Identification des facteurs de risque

Michèle Sebag

TAO – CNRS – INRIA – LRI

Michele.Sebag@lri.fr

Digiteo, 16 novembre 2005

Cadre

Apprentissage supervisé

Prédiction de durée de vie.

Risques

Descripteurs — Feature selection

Exemples — Outliers

Travail joint :

Jérome Azé, Mathieu Roche, Jérémie Mary, Antoine Cornuéjols LRI

Elena Marchiori, Kees Jong Vrije Universiteit Amsterdam

Plan

- Etat de l’art
- Receiver Operating Characteristics (ROC) Analysis
- Optimiser l’aire sous la courbe ROC
- Applications
  - Visualisation du risque
  - Feature Selection
  - Chimiometrie

Motivations

Avant l'apprentissage : décrire les données...

- Une description trop pauvre — on ne peut rien faire
- Une description trop riche — on doit filter les descripteurs

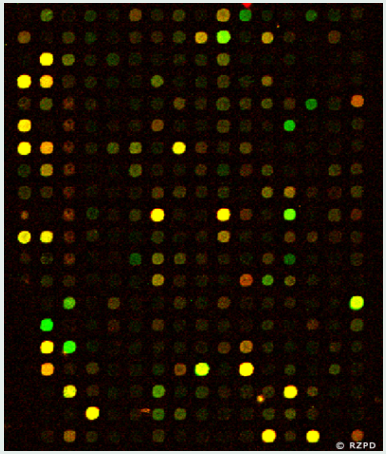
Pourquoi ?

- L'apprentissage n'est pas un problème bien posé
- $\Longrightarrow$ Rajouter de l'information inutile peut dégrader les hypothèses obtenues.

Le but caché : sélectionner ou construire des descripteurs ?

- Feature Construction : construire les bons descripteurs
- A partir desquels il sera facile d'apprendre
- Les meilleurs descripteurs = les bonnes hypothèses...

Quand l’apprentissage c’est la sélection d’attribut



Bio-informatique

- 30 000 gènes
- peu d’exemples (chers)
- but : trouver les gènes pertinents

Position du problème

Buts

- Sélection : trouver un sous-ensemble d'attributs
- Ordre/Ranking : ordonner les attributs

Formulation

Soient les attributs $\mathcal{A} = \{a_1, ..a_d\}$. Soit la fonction :

$$\begin{aligned} \mathcal{F} : \mathcal{P}(\mathcal{A}) &\mapsto \mathbb{R} \\ A \subset \mathcal{A} &\mapsto Err(A) = \text{ erreur min. des hypothèses fondées sur } A \end{aligned}$$

Trouver $Argmin(\mathcal{F})$

Difficultés

- Un problème d'optimisation combinatoire (2^d)
- D'une fonction $\mathcal{F}$ inconnue...

Approches

Filter — méthode univariée

Définir $score(a_i)$; ajouter itérativement les attributs maximisant $score$
ou retirer itérativement les attributs minimisant $score$

$+$ simple - pas cher

$-$ optima très locaux

Rq : on peut bactracker : meilleurs optima, mais plus cher

Wrapping — méthode multivariée

Mesurer la qualité d'attributs en rapport avec d'autres attributs :

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

$-$ cher : une estimation = un pb d'apprentissage.

$+$ optima meilleurs

Approches filtre

Notations

Base d'apprentissage : $\mathcal{E} = \{(x_i, y_i), i = 1..n, y_i \in \{-1, 1\}\}$

$a(x_i)$ = valeur attribut a pour exemple (x_i)

Gain d'information

arbres de décision

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

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

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

Corrélation

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

Approches wrapper

Principe générer/tester

Etant donné une liste de candidats $\mathcal{L} = \{A_1, .., A_p\}$

- Générer un candidat A
- Calculer $\mathcal{F}(A)$
  - apprendre h_A à partir de $\mathcal{E}_{|A}$
  - tester h_A sur un ensemble de test $= \hat{\mathcal{F}}(A)$
- Mettre à jour $\mathcal{L}$.

Algorithmes

- hill-climbing / multiple restart
- algorithmes génétiques Vafaie-DeJong, IJCAI 95
- (*) programmation génétique & feature construction.

Krawiec, GPEH 01

Approches a posteriori

Principe

- Construire des hypothèses
- En déduire les attributs importants
- Eliminer les autres
- Recommencer

Algorithme : SVM Recursive Feature Elimination

Guyon et al. 03

- SVM linéaire $\rightarrow h(x) = sign(\sum w_i.a_i(x) + b)$
- Si $|w_i|$ est petit, a_i n'est pas important
- Eliminer les k attributs ayant un poids min.
- Recommencer.

Limites

Hypothèses linéaires

- Un poids par attribut.

Quantité des exemples

- Les poids des attributs sont liés.
- La dimension du système est liée au nombre d’exemples.

Or le pb de FS se pose souvent quand il n’y a pas assez d’exemples

Plan

- Etat de l’art
- Receiver Operating Characteristics (ROC) Analysis
- Optimiser l’aire sous la courbe ROC
- Applications
  - Visualisation du risque
  - Feature Selection
  - Chimiometrie

Receiver Operating Characteristics

Principe — traitement du signal, médecine

Soit $h(x)$ mesurant le risque du patient x. $h : X \mapsto \mathbb{R}$

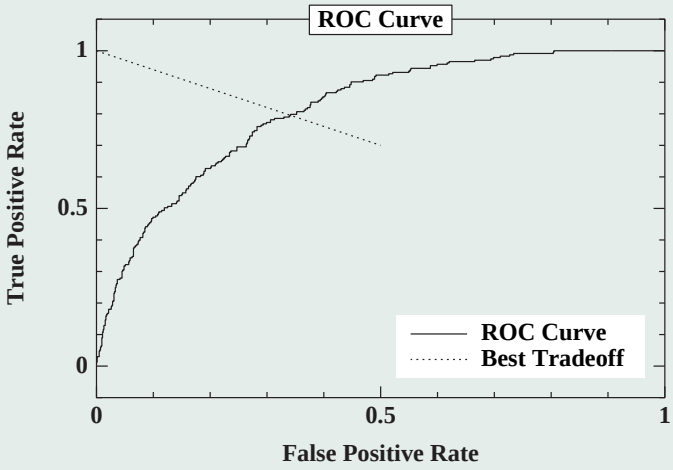
$$t \in \mathbb{R} \mapsto h_t(x) = \begin{cases} malade & si\ h(x) > t \\ OK & sinon \end{cases}$$

Pour h_t, définir:

- TP(t) : true positive rate, $Pr(h_t(x) = malade | x\ malade)$
- FP(t) : false positive rate, $Pr(h_t(x) = malade | x\ pas\ malade)$.

Tracer la courbe $(TP(t), FP(t), t \in \mathbb{R})$.

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 minimized

Plan

- Etat de l'art
- Receiver Operating Characteristics (ROC) Analysis
- Optimiser l'aire sous la courbe ROC (AUC)
- Applications
  - Visualisation du risque
  - Feature Selection
  - Chimiometrie

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

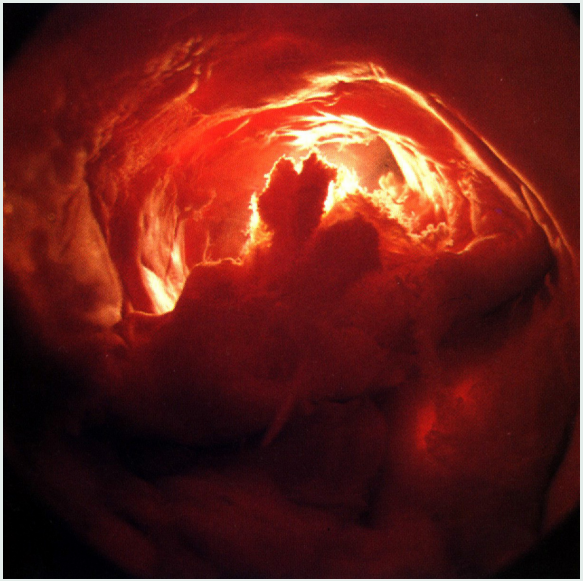
For each split, 1 SVMTorch run, 21 ROGER runs (take median)

EA 03, IEEE ICDM 03

Plan

- Etat de l’art
- Receiver Operating Characteristics (ROC) Analysis
- Optimiser l’aire sous la courbe ROC (AUC)
- Applications
  - Visualisation du risque
  - Feature Selection
  - Chimiometrie

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

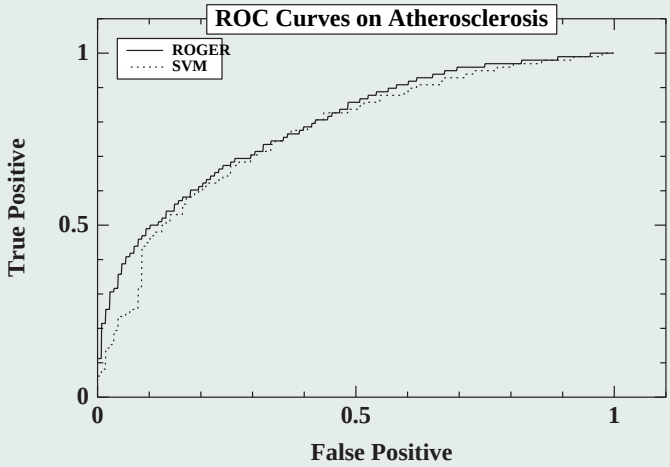
$$\text{Patient} \mapsto \mathbb{R}$$

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

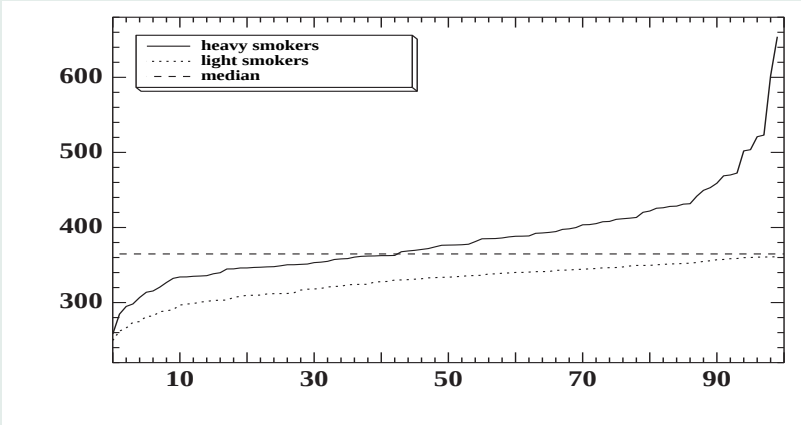
Procedure

A = { 100 non smoking individuals }

B = { 100 most smoking individuals }

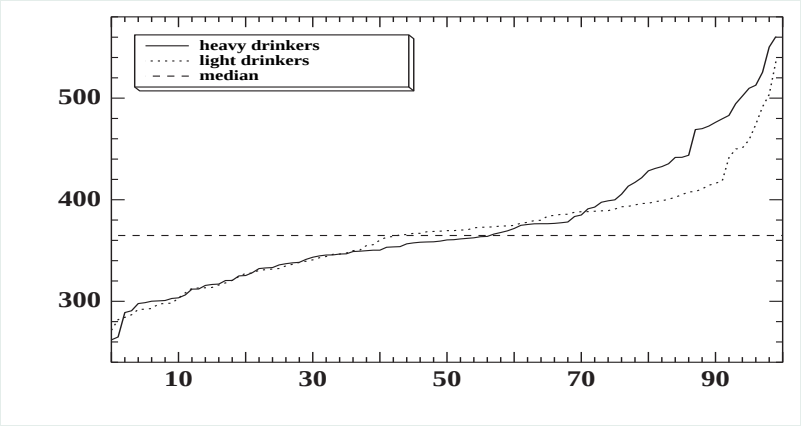
Sort A and B by increasing value of the risk

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))



Visualizing the risk

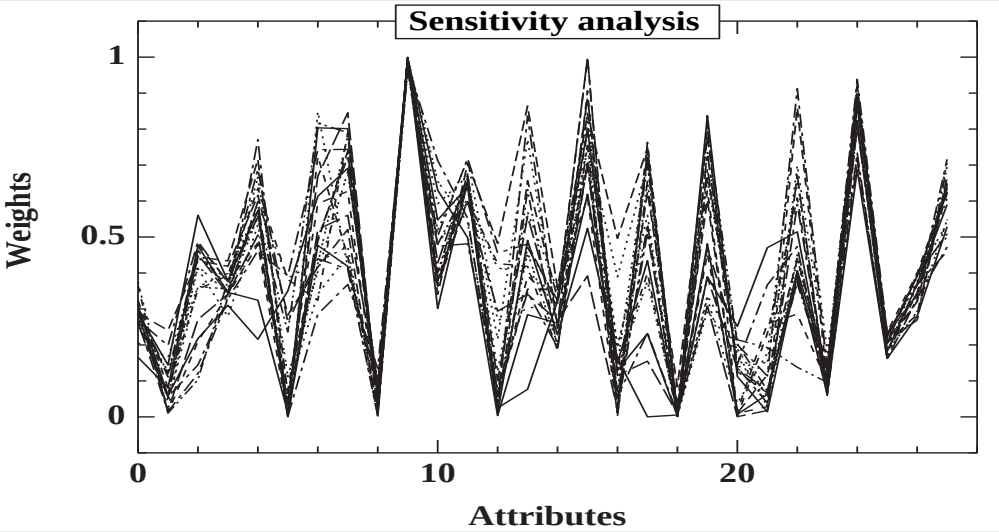
- Good predictive performances
- Affordable complexity
- UNDERSTANDABLE RESULTS

Using Vision to Think, Card et al. 2001

Plan

- Etat de l’art
- Receiver Operating Characteristics (ROC) Analysis
- Optimiser l’aire sous la courbe ROC (AUC)
- Applications
  - Visualisation du risque
  - Feature Selection
  - Chimiometrie

Sensitivity Analysis - For free



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

Un espace d'hypothèses plus intéressant

Espace linéaire

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

Espace non linéaire pauvre

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

Intérêt

hypothèses non linéaires
espace de recherche linéaire $\mathbb{R}^{2d}$
$score(atribut\ a_i) = w_i$.

Roger - Feature Selection

Principle

T runs $\rightarrow$ T hypotheses

Ensemble learning for free

Quality of a feature

proportional to its |weight|

averaged over all hypotheses

Algorithme de référence

Stoppiglia et al., JMLR 2003

Score d'un attribut

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

Projection itérative de Gauss

- Trouver le meilleur attribut a
- Projeter les données et le concept sur l'espace ortogonal à a

Mesure de performance

Qualité pour une sélection itérative

p_b probabilité du top d'être pertinent

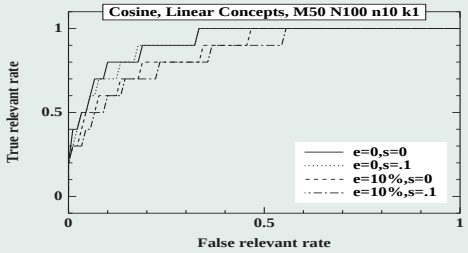
Qualité pour une élimination itérative

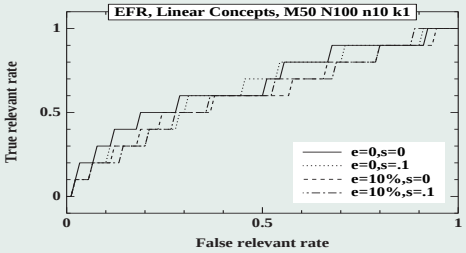
p_w pire rang d'un attribut pertinent

Compromis

Taux de vrais pertinents *vs* taux de faux pertinents : AUC.

Comparaison sur des concepts linéaires



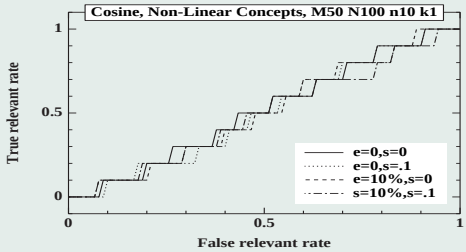


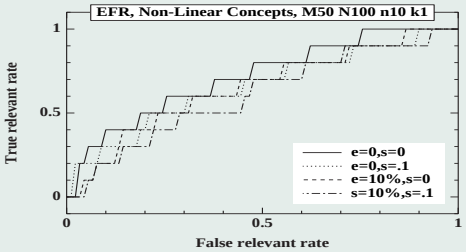
Stoppiglia ROGER

Stoppiglia >> ROGER >> Random

nb descripteurs 100; nb exemples 50; nb desc. pertinents: 10;

Comparaison sur des concepts non linéaires



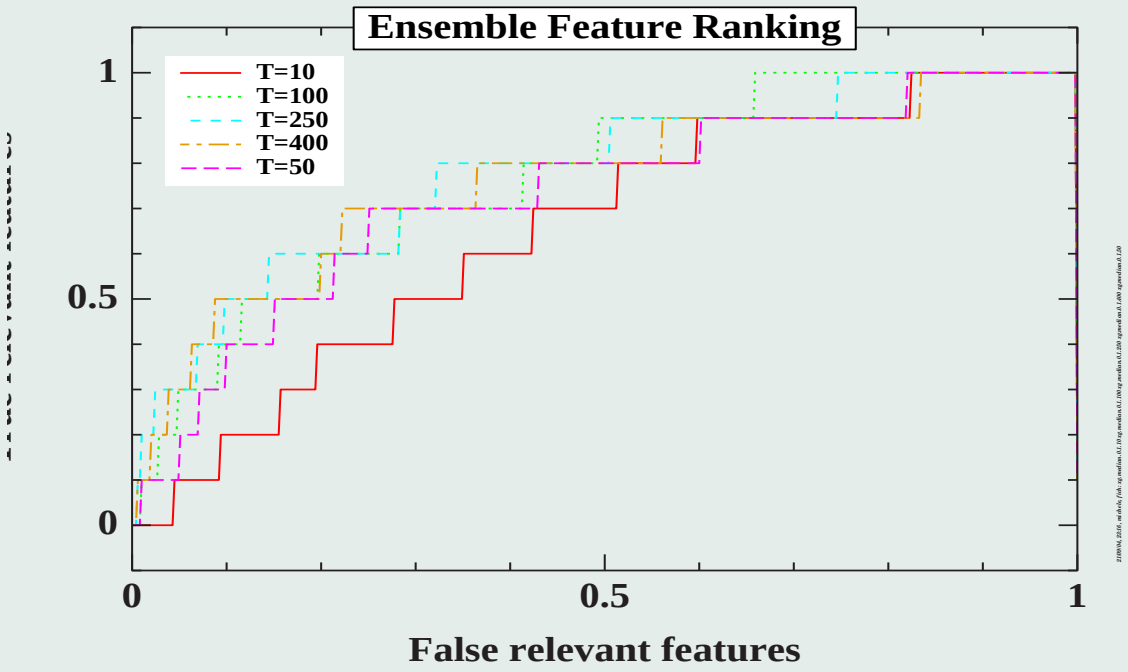


Stoppiglia ROGER

ROGER >> Stoppiglia = Random

nb descripteurs 100; nb exemples 50; nb desc. pertinents: 10;

Ensemble Feature Ranking



Quand on augmente T : de 10 à 400.

Plan

- Etat de l’art
- Receiver Operating Characteristics (ROC) Analysis
- Optimiser l’aire sous la courbe ROC
- Applications
  - Visualisation du risque
  - Feature Selection
  - Chimiometrie

Chimiométrie 2005

Challenge GFC 05

215 exemples, 2901 descripteurs réels
4 classes
43 exemples témoins.

Roger, induction constructive

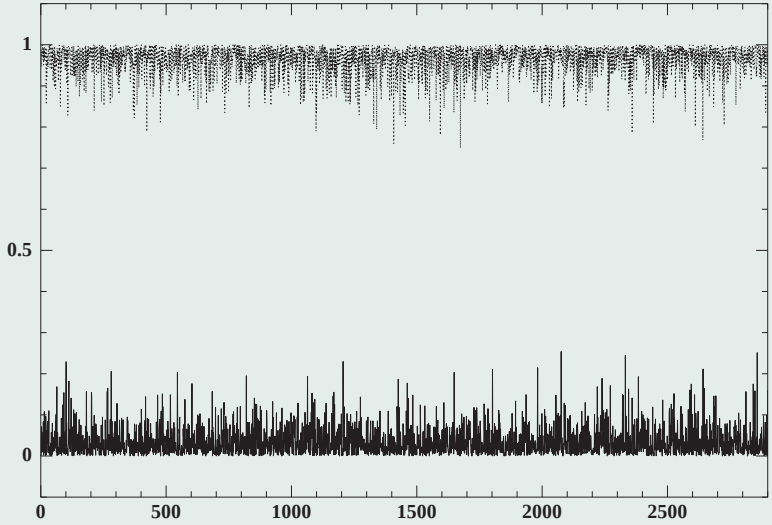
Une classe contre les trois autres $\rightarrow$ une hypothèse
n runs
$\rightarrow 4n$ nouveaux descripteurs.

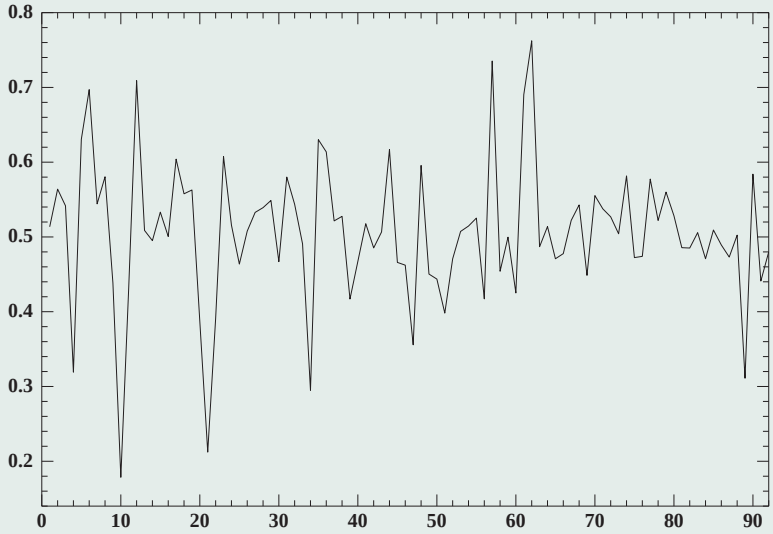
Usage

plus-proche voisins.

Références

Methods	chemo	chemopre
AdaBoost boosting=100 weak_learner=lgg	13.25	12.91
AdaBoost boosting=1000 weak_learner=lgg	-	12.45
AdaBoost boosting=1000 weak_learner=qlgg	12.29	-
AdaBoost boosting=2000 weak_learner=qlgg	7	11.56
ADTree mode=boostexter boosting=100	12.94	9.13
C4.5	19.75	16.88
C5 boosting=100	11.91	8.81
DLG weight uniform	23.65	-
GloBo mode=fast weight uniform	20.32	-
GloBoost boosting=100 weight frequency	16.11	12.14
GloBoost boosting=1000 weight frequency	-	12.32
GloBoost boosting=1000 mode=fast weight frequency	13.52	-
GloBoost boosting=2000 mode=fast weight frequency	13.73	-
RISE	-	10.28
Average	15.37	11.86


1
0.5
0
0
500
1000
1500
2000
2500


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
10
20
30
40
50
60
70
80
90

Conclusion

Contributions

- EC enables ensemble methods “for free”
  for feature selection
  for feature construction

Limites

- Concepts conjonctifs seulement.

Next

- Multi-modal evolution / several hypotheses in a population.
- Developper la visualisation

Evolution artificielle et méthodes d'ensemble, 2

Algorithme stochastique

- Chaque run $\rightarrow$ une hypothèse indépendante.
- Chaque hypothèse $\rightarrow$ un ordre sur les attributs.

Ordre faible

- Soit $\{a_1, ..a_N\}$ l'ordre parfait
- h_t : induit un ordre $<_t$ sur les attributs
- Supposons un ordre faible :

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

Agrégation

- On définit $<_*$ comme :

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

Evolution artificielle et méthodes d'ensemble, 3

L'ordre agrégé est bien un ordre

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

.. et tend vers l'ordre parfait

- Soit $O_*(i) = |\{j / i <_* j\}|$ alors

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

Validation

Difficulté

- Validation d'un ens d'attributs == qualité de la meilleure hypothèse fondée sur ces attributs

$\Rightarrow$ Pas moyen de tester une méthode de sélection en soi.

Approche

- Pbs artificiels
- On connait la solution ; est-ce qu'on la retrouve ?
- Permet étude de "Lésions" : bruit, passage à l'échelle % nb exemples, nb attributs...

Problèmes artificiels

Paramètres d'ordre

- Nb attributs $d = 100, 200, 500$
- Nombre d'exemples $n = d/2, d, 2d$
- Nombre d'attributs pertinents $r = d/20, d/10, d/5$
- Type de concept à apprendre : Linéaire ou Non.
- Bruit de classe $e = 0, 5, 10\%$
- Bruit d'attribut $\sigma = 0, 0.05, 0.1$

Construire un pb artificiel (d, n, r, l, e, σ)

Se donner les attributs pertinents : $\{1, 2, ...r\}$ parmi $\{1, ..d\}$

Pour chaque exemple x_j

- Pour $i = 1..d$, tirer $a_i(x_j)$ uniformément ds $[0, 1]$

Construction de y_j

- Cas linéaire :

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

- Cas non-linéaire :

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

Pbs artificiels, suite

Perturbation

- $y_j = -y_j$ avec probabilité e
- $a_i(x_j) += \mathcal{N}(0, \sigma)$

Méthodologie expérimentale

Pour chaque (d, n, r, l, e, σ), construire 20 problèmes

  Pour chaque problème, apprendre 20 hypothèses — 20 runs

  Agréger les poids des 20 hypothèses

  Comparer l'ordre obtenu à l'ordre désiré

Moyenner l'erreur sur les 20 problèmes