Evolutionary Computation
A robut optimisation framework for ML

Michèle Sebag

TAO Group – Equipe Inférence et Apprentissage
Laboratoire de Recherche en Informatique
Université Paris-Sud, CNRS-UMR 8623 – Orsay
Michele.Sebag@lri.fr

Bari – Marzo 10, 2005

Part II : Feature Selection

Contents

- Motivations
- State of the art
- An (evolutionary) ensemble approach
  - A combinatorial optimization criterion, ROC
  - An application: Risk for Cardio Vascular Diseases
  - Exploiting diverse hypotheses for free with EC

Motivations

Before learning: find a representation of the data...

- Too poor representation — nothing can be learned
- Too detailed representation — feature selection needed

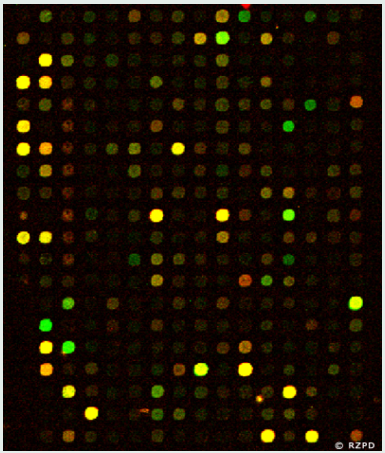
Why ?

- Machine Learning is not a well-posed problem
- $\Longrightarrow$ Adding irrelevant information can harm learning.

What is the goal: Feature Selection / Feature Construction ?

- Feature Construction: define relevant features
- ... = that enable learning
- but: Best features = good hypotheses...

When ML = Feature Selection



Bio-informatics

- 30 000 genes
- few examples (costly)
- Goal: finding the relevant genes

Position of the problem

Training set $\mathcal{E} = \{(\mathbf{x}_i, y_i), \mathbf{x}_i \in X = \mathbb{R}^d, y_i = \pm 1, i = 1..n\}$
Set of attributes $\mathcal{A} = \{a_1, ..a_d\}$

Goals

- Feature Selection: find a subset of $\{a_1, ..., a_d\}$
- Feature Ranking : find an order on $\{a_1, ..., a_d\}$

Formally

Given

$$\begin{aligned} \mathcal{F} : \mathcal{P}(\mathcal{A}) &\mapsto \mathbb{R} \\ A \subset \mathcal{A} &\mapsto Err(A) = \text{ min error of hypotheses based on } A \end{aligned}$$

Find $Argmin(\mathcal{F})$

Difficulties

- A combinatorial optimization problem (2^d)
- defined for an unknown $\mathcal{F}$ ...

Approaches

Filter — univariate methods

Define $score(a_i)$; iteratively, add the features maximizing $score$
or remove the features minimizing $score$

PRO: simple & tractable

CON: (very) local optima

Rk : backtracking $\rightarrow$ better optima, more expensive method

Wrapping — multivariate methods

Measure the quality of sets of features

estimate $\mathcal{F}(a_{i1}, \ldots a_{ik})$

CON: expensive: one estimation = one learning pb

PRO: better optima

Filter Approaches

Notations

Training set : $\mathcal{E} = \{(x_i, y_i), i = 1..n, y_i \in \{-1, 1\}\}$

$a(x_i)$ = value of feature a for example (x_i)

Information gain

decision trees

$$p([a = v]) = Pr(y = 1 | a(x_i) = v)$$

$$QI([a = v]) = -p \log p \; - (1 - p) \log (1 - p)$$

$$QI = \sum_v p(v) QI([a = v])$$

Correlation

$$corr(a) = \frac{\sum_i a(x_i).y_i}{\sqrt{\sum_i (a(x_i))^2 \; \times \; \sum_i y_i^2}} \propto \sum_i a(x_i).y_i$$

Wrapper Approaches

Generate and test

Given a list of candidates $\mathcal{L} = \{A_1, .., A_p\}$

– Generate candidate A

– Compute $\mathcal{F}(A)$

—— learn h_A from $\mathcal{E}_{|A}$

—— test h_A on some test set $= \hat{\mathcal{F}}(A)$

– Update $\mathcal{L}$.

Algorithms

- hill-climbing / multiple restart
- genetic algorithms Vafaie-DeJong, IJCAI 95
- (*) genetic programming & feature construction. Krawiec, GPEH 01

A posteriori Approaches

Principle

- Construct hypotheses
- Induce which are the relevant features
- Prune the (most) irrelevant features
- Iterate.

Algorithm : SVM Recursive Feature Elimination Guyon et al. 03

- Linear SVM $\rightarrow h(x) = sign(\sum w_i.a_i(x) + b)$
- If $|w_i|$ is small, a_i is not relevant
- Prune the k features with minimal absolute weight
- Iterate.

Limitations

Linear Hypotheses

- One weight per feature.

Amount of examples

- The features weights are not independent.

[algebraically, the weight vector lies
in the subspace defined from the examples

But FS is when the number of examples is insufficient...

An evolutionary approach: ROGER

Overview

- A combinatorial optimization formulation for ML
  tackled by Evolutionary Computation
- A real-world application: Risk for Cardio-Vascular Diseases
  and some remarks on the risk attached to tobacco and alcool...
- Defining more complex hypotheses
  i) non-linear; ii) allowing for scoring the features
- Exploit the variability of solutions provided by a stochastic optimization method (GA) $\Longrightarrow$ Ensemble learning.

ROC criterion
Receiver Operating Characteristics

Principle — signal processing, medical data analysis

Let $h(x)$ be the risk of patient x. $h : X \mapsto \mathbb{R}$

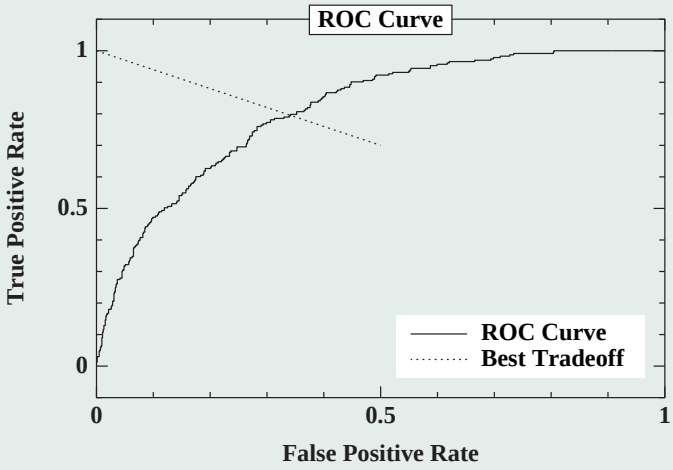
$$t \in \mathbb{R} \quad \mapsto \quad h_t(x) = \begin{cases} ill & if\ h(x) > t \\ OK & otherwise \end{cases}$$

For any h_t, let:

- TP(t) : true positive rate, $Pr(h_t(x) = ill | x\ ill)$
- FP(t) : false positive rate, $Pr(h_t(x) = ill | x\ not\ ill)$.

Plot the curve $(TP(t), FP(t), t \in \mathbb{R})$.

ROC Curve


ROC Curve
True Positive Rate
False Positive Rate
0
0.5
1
0
0.5
1
ROC Curve
Best Tradeoff

ROC Curve, 2

ROC depicts the trade-off False Positive / True Positive.

Standard: misclassification cost (Domingos, KDD 99)

$$\mathcal{F} = \#\text{ false positive} + c \times \#\text{ false negative}$$

In a multi-objective perspective, ROC = Pareto front.

Best solution: intersection of Pareto front with $\Delta(-c, -1)$

ROC: Extensively Used by Physicians

ROC Curve, 3

Used to compare learners — Bradley 97

multi-objective-like

insensitive to imbalanced distributions

shows sensitivity to error cost.

Used as learning criterion: Area under the ROC curve

Given Dataset = $\{(\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}\}$

Genotype: hypothesis $h \mapsto$ Phenotype: ordered examples

$+ + + - + + - + + + + - - - + - - - + - - - - - - - - - -$

$\mathcal{F}(h)$ = sum of ranks of positive examples. AUC : to be maximized

Area Under the ROC Curve

Previous

EP-based NN optimization	Fogel+, 1998
GA-based linear optimization	Mozer+, 2001
greedy Decision Tree optimization	Ferri-Flach, 2002

ROGER: ROC-based Genetic Evolutionary LearneR

$(\mu + \lambda)$-ES (Evolution Strategy)

Parameters

population size	# parents μ	10
	# offspring λ	50
max nb evaluations		10,000
crossover	uniform	rate .6
mutation	self-adaptive	rate 1

Experiments

Reference results: Support Vector Machines (SVMTorch)

Search space: linear classifiers : $\mathbb{R}^d$

Datasets from Irvine repository

	#att	#weight	#Train	#Test
Br. Canc.	9	42	189	97
Crx	15	47	70	620
German	25	25	100	900
Promoters	59	229	70	36
Satimage	36	36	139	1237
Vehicle	18	18	125	291
Votes	16	32	287	148
Waveform	22	22	211	3321

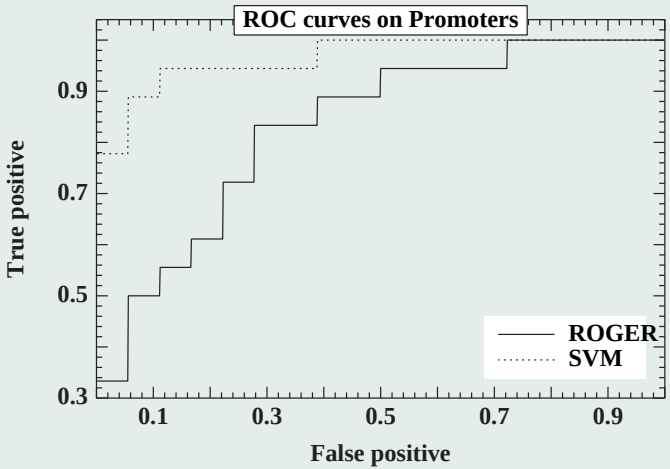
	ROGER		SVMTorch	
	AUC	time	AUC	time
Br. Canc.	$.674 \pm .05$	7”	$.672 \pm .05$	1”
Crx	$.816 \pm .06$	7”	$.839 \pm .04$	886”
German	$.712 \pm .03$	6”	$.690 \pm .02$	96”
Promoters	$.863 \pm .07$	2”	$.974 \pm .02$	< 1”
Satimage	$.918 \pm .01$	4”	$.876 \pm .02$	14”
Vehicle	$.994 \pm .005$	1”	$.993 \pm .007$	< 1”
Votes	$.993 \pm .004$	7”	$.989 \pm .005$	$> 1{,}000$
Waveform	$.971 \pm .004$	4”	$.963 \pm .008$	2”

Experimental setting

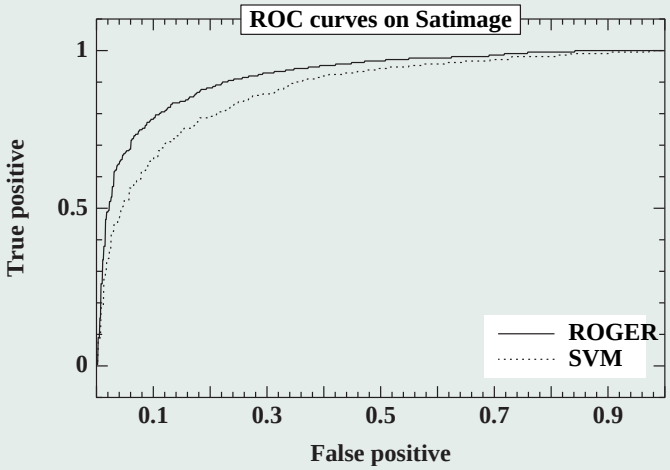
10 train/test splits

For each split, 1 SVMTorch run, 21 ROGER runs (take median)

ROC Curve, Promoters



ROC Curve, Satimage



Partial conclusions - ML aspects

PROS

- Competitive wrt state of art, SVM.
- Affordable cost, fitness computation $nlog(n)$
- Learning stability wrt imbalanced distribution, error cost

CONS

- Does not scale up well with # attributes

An evolutionary approach: ROGER

Overview

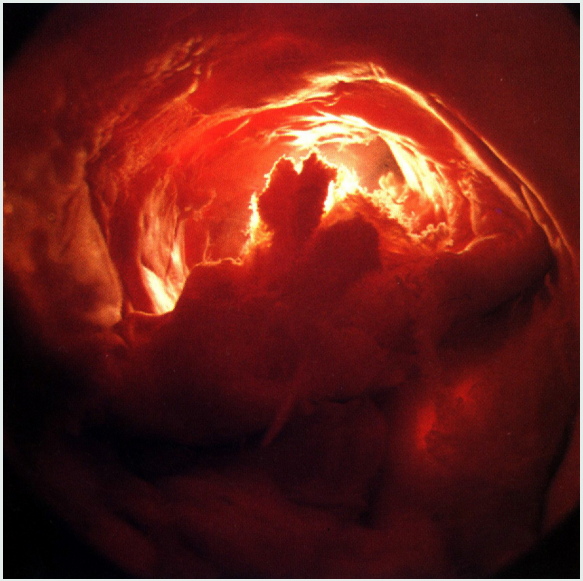
- A combinatorial optimization formulation for ML
  tackled by Evolutionary Computation

- A real-world application: Risk for Cardio-Vascular Diseases
  and some remarks on the risk attached to tobacco and alcool...

- Defining more complex hypotheses
  i) non-linear; ii) allowing for scoring the features

- Exploit the variability of solutions provided by a stochastic optimization method (GA) $\Longrightarrow$ Ensemble learning.

A Medical Data Mining Application



Understanding Cardio-vascular Diseases

PKDD 2002-2003 Challenge

- Study Atherosclerosis Risk Factors First death cause in Western countries

Data

- ENTRY database (medical cliché, 1419 men, 219 attributes, 1976)
- CONTROL database (longitudinal study of a sample, 1976-1996)

First goal

- Given the medical cliché at t_0, predict health state at $t_0 + 20$.

Some limitations of the data

Initial description :

very detailed — diseases 1st..4th brother, 1st..4th sister

...not usable... — 4th sister INF MYOCARD....

What cannot be learned : sufficient conditions for diseases

(1) If father or mother diabetic
(2) And high stress
(3) And does not laugh once a day
Then disease

... (Condition 3 likely missing in hospital db)

$\rightarrow$ find at best necessary conditions

Changing the problem

Initial goal: classification — predefined classes

$$\text{Patient} \mapsto \{ \text{normal, at risk, pathological} \}$$

Alternative: ranking — Mr X is more at risk than Ms Y

$$(\text{Patient} \times \text{Patient}) \mapsto \{true, false\}$$

concept is smoother (frontier between normal and pathological)
more flexible (medical / economical concerns)

Proposed: "underconstrained regression" — Risk(Mr X) is 3.7

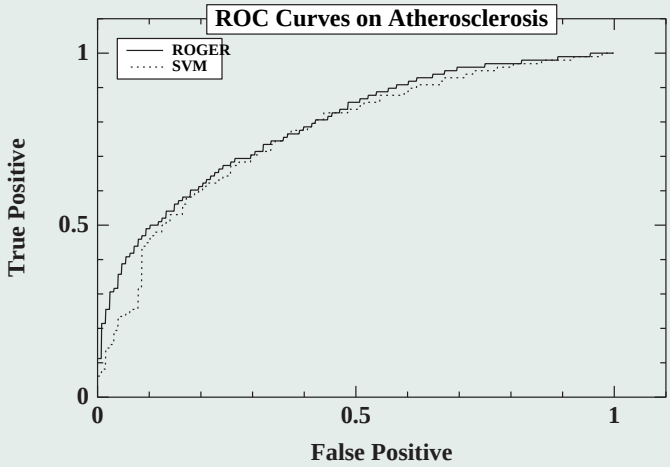
$$\text{Patient} \mapsto \mathbb{R}$$

Atherosclerosis

Experimental setting: 2/3 training, 1/3 test $\times$ 10

On each training set, 21 independent runs

Display the median ROC curve



Influence Analysis - The tobacco factor

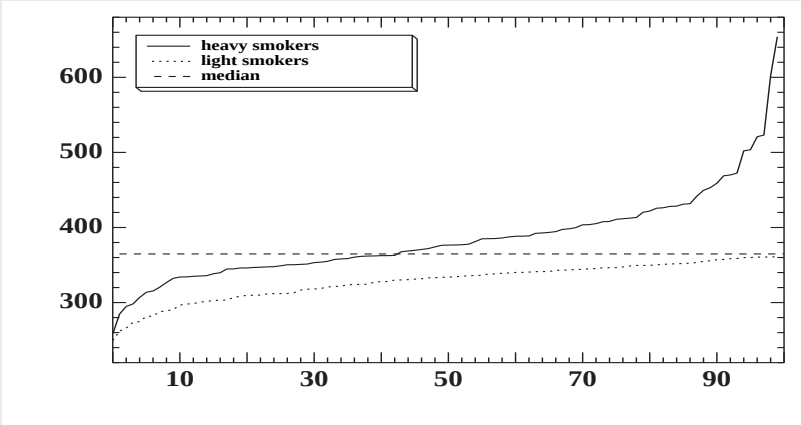
Procedure

A = { 100 non smoking individuals }

B = { 100 most smoking individuals }

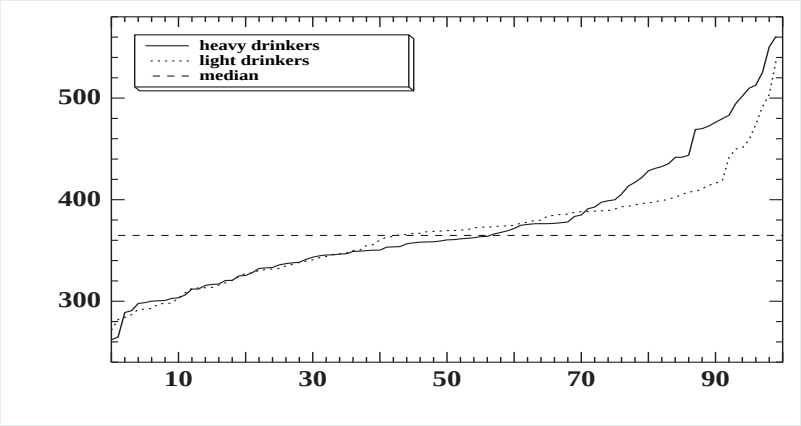
Sort A and B by increasing value of the risk

Plot (i, risk(i))

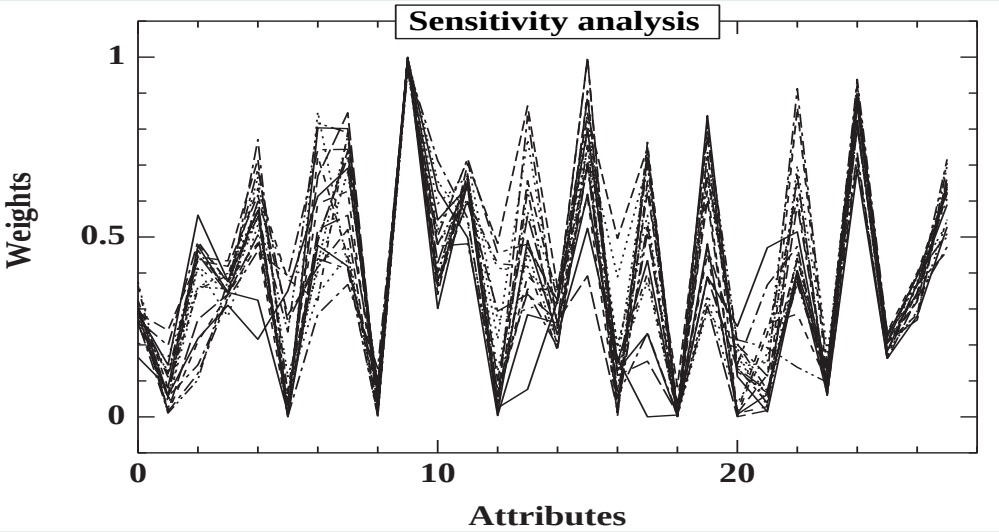


Influence Analysis - The alcool factor

A = { 100 light drinkers }
B = { 100 heavy drinkers }
Sort A and B by increasing value of the risk
Plot (i, risk(i))



Sensitivity Analysis - For free



21 runs, 21 solutions, 21 curves: $(i, weight(attribute_i))$

Conclusions - Perspectives

ICDM 2003, AE 2003

Present

- Good predictive performances
- Affordable complexity
- UNDERSTANDABLE RESULTS

Using Vision to Think, Card et al. 2001

Next

2004

- Extend to kernel spaces
- Use for constructive induction

An evolutionary approach: ROGER

Overview

- A combinatorial optimization formulation for ML
  tackled by Evolutionary Computation
- A real-world application: Risk for Cardio-Vascular Diseases
  and some remarks on the risk attached to tobacco and alcool...
- Defining more complex hypotheses
  i) non-linear; ii) allowing for scoring the features
- Exploit the variability of solutions provided by a stochastic optimization method (GA) $\Longrightarrow$ Ensemble learning.

A more complex hypothesis space

Linear space

$$h(x) = \sum_i w_i a_i(x) \qquad h \equiv w \in \mathbb{R}^d$$

Coarse non linear space

$$h(x) = \sum_i w_i |a_i(x) - c_i| \qquad h \equiv (w, c) \in \mathbb{R}^{2d}$$

Advantages

- non linear hypotheses
- linear search space $\mathbb{R}^{2d}$
- $score(atribut\ a_i) = w_i$.

An evolutionary approach: ROGER

Overview

- A combinatorial optimization formulation for ML
  tackled by Evolutionary Computation
- A real-world application: Risk for Cardio-Vascular Diseases
  and some remarks on the risk attached to tobacco and alcool...
- Defining more complex hypotheses
  i) non-linear; ii) allowing for scoring the features
- Exploit the variability of solutions provided by a stochastic optimization method (GA) $\Longrightarrow$ Ensemble learning.

Evolutionary computation and Ensemble Learning

Ensemble learning

$\mathcal{H}$: hypothesis space

- Error = bias + variance
- Bias : the best one can do on $\mathcal{H}$

$$Err(h^*) = Argmin\{Err(h),\ h \in \mathcal{H}\}$$

- Variance : we don't get h^*, alas.
  But instead $\hat{h}_n$, depending on the n training examples

Ensemble learning, 2

Principle : reducing the variance

- Learn $h_1, ..h_T$, decorrelated
- With a “reasonably low” error weak learning

$$Pr(h_i(x) = y) = \frac{1}{2} + \eta$$

- The vote (or linear aggregation of the h_i, improves on the best h_i

Sketch of proof : Hoeffding inequality

- Let V_i be random independent boolean variables, with probability p.
- Let Y_T be the sum of $V_1, ..V_T$

$$Pr(|Y_T - T \times p| > \epsilon) < exp^{-2\epsilon T^2}$$

Evolutionary computation and Ensemble Methods, 2

Stochastic Algorithm

- Each run $\rightarrow$ a hypothesis.
- Each hypothesis $\rightarrow$ an order on the features.

Weak order

- Let $\{a_1, ..a_N\}$ denote the target order
- h_t : induces an order relation $<_t$ on features
- Assume these are weak orders:

$$P(a_i <_t a_j | i < j) > \frac{1}{2} + \eta$$

Agregating weak orders

- Define $<_*$ as:

$$(a_i <_* a_j) \Longleftrightarrow |\{t / a_i <_t j\}| > \frac{T}{2}$$

Evolutionary computation and Ensemble Methods, 3

The agregated order is an order

$$Pr(i <_* k | i <_* j \text{ et } j <_* k) \to 1 \text{ as } T \to \infty$$

.. which goes toward the target order as T goes to ∞

- Let $O_*(i) = |\{j / i <_* j\}|$, then

$$Pr(|O_*(i) - i| > \tau) \to 0$$

Validation

Difficulty

- Validation of a feature subset ==
  generalization error of the best hypothesis based on these features

$\Rightarrow$ No way, for validating a feature selection/ranking method *per se*

Approache

- Artificial datasets
- Whose solution is known: can the method find the solution ?
- Enable “Lesions studies” :
  noise, scalability, wrt nb examples, features...

Artificial datasets

Order parameters

- Nb features $d = 100, 200, 500$
- Nb examples $n = d/2, d, 2d$
- Nb relevant features $r = d/20, d/10, d/5$
- Type of target concept : Linear / Non Linear
- Class noise $e = 0, 5, 10\%$
- Feature noise $\sigma = 0, 0.05, 0.1$

Construct an artificial dataset (d, n, r, l, e, σ)

Select the relevant features : $\{1, 2, ...r\}$ among $\{1, ..d\}$

For each example x_j

- For $i = 1..d$, draw $a_i(x_j)$ uniformly in $[0, 1]$

Construct y_j

- Linear target concept:

$$y_j = (\sum_{i=1}^{r} a_i(x_j) > \frac{r}{2})$$

- Non-linear target concept:

$$y_j = (\sum_{i=1}^{r} |a_i(x_j) - .5| < \frac{r}{12})$$

Construct an artificial dataset $(d, n, r, l, e, \sigma), 2$

Perturbations

- $y_j = -y_j$ with probability e
- $a_i(x_j) \mathrel{+}= \mathcal{N}(0, \sigma)$

Experimental setting

For each tuple (d, n, r, l, e, σ), construct 20 datasets

Foreach dataset, learn 20 hypotheses — 20 runs

Agregate the orders based on the 20 hypotheses

Compare the agregate order with the target order

Average the ranking error over the 20 datasets.

Baseline Algorithm

Stoppiglia et al., JMLR 2003

Score of a feature

- Cosine : score(a) = $\sum_i a(x_i).y_i$

Gauss-Schmidt iterative projection

- Find the best feature a
- Project the dataset and the concepts on the subspace orthogonal to a.

Performance measure

For iterative selection

p_b probability for the top-ranked feature to be relevant

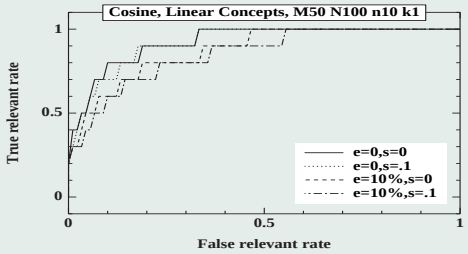
For iterative elimination

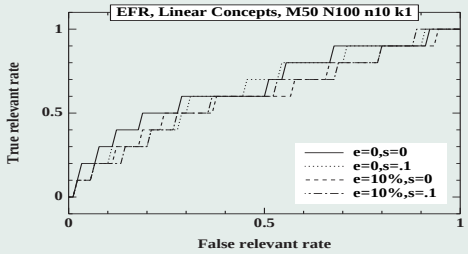
p_w rank of the last relevant feature

Tradeoff

True relevant rate *vs* False relevant rate (ROC)

Comparison on linear concepts



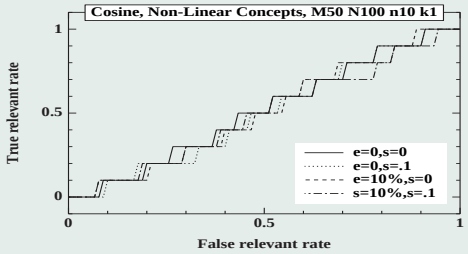


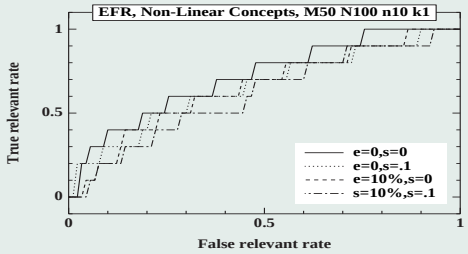
Stoppiglia ROGER

Stoppiglia >> ROGER >> Random

Here : d = 100, $n = d/2$, $r = d/10$

Comparison on non linear concepts



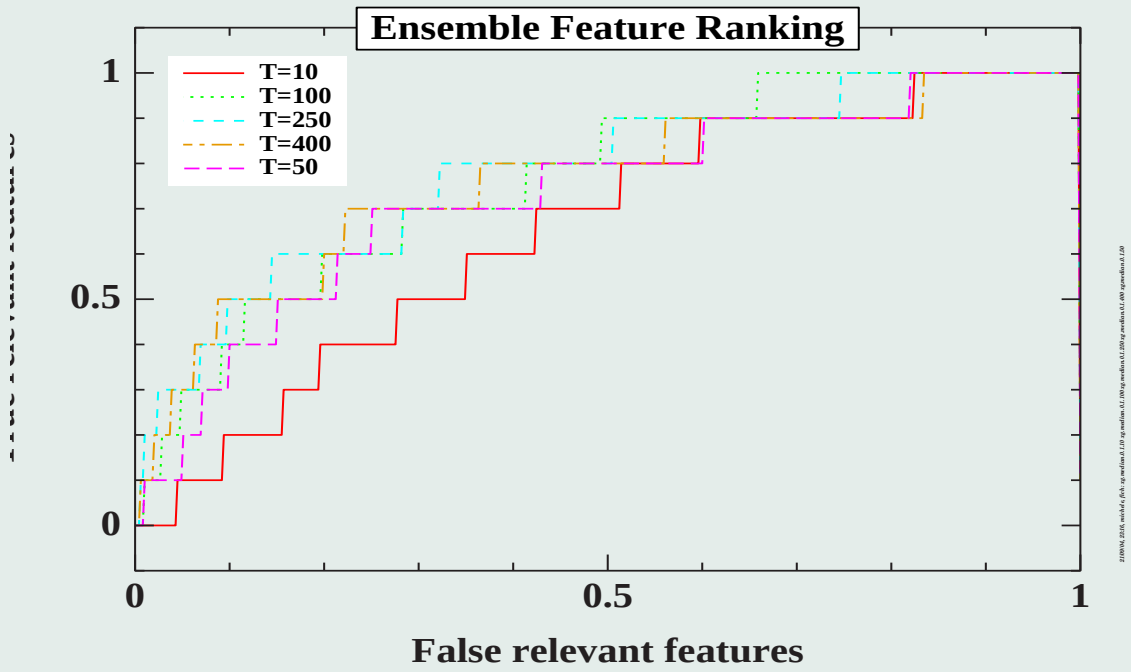


Stoppiglia ROGER

ROGER $>>$ Stoppiglia = Random

Here : d = 100, $n = d/2$, $r = d/10$

Ensemble Feature Ranking



When T increases from 10 to 400.

Conclusion

Contributions

- weak ranking $\Rightarrow$ strong ranking
- EC enables ensemble methods “for free”
- a principled framework for evaluating feature ranking/selection

Limits

- Only conjunctive concepts.

Next

- Multi-modal evolution / several hypotheses in a population.