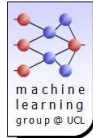


Outline

- 1 DFA induction
- 2 Feature selection

Two combinatorial optimization problems in machine learning

Pierre.Dupont@uclouvain.be



ICTEAM Institute
Université catholique de Louvain – Belgium

May 17, 2011

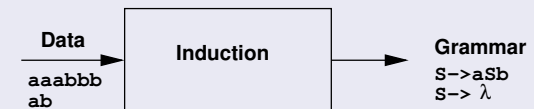
Talk objectives

- Describe 2 “simple” ML problems and formulate them as **constrained combinatorial optimization problems**
 - 1 DFA induction
 - 2 Feature selection
- Trigger discussions to see whether a **CP** approach may help to **better address them**
 - ▶ and possibly bootstrap from ML to CP

Grammar induction

Also known as grammatical inference

- **Grammar induction** is about **learning a formal grammar** from a set of positive strings from its language, and possibly negative strings as well

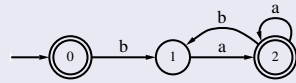


- The positive and negative strings form a **learning sample** and the grammar, or an alternative representation, **generalizes it**
- Often a **simplest generalization** is sought (Ockham's razor formalized in computational learning theory)
- Probabilistic extensions and statistical estimation algorithms have been proposed

The minimal DFA consistency problem

- Learning a **regular language** is the most studied case
 - Interesting applications in computational biology, natural language processing, software engineering, ...
- A regular language can be equivalently represented by a canonical **Deterministic Finite-state Automaton** (DFA)

$$L = (ba^*a)^* \quad \left\{ \begin{array}{l|l} S \rightarrow \lambda & bA \\ A \rightarrow a & aB \\ B \rightarrow aB & bA \end{array} \right.$$



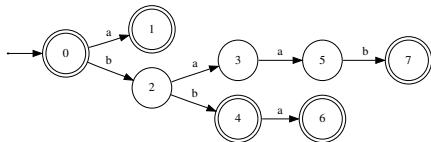
- Learning a regular language \Rightarrow solving the minimal DFA consistency problem

Note: theoretical results show that solving this NP-hard problem on a growing learning sample leads to the correct language identification in finite time [5, 1]

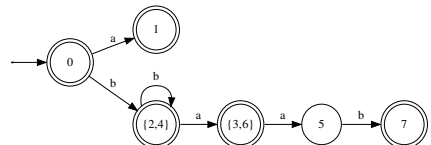
Language generalization through state merging

$$S_+ = \{\lambda, a, baab, bb, bba\} \quad S_- = \{b, ab\}$$

Prefix-Tree Acceptor (PTA)



Quotient automaton



Under reasonable assumptions, the target machine is in the PTA partition set [4]

State-merging DFA induction algorithm

Algorithm STATE-MERGING DFA INDUCTION ALGORITHM

Input: A positive and negative sample (S_+, S_-)

Output: A DFA A consistent with (S_+, S_-)

// Compute a **PTA**, let N denote the number of its states

$PTA \leftarrow$ Initialize(S_+); $\pi \leftarrow \{\{0\}, \{1\}, \dots, \{N-1\}\}$

// Main state-merging loop

while (B_i, B_j) \leftarrow ChoosePair(π) **do**

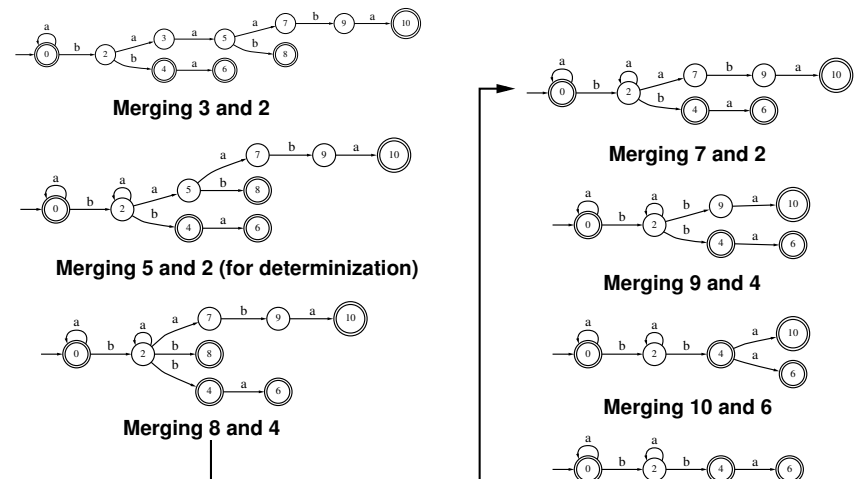
$\pi_{new} \leftarrow$ Merge(π, B_i, B_j)

if Compatible($PTA/\pi_{new}, S_-$) **then**

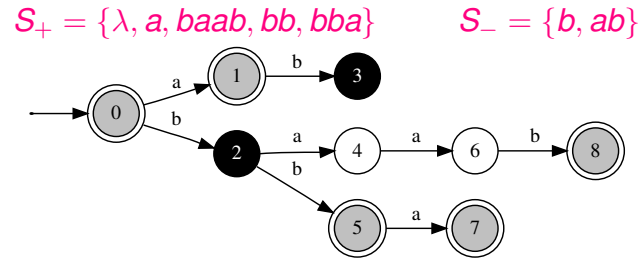
$\pi \leftarrow \pi_{new}$

return PTA/π

The Merge function also reduces non-determinism



An alternative representation: graph coloring problem



- Augmented PTA with **positively accepting states** (= grey) and **negatively accepting states** (= black)
- The Merge function reduces non-determinism **and** checks such **coloring constraints**
 - States having different colors **may not** be merged

A minimal graph coloring problem

Find a deterministic graph with a minimal number of nodes satisfying the coloring constraints [2, 3]

THE STAMINA COMPETITION

Learning regular languages with large alphabets

[any question?](#) [create an account](#) [login](#)



[Home](#) [Protocol](#) [Participate](#) [Download](#) [Baseline](#) [Scientific Committee](#)

This page describes the competition protocol in details. It also provides information about the **target machines**, the **training and test samples** and the **evaluation** of submitted test results.

Challenge

Solve a grid of DFA induction problems for **increasing alphabet size** and **decreasing learning sample sizes** [14]

5 PROBLEMS TO SOLVE IN EACH CELL

Alphabet size	Sparsity of the training sample			
	100%	50%	25%	12.5%
2	1 - 5	6 - 10	11 - 15	16 - 20
5	21 - 25	26 - 30	31 - 35	36 - 40
10	41 - 45	46 - 50	51 - 55	56 - 60
20	61 - 65	66 - 70	71 - 75	76 - 80
50	81 - 85	86 - 90	91 - 95	96 - 100

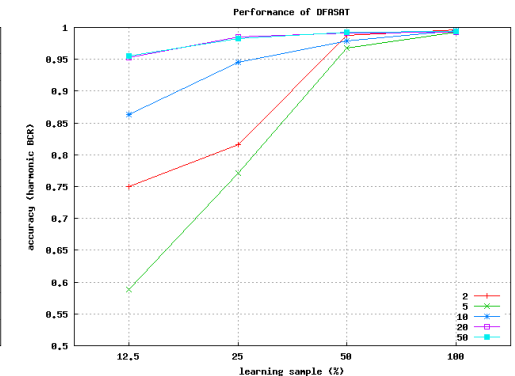
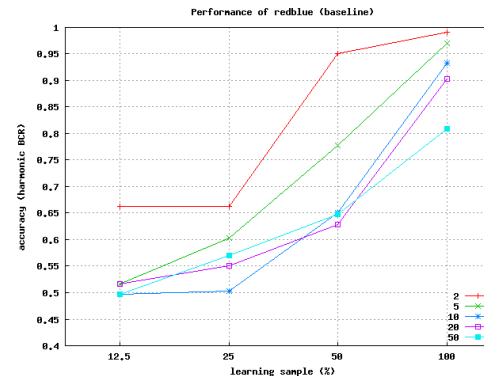
Note: the competition is over but you can still try to outperform the winning algorithm!

stamina.chefbe.net

The Stamina winner

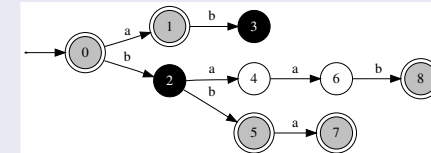
The winning algorithm DFASAT

- from Marijn Heule and Sicco Verwer (Delft, Leuven)
- a preliminary version was proposed in [9]
- a mix of state merging + graph coloring + reduction to SAT



Why CP looks interesting to tackle this problem?

- This combinatorial optimization problem is a **CSP**
- DFASAT uses some **ingredients** also found in **CP**
 - A dedicated problem representation
 - Redundant clauses
 - Symmetry
- Constraint propagation** is natural in this problem



$\{2, 8\}$ implies $\{0, 6\}$ incompatibility!

- Some form of **cannot-link constraints** on a graph structure + determinism
- Mandatory merge (must link) constraints have also been proposed [10]

Why is it challenging?

The search space is the **partition set** of the PTA state set

A concrete example from Stamina:

- $\approx 15,000$ positive and negative learning strings
- The augmented PTA has $\approx 50,000$ states
- The number of partitions of a set of m elements into k non empty subsets is $\approx O\left(\frac{k^m}{k!}\right)$ (exact computation through a Stirling number)
 $m = 50,000$; $k = 50 \approx 10^{10^5}$ machines with 50 states
 (and yet, one should search other target sizes as well)

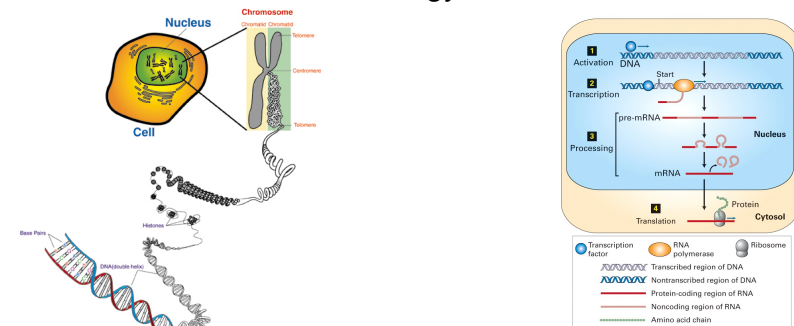
The **more learning data** you get the **larger the search space** while it **should be simpler** from a ML viewpoint

Outline

- 1 DFA induction
- 2 Feature selection

A motivating example

Molecular biology in **one** slide!

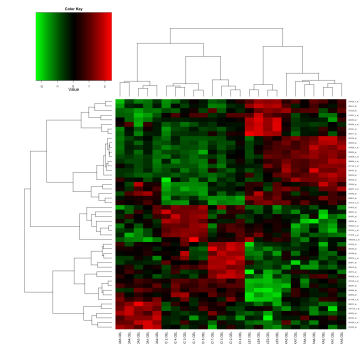
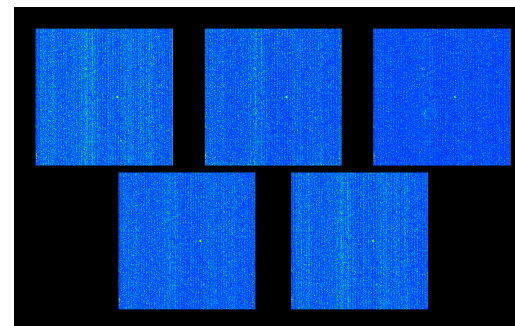


Gene expressions **vary** for many different, but possibly dependent, reasons:

- chemical or physical environment of the cells,
- growth of the organism,
- regulation of complex metabolic processes,
- susceptibility to develop some illness or to respond to a treatment,...

DNA microarrays

- DNA chips measure the level of expression of **all genes** in a **single experiment**



Gene selection: a machine learning viewpoint

	gene 1	gene 2	...	gene d	class label
sample 1	$x_{1,1}$	$x_{1,2}$...	$x_{1,d}$	y_1
sample 2	$x_{2,1}$	$x_{2,2}$...	$x_{2,d}$	y_2
...
sample n	$x_{n,1}$	$x_{n,2}$...	$x_{n,d}$	y_n

- The number d of **genes** or **probe sets** $\approx 55,000$
- The number n of **samples** (tissues, patients) ≤ 100
- The class labels y come from clinical status:
responsive or not to treatment,
good or bad diagnosis/prognosis, type of pathology, ...

Biomarker Selection

Find a small subset of (≈ 50) genes to predict the outcome y of **new samples** $\Rightarrow C_{55,000}^{50} \approx 3.10^{172}$ possibilities...

A basic feature selection: t-Test relevance index

Many feature selection methods have been proposed [6, 13]

A basic approach

- Compute the **mean feature values** in each class
- Assess whether the means **significantly differ** between classes
 - ▶ select the top ranked features according to the p-values of a t-Test

Issues with t-Test selection

- you need **reliable estimates** of the **means** and **variances** of each feature (hard with few samples)
- you need to correct for **multiple testing**
- the selection is **univariate**
 \Rightarrow the **dependencies** between features are not considered

Maximum relevance minimum redundancy [12]

Multivariate objective

- Find a subset $S \subseteq X$ of k **maximally relevant** and **minimally redundant** features
- **Relevance** can be measured by the mutual information with the response $I(S; Y)$
- **Redundancy** can be measured by the mutual information between variables $I(S_1, \dots, S_k)$

Notes:

- Mutual information is difficult to estimate in high dimensions but approximations or alternative measures (e.g. rank correlation) exist
- The mRmR approach uses a greedy search to optimize this objective

Question: Could CP help to better solve this CSP?

If standard feature selection looks too simple to you

Questions

- How to use some uncertain and partial **prior knowledge** about relevant features? or about feature dependencies?
- How to select **common** features on **distinct** but related **tasks** (transfer or multi-task learning)?

Note: mathematical programming approaches have been proposed to address those problems [8, 7, 11]. Can CP complement or outperform those methods?

Conclusions

- DFA induction and feature selection can be formulated as CSPs
- These are combinatorial optimization problems with very large search spaces, even with relatively small learning samples
- Those problems could trigger additional collaborations between CP and ML

Some references I

- [1] D. Angluin, *On the complexity of minimum inference of regular sets*, Information and Control **39** (1978), 337–350.
- [2] F. Coste and J. Nicolas, *How considering incompatible state mergings may reduce the DFA induction search tree*, Grammatical Inference, ICGI'98 (Ames, Iowa), Lecture Notes in Artificial Intelligence, no. 1433, Springer Verlag, 1998, pp. 199–210.
- [3] P. Dupont, B. Lambeau, C. Damas, and A. van Lamsweerde, *The QSM algorithm and its application to software behavior model induction*, Applied Artificial Intelligence **22** (2008), 77–115.
- [4] P. Dupont, L. Miclet, and E. Vidal, *What is the search space of the regular inference ?*, Grammatical Inference and Applications, ICGI'94 (Alicante, Spain), Lecture Notes in Artificial Intelligence, no. 862, Springer Verlag, 1994, pp. 25–37.

Some references II

- [5] E.M. Gold, *Complexity of automaton identification from given data*, Information and Control **37** (1978), 302–320.
- [6] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh (eds.), *Feature extraction, foundations and applications*, Series Studies in Fuzziness and Soft Computing, vol. 207, Springer, 2006.
- [7] T. Helleputte and P. Dupont, *Feature selection by transfer learning with linear regularized models*, European Conference on Machine Learning, Lecture Notes in Artificial Intelligence, no. 5781, 2009, pp. 533–547.
- [8] T. Helleputte and P. Dupont, *Partially supervised feature selection with regularized linear models*, International Conference on Machine Learning, 2009.

Some references III

-  [9] M. Heule and S. Verwer, *Exact DFA identification using SAT solvers*, International Colloquium on Grammatical Inference, Lecture Notes in Artificial Intelligence, vol. 6339, Springer Verlag, 2010, pp. 66–79.
-  [10] B. Lambeau, C. Damas, and P. Dupont, *State-merging DFA induction algorithms with mandatory merge constraints*, Lecture Notes in Artificial Intelligence, vol. 5278, 2008, pp. 139–153.
-  [11] G. Obozinski, B. Taskar, and M.I. Jordan, *Joint covariate selection and joint subspace selection for multiple classification problems*, Statistics and Computing (2009), 1–22.

Some references IV

-  [12] H. Peng, F. Long, and C. Ding, *Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy*, IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005), no. 8, 1226–1238.
-  [13] Y. Saeys, I. Inza, and P. Larrañaga, *A review of feature selection techniques in bioinformatics*, Bioinformatics **23** (2007), no. 19, 2507–2517.
-  [14] N. Walkinshaw, K. Bogdanov, C. Damas, B. Lambeau, and P. Dupont, *A framework for the competitive evaluation of model inference techniques*, 1st International workshop on Model Inference In Testing, 2010.