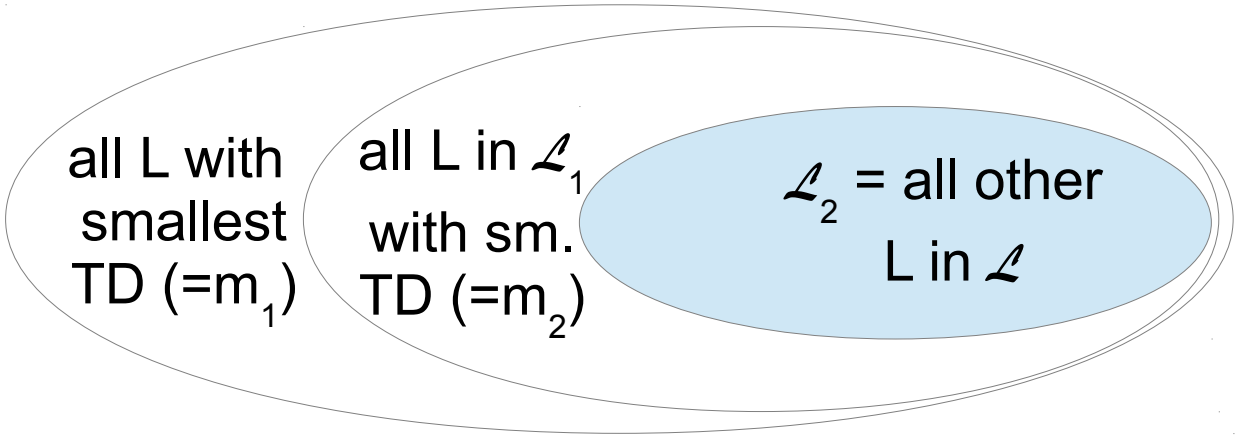


University of Regina

Recursive teaching dimension (RTD)

[Z, Lange, Holte, Zinkevich 2011]

the recursive teaching dimension of $\mathcal{L}$:

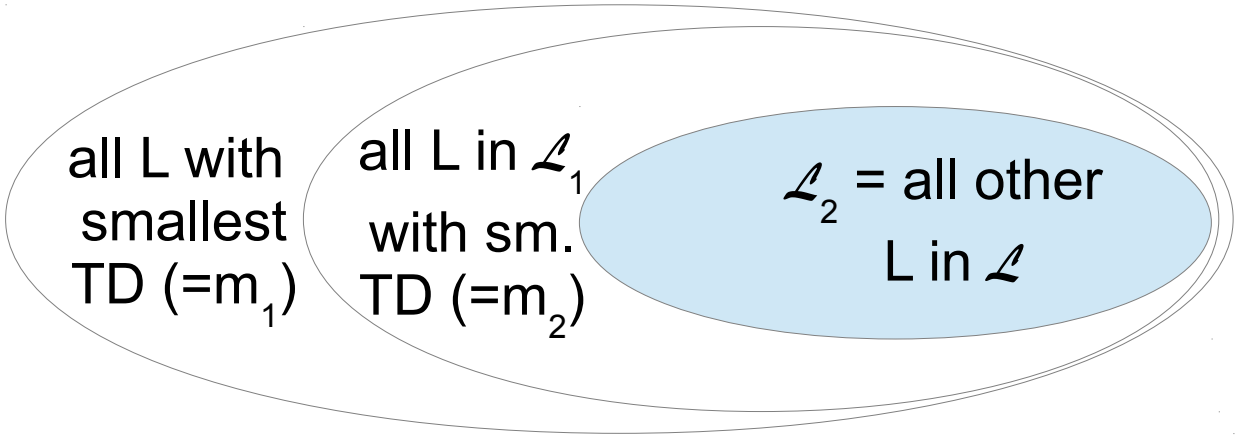


University of Regina

Recursive teaching dimension (RTD)

[Z, Lange, Holte, Zinkevich 2011]

the recursive teaching dimension of $\mathcal{L}$:



proceed recursively ... $\rightsquigarrow$ $\mathrm{RTD}(\mathcal{L}) = \sup_i m_i$

University of Regina

Recursive teaching dimension (RTD)

[Z, Lange, Holte, Zinkevich 2011]

	ϵ	a	b	aa	ab	ba	$\cdots$	TD$(L, \mathcal{L})$	RTD$(L, \mathcal{L})$
L_{empty}	0	0	0	0	0	0	0	∞	0
L_1	1	0	0	0	0	0	0	1	1
L_2	0	1	0	0	0	0	0	1	1
L_3	0	0	1	0	0	0	0	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	TD$(\mathcal{L}) = \infty$	RTD$(\mathcal{L}) = 1$

Recursive teaching dimension (RTD)

	ϵ	a	b	aa	TD$(L, \mathcal{L})$	RTD$(L, \mathcal{L})$
L_{empty}	0	0	0	0	4	0
L_1	1	0	0	0	4	1
L_2	0	1	0	0	4	1
L_3	0	0	1	0	4	1
L_4	0	0	0	1	4	1
L_{12}	1	1	0	0	2	2
L_{13}	1	0	0	1	2	2
L_{14}	1	0	0	1	2	2
L_{23}	0	1	1	0	2	2
L_{24}	0	1	0	1	2	2
L_{34}	0	0	1	1	2	2
					TD$(\mathcal{L}) = 4$	RTD$(\mathcal{L}) = 2$

RTD related to other important complexity parameters in learning theory [Doliwa, Simon, Z 2010; Darnstädt, Doliwa, Simon, Z 2013]

University of Regina

Central questions of our paper

Given some structurally interesting class $\mathcal{L}$ of pattern languages, what are the values for TD($\mathcal{L}$) and RTD($\mathcal{L}$)?

Do these values depend on the alphabet size?

The class of all pattern languages

For the class of all pattern languages:

- $\mathrm{TD}(L, \mathcal{L})$ is finite for any Σ and any L:
  L generated by a pattern of length k
  $\Rightarrow$ L contains a word w of length k
  $\Rightarrow$ only finitely many pattern languages $L_1, \ldots, L_j$ contain w
  $\Rightarrow$ some set $\{(w, 1), (w_1, \ell_1), \ldots, (w_j, \ell_j)\}$ uniquely determines L,
  where $w_i \in L \Delta L_i$
- $\mathrm{TD}(\mathcal{L}) = \infty$ for any Σ
  (no upper bound on $\mathrm{TD}(L, \mathcal{L})$ for $L \in \mathcal{L}$)
- $\mathrm{RTD}(\mathcal{L}) = 2$ for infinite Σ
- $\mathrm{RTD}(\mathcal{L})$ unknown for finite Σ

$\Rightarrow$ for infinite alphabets, $\mathrm{TD}(\mathcal{L}) = \infty$ and $\mathrm{RTD}(\mathcal{L}) = 2$

$\mathrm{TD}(\mathcal{L}) = \infty$ and $\mathrm{RTD}(\mathcal{L}) = 2$ ...

... also holds for a very restricted class of pattern languages over finite non-singleton alphabets:

Observation

Let $2 \leq |\Sigma| < \infty$ and let $\mathcal{L}$ be the class of all languages generated by patterns of the form

$$x_1 \cdots x_k \text{ for } k \geq 1$$

or of the form

$$x_1 \cdots x_i \cdots x_{j-1} x_i x_j \cdots x_k \text{ for } 1 \leq i < j \leq k\,.$$

Then $\mathrm{TD}(\mathcal{L}) = \infty$ and $\mathrm{RTD}(\mathcal{L}) = 2$.

$\mathrm{TD}(\mathcal{L}) = \infty$ and $\mathrm{RTD}(\mathcal{L}) = 2$ ...

... also holds for a very restricted class of pattern languages over finite non-singleton alphabets:

Observation

Let $2 \leq |\Sigma| < \infty$ and let $\mathcal{L}$ be the class of all languages generated by patterns of the form

$$x_1 \cdots x_k \text{ for } k \geq 1$$

or of the form

$$x_1 \cdots x_i \cdots x_{j-1} x_i x_j \cdots x_k \text{ for } 1 \leq i < j \leq k\,.$$

Then $\mathrm{TD}(\mathcal{L}) = \infty$ and $\mathrm{RTD}(\mathcal{L}) = 2$.

Here we use:

- unbounded number of distinct variables
- repetition of a variable

What if neither is allowed?

University of Regina

Sub-classes of pattern languages

- bounded number of distinct variables OR
- no repetitions of variables

Definition [Angluin 1980, Shinohara 1982]

A pattern is called

- one-variable if it contains at most one variable (possibly multiple times),
- regular if it has no repeated variables.

Languages generated by one-variable (regular) patterns are called one-variable (regular) pattern languages.

Summary of Results

	$2 \leq \lvert\Sigma\rvert \leq \infty$	$\lvert\Sigma\rvert = 1$	$\lvert\Sigma\rvert = \infty$
arbitrary patterns	$TD = \infty$ $RTD \geq 2$	$TD = \infty$ $RTD \geq 2$	$TD = \infty$ $RTD = 2$
one-variable patterns	$TD = \infty$ $RTD = 2$	$TD = \infty$ $RTD = 2$	$TD = \infty$ $RTD = 2$
regular patterns	$TD \geq 5$ $RTD = 2$	$TD = 3$ $RTD = 2$	$TD = 5$ $RTD = 2$

To show that TD$=\infty$ for the class $\mathcal{L}$ of all one-variable pattern languages:

Suppose TD$= m-1$.
Let $k = p_1 \cdot \ldots \cdot p_m$ for m pairwise distinct primes $p_1, \ldots, p_m$ and $\pi = x^k$.
Some set S of $m-1$ examples uniquely determines $L(\pi)$ in $\mathcal{L}$.

$L(\pi_1), \ldots, L(\pi_m)$, for $\pi_i = x^{k/p_i}$, are m pairwise distinct proper supersets of $L(\pi)$.

For $1 \leq i \leq m$, S contains $(w_i, 0)$ for some $w_i \in L(\pi_i) \setminus L(\pi)$.

$\exists i, j$: $i \neq j$ and $w_i = w_j$, in contradiction to $L(\pi_i) \cap L(\pi_j) = L(\pi)$.

Suppose TD$= 2$, i.e., $m = 3$.
Consider $k = 2 \cdot 3 \cdot 5 = 30$ and $\pi = x^{30}$.

Some set S of size 2 uniquely determines $L(x^{30})$ in $\mathcal{L}$.

$L(x^{15})$, $L(x^{10})$, $L(x^6)$ are pairwise distinct proper supersets of $L(x^{30})$.

S contains $w_1 \in L(x^{15}) \setminus L(x^{30})$,
$w_2 \in L(x^{10}) \setminus L(x^{30})$,
$w_3 \in L(x^6) \setminus L(x^{30})$.

Some two w_i must be equal since $|S| = 2$ ... contradiction.

Conclusions

- in all resolved cases, RTD $<$ TD
- alphabet sometimes affects TD and RTD values
- proof techniques (and resulting “teaching sets”) often vary with varying alphabet size

  $\Rightarrow$ insights into structural properties of language families