

Extraction de la terminologie

Seconde partie, une approche supervisée

Mathieu Roche

Cours Fouille de Données

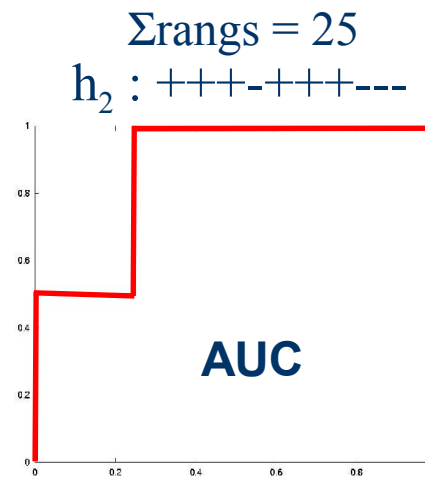
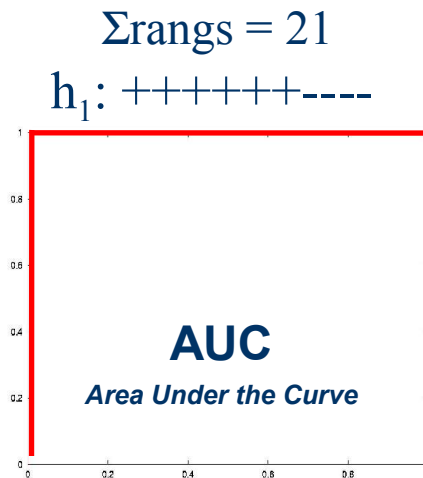
février 2007

Approche supervisée pour apprendre une mesure [Roche *et al.*, ROCAI'04 ; Azé *et al.*, ICCI'04]

- **Entrée** : quelques collocations étiquetées (positives ou négatives).
- **Sortie** : fonction de rang [Cohen *et al.* 1999]
- **Évaluation d'une fonction de rang** : somme des rangs des exemples positifs.

Approche supervisée pour apprendre une mesure [Roche *et al.*, ROCAI'04 ; Azé *et al.*, ICCI'04]

- Minimiser la somme des rangs des exemples positifs \Leftrightarrow maximiser l'aire sous la courbe ROC



- **Avantage : pas de sensibilité** dans le cas d'un **déséquilibre** entre les classes.

Protocole expérimental (1/2)

- Données utilisées

	# collocations	% collocations pertinentes	% collocations non pertinentes
CV, fréquents	376	85.7	14.3
CV, rares	2822	56.6	43.4
Biologie	1028	90.9	9.1

Protocole expérimental (2/2)

Critères statistiques	AUC <i>collocations fréquentes</i> corpus de CVs	AUC <i>collocation fréquentes</i> corpus de Biologie
OCC_{RV} - Occurrence + RV [Roche <i>et al.</i> 2004]	0.58	0.57
RV - Rapport de Vraisemblance [Dunning 1993]	0.43	0.42
I³ - Information Mutuelle au cube [Daille <i>et al.</i> 1998]	0.40	0.35
Dice - Coefficient de Dice [Smadja <i>et al.</i> 1996]	0.39	0.31
I - Information Mutuelle [Church and Hanks 1990]	0.31	0.30

- **Combinaison de mesures**

Algorithme ROGER (ROC based GENetic learner) (1/3)

Approche linéaire

$$h(\text{Coll}) = \sum w_i x \text{ mes}_i (\text{Coll}) \text{ avec } (\text{Coll}, +/-)$$

Approche non linéaire

$$h(\text{Coll}) = \sum w_i x | \text{mes}_i (\text{Coll}) - c_i | \text{ avec } (\text{Coll}, +/-)$$

Hypothèses : Aire sous la courbe ROC

$h \rightarrow (\text{rang}(\text{Coll}), \text{Etiqu}(\text{Coll}))$

classer les exemples par rangs croissants

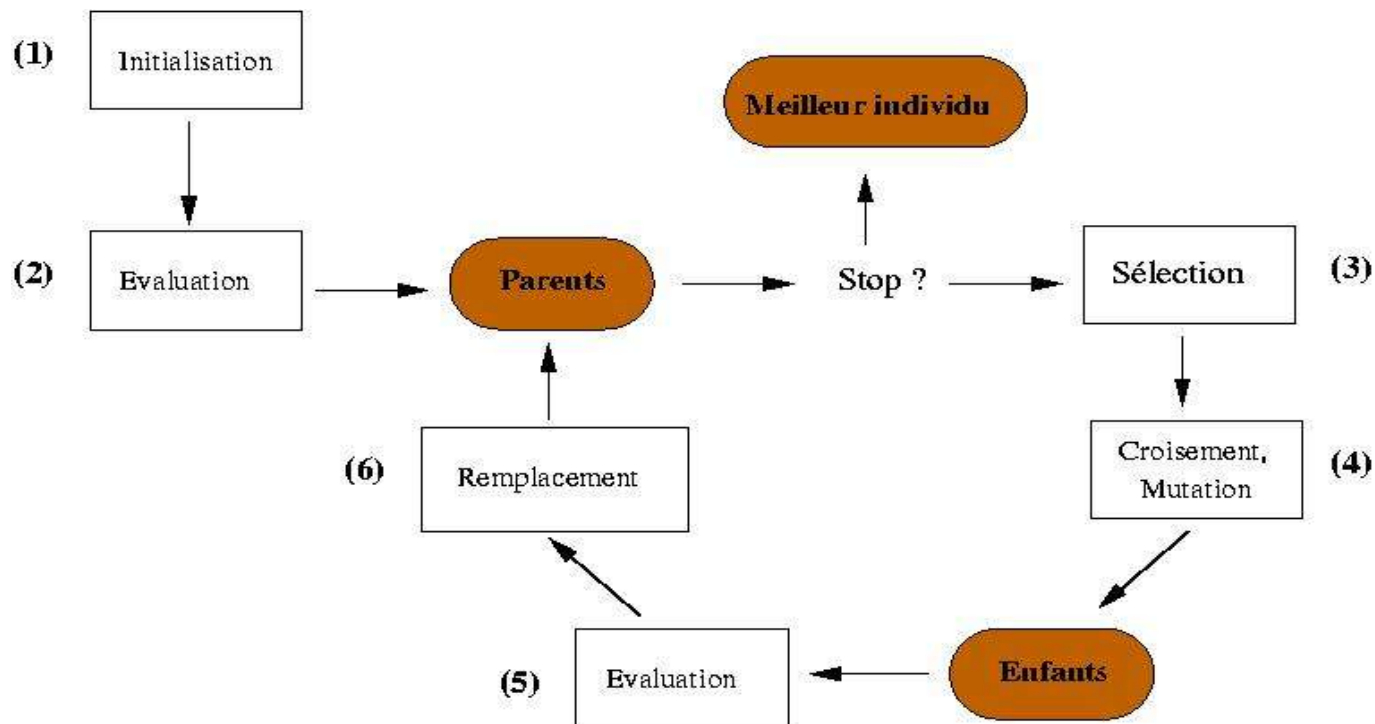


+ : collocation pertinente

- : collocation non pertinente

Algorithme ROGER (2/3)

- Utilisation des algorithmes génétiques



Algorithme ROGER (3/3)

- **Protocole expérimental**

- 90% Apprentissage, 10% Test, 10 validations croisées
- 21 exécutions indépendantes
- Soit h_1, \dots, h_T les meilleurs hypothèses retenues à partir de T ($T=21$) exécutions indépendantes de ROGER.

$$Bh(x) = \text{Médiane}(\{h_t(x), t=1 \dots T\})$$

Algorithme ROGER (2/2)

- Validation expérimentale sur les ensembles tests

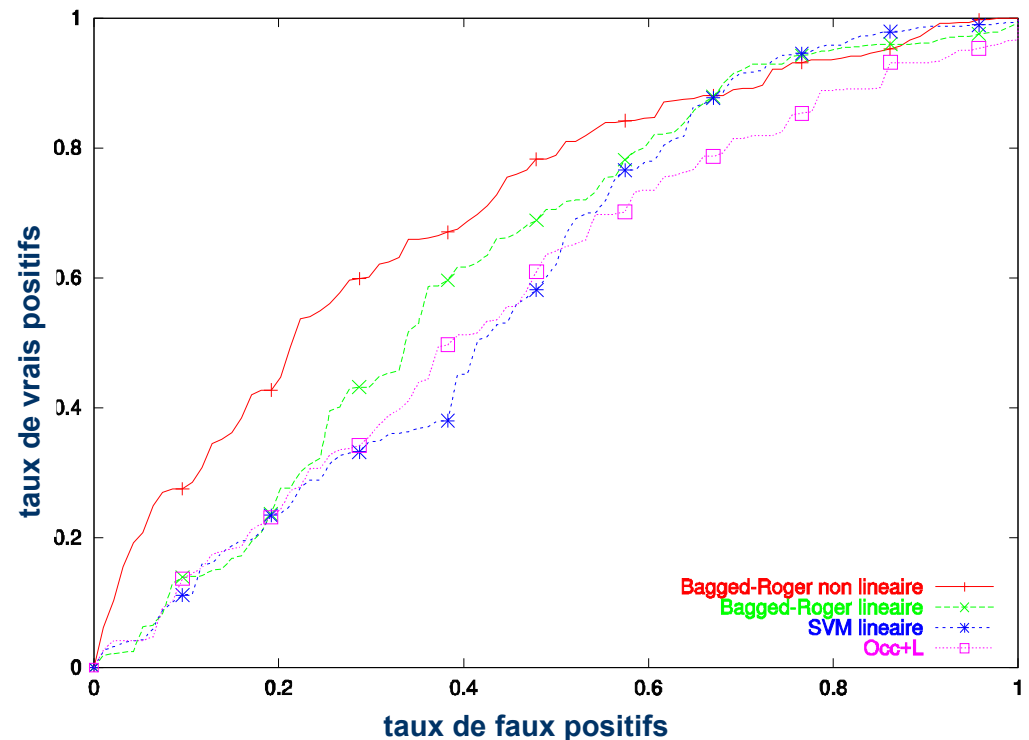
	OCC_{RV}	Bagged-Roger		Bagged-SVM		
		<i>Linéaire</i>	<i>Non linéaire</i>	<i>Linéaire</i>	<i>Gaussien</i>	<i>Quadratique</i>
Biologie, fréquents	0.57	0.61 ± 0.04	0.67 ± 0.05	0.51 ± 0.13	0.54 ± 0.12	0.32 ± 0.07
CV, fréquents	0.58	0.59 ± 0.10	0.61 ± 0.11	0.46 ± 0.13	0.42 ± 0.14	0.52 ± 0.07

- Etude de généralité
 - différents domaines
 - différentes langues
 - différentes fréquences des collocations

Étude de généralité (1) : apprentissage CVs / application Biologie (fréquents)

	AUC Collocation fréquentes Corpus de Biologie
Occ_{RV}	0.57
RV	0.42
β^3	0.35
$Dice$	0.31
I	0.30

SVM	Bagged-ROGER	
Linéaire	Linéaire	Non Linéaire
0.59	0.63	0.71

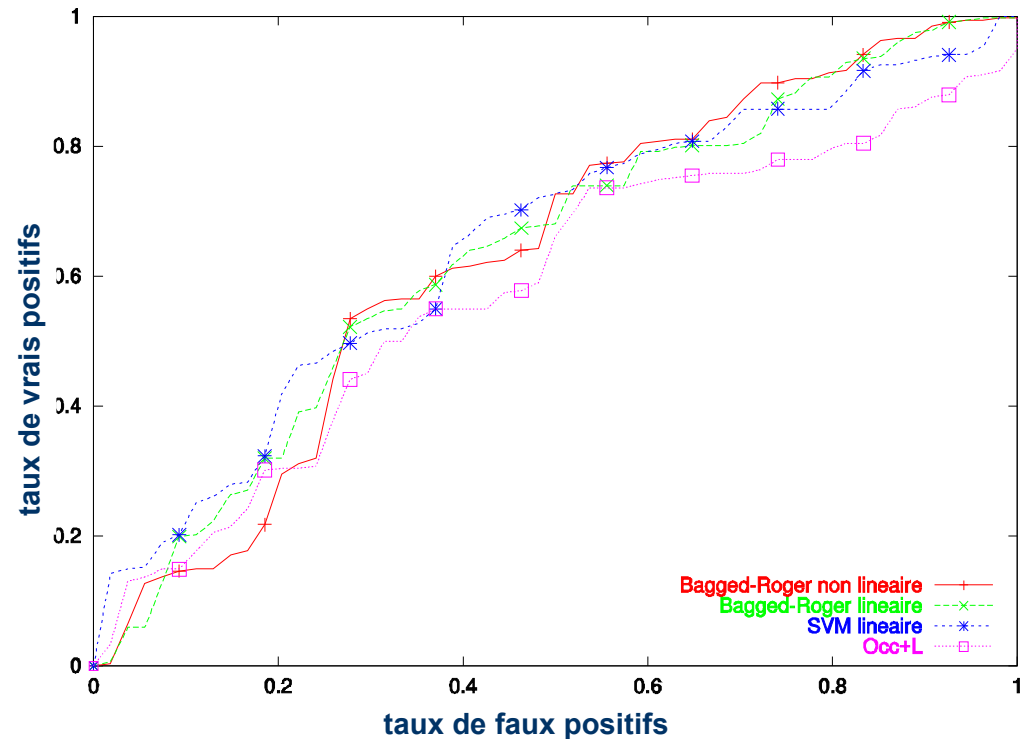


Autres noyaux donnent des résultats plus faibles

Étude de généralité (2) : apprentissage Biologie / validation CVs (fréquents)

	AUC Collocations fréquentes Corpus de CVs
Occ_{RV}	0.58
RV	0.43
β^3	0.40
$Dice$	0.39
I	0.31

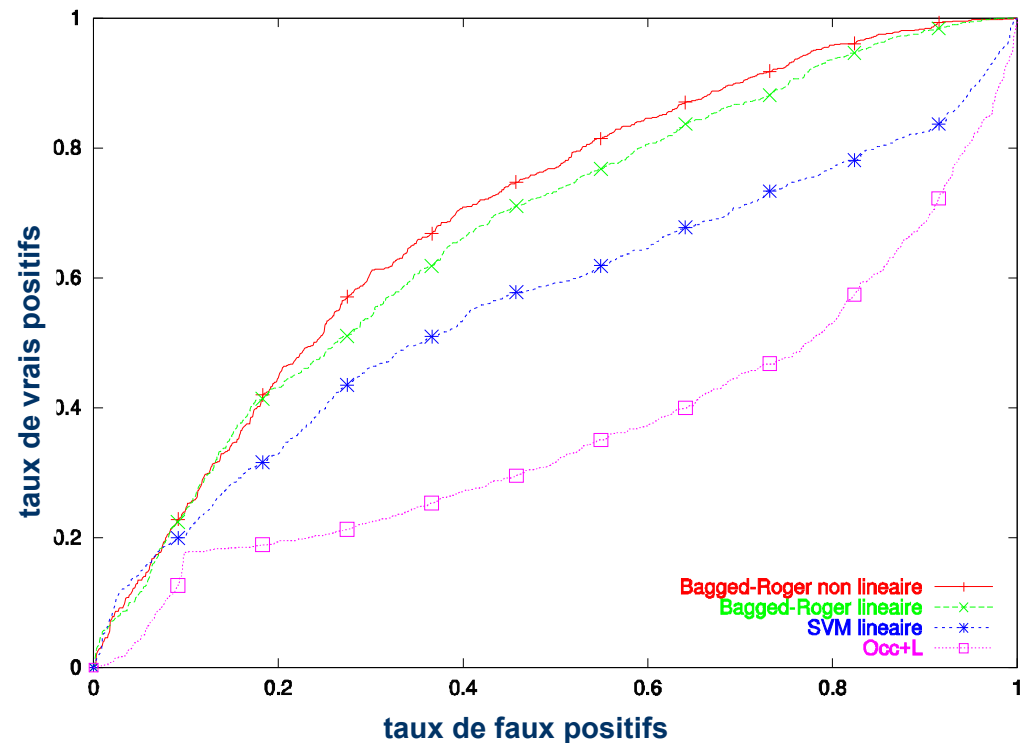
SVM	Bagged-ROGER	
<i>Linéaire</i>	<i>Linéaire</i>	<i>Non Linéaire</i>
0.65	0.64	0.63



Étude de généralité (3) : apprentissage coll. fréquentes / application coll. rares (CVs)

	AUC Collocations rares Corpus de CVs
Occ_{RV}	0.37
Dice	0.32
RV	0.30
I^{β}	0.30
I	0.29

SVM	Bagged-ROGER	
Linéaire	Linéaire	Non Linéaire
0.56	0.67	0.70



Perspectives

- **Apprentissage actif** : demander à l'expert de valider un nombre restreint de collocations à chaque exécution de ROGER.