Développement d'une base documentaire sur le VIH : l'intelligence artificielle au service de la recherche biomédicale.

Marie-Jean Meurs[1,2]

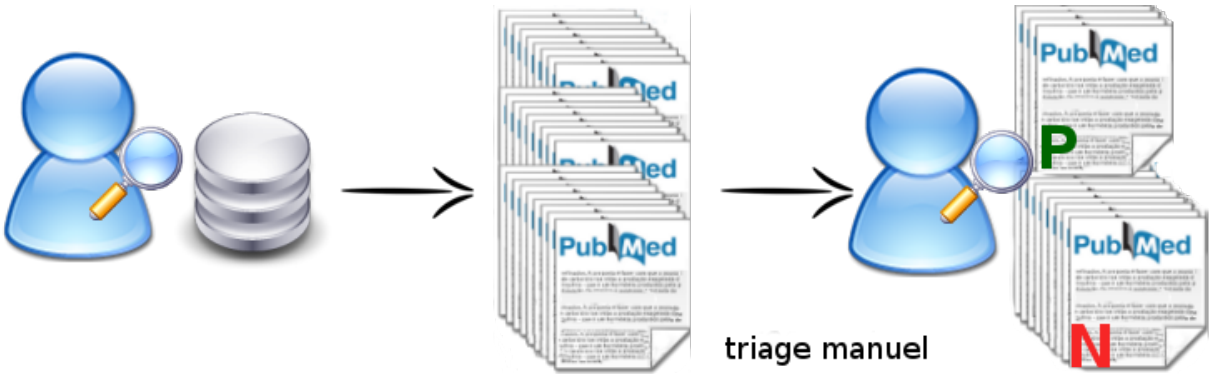
[1] Université du Québec à Montréal

Collaboration: Hayda Almeida[2], Leila Kosseim[3], Adrian Tsang[2]
[2] Centre for Structural and Functional Genomics
[3] Dept. of Computer Science and Software Engineering
Concordia University

Montréal, QC, Canada

84è Congrès de l'ACFAS, Jeudi 12 mai 2016

Curation de la littérature biomédicale



- Beaucoup de documents – peu sont sélectionnés
- Triage : difficile, long, sujet à erreurs
- Aucune garantie d'exhaustivité
- Important goulot d'étranglement pour la curation manuelle

Synthesized HIV Research Evidence (SHARE)

- Base de données accessible en ligne
- Synthèses des connaissances sur le VIH : études systématiques
  - Découvertes de recherche, guides de traitement, protocoles
- Ressource pour les réseaux de traitement du VIH
  - Soignants, décideurs gouvernementaux, chercheurs
- Études systématiques dans SHARE : mises à jour periodiques

résultats: *"HIV"* $\rightarrow 290{,}000^{+}$ *"AIDS"* $\rightarrow 230{,}000^{+}$

https://aidsevidence.ca

Approche par apprentissage machine

Classification automatique de textes

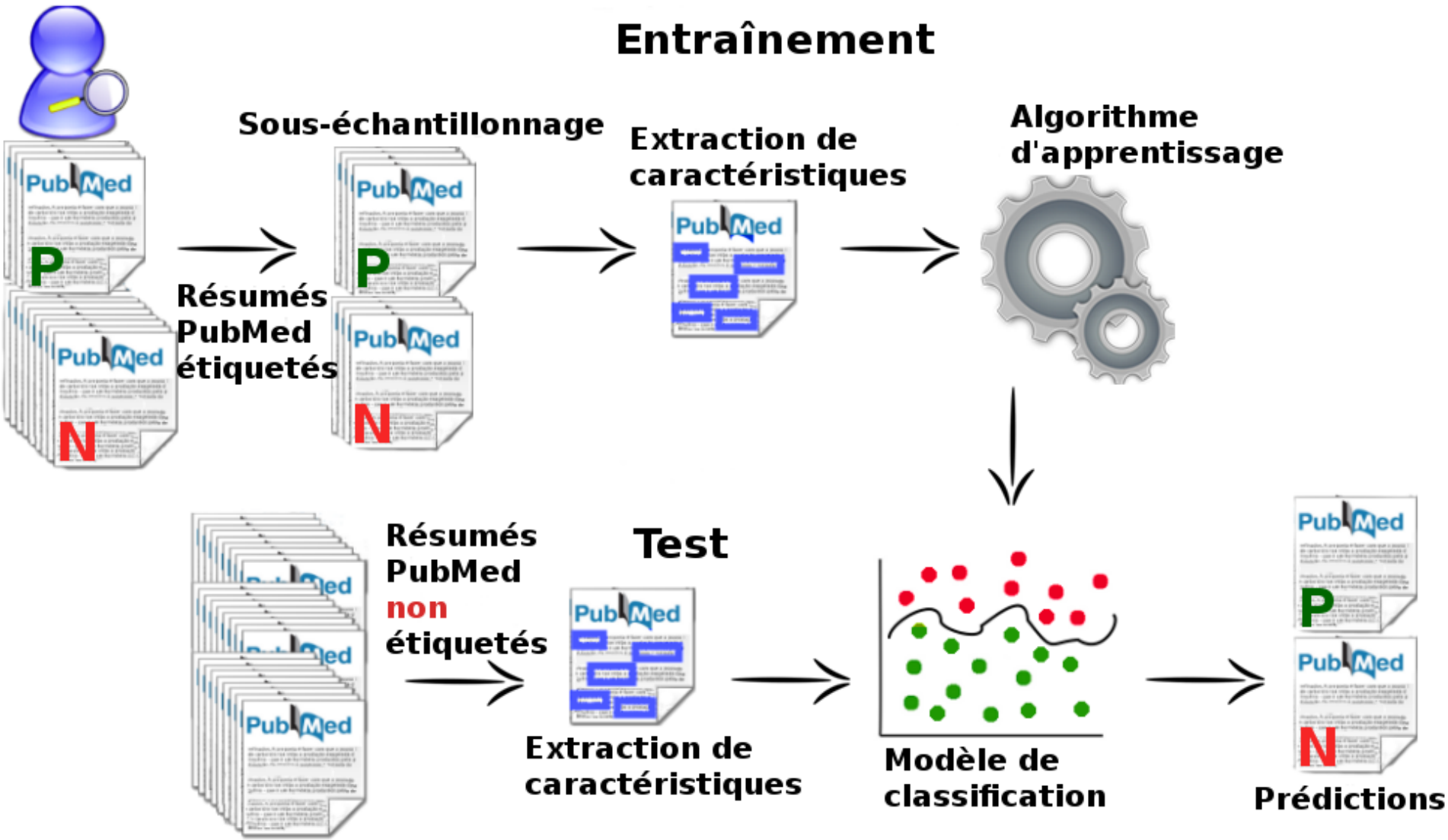
- Faciliter la tâche de curation/triage aux scientifiques
- Évaluer plus de documents en moins de temps
- Moins de chance de manquer des documents pertinents

But : réduire l'effort en signalant les documents candidats

Triage par apprentissage supervisé

- Corpus de documents correctement étiquetés
- Exemples d'entraînement donnés au classifieur
- Modèle appris → étiquetage des nouveaux documents
- Exemples de test → évaluation du modèle

Triage par apprentissage supervisé



Classification de texte pour la biocuration : problèmes spécifiques

1. Distribution déséquilibrée des classes
   - Grand corpus mais peu de documents pertinents
   - Majorité non pertinente = bruit
   - Distribution déséquilibrée des classes ⇒ performances ↘

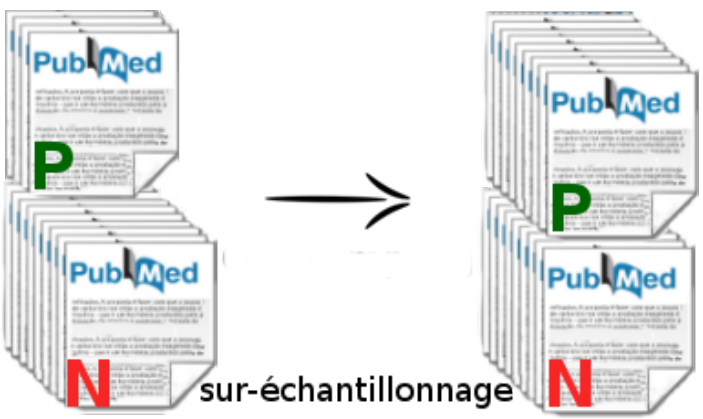
   ⇒ Réduction des biais indispensable

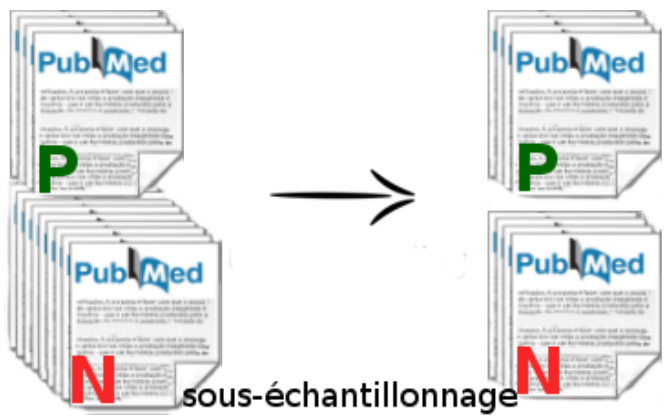
2. Ensemble de caractéristiques très vaste
   - Trop de caractéristiques ⇒ sur-apprentissage
   - Faible pouvoir discriminant ⇒ contribution médiocre
   - Plus de caractéristiques ⇒ plus de resources de calcul

   ⇒ Identification des meilleures caractéristiques pour la tâche

Échantillonnage des données

- Sélection d'un sous-ensemble spécifique du corpus
- Étape préliminaire
- Sur-échantillonnage : ajouter des instances synthétiques pour équilibrer la distribution
- Sous-échantillonnage : retirer des instances de la classe majoritaire pour équilibrer la distribution





Composition du corpus

Base de données SHARE → `http://www.aidsevidence.ca`

- 27 291 documents examinés [L1]
- 1 758 inclus [L3]
- 26 968 instances uniques (aucun doublon)
- Résumés scientifiques retrouvés à partir de PubMed

→ 18 703 instances uniques avec PMID

Distribution des classes dans le corpus

- Instances étiquetées comme **exclues**: 17,402 (**93,05%**)

  Exemples **negatifs** : classe majoritaire

- Instances étiquetées comme **incluses**: 1,301 (**6,95%**)

  Exemples **positifs** : classe minoritaire

- Distribution typique d'une tâche de triage
- Déséquilibre $\Rightarrow$ impact sur les décisions

  [maximisation de l'exactitude générale
  $\Rightarrow$ non prise en compte des instances positives]

Observations sur le corpus

Attribus	**Nombre**	**%**
Nombre total d'instances	18,703	100%
Instances négatives	17,402	93.05%
Instances positives	1,301	6.95%
Mots uniques dans les résumés	31,632	-
Mots uniques dans les titres	6,821	-
Termes MeSH uniques	17,971	-

Méthodologie globale

- Corpus représentatif de la curation de documents sur le VIH
- Étude des facteurs de sous-échantillonnage
- Évaluation de méthodes de sélection de caractéristiques
- Évaluation de différents algorithmes de classification
- Comparaison de configurations
  - meilleure distribution de classes
  - ensemble de caractéristiques le plus discriminant

Stratégie d'apprentissage en contexte déséquilibré

Ensembles d'entraînement

- Plusieurs distributions de classes
- Retrait aléatoire d'instances de la classe majoritaire
- Sous-échantillonnage progressif
- Comparaison des différentes distributions

Ensemble de test

- 15% du corpus
- Sélection aléatoire des instances
- Distribution de la tâche réelle de triage

– instances : ≈7% positives et ≈93% négatives

Extraction de caractéristiques

<AbstractText>AIDS has emerged as a serious public health threat (...) </AbstractText>(...)

<MeshHeadingList>
<MeshHeading>
<DescriptorName (...)>Adolescent </DescriptorName>
</MeshHeading>
<MeshHeading>
<DescriptorName (...)>HIV Infections </DescriptorName>
<QualifierName (...)>etiology </QualifierName>
<QualifierName (...)>prevention control </QualifierName>
</MeshHeading>
</MeshHeadingList>

- Termes MeSH:

  [adolescent, descriptorname] [hiv infections, descriptorname]

  [etiology, qualifiername] [prevention control, qualifiername]

- Sac de mots:

  [aids, 1] [emerged, 1] [serious, 1] [public, 1] [health, 1] [threat, 1]

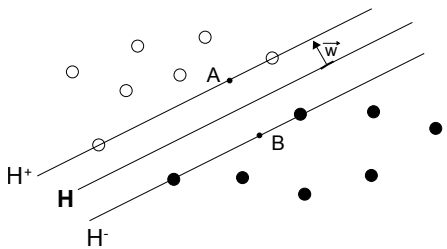
Algorithmes de classification

Réseau bayésien naïf (NB)

- Algorithme de base pour le triage
- Méthode naïve pour évaluer notre approche

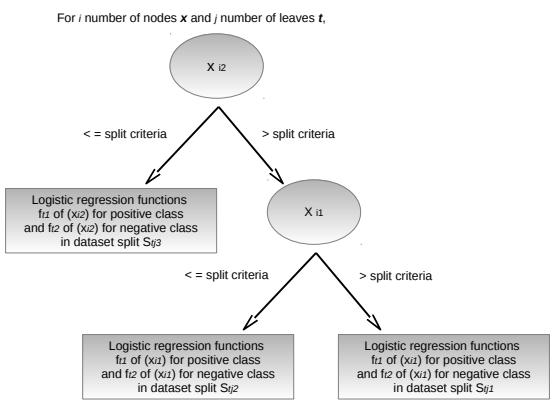
Séparateur à vaste marge (SVM)

- Application fréquente en contexte déséquilibré
  (Mountassir et al., 2012)



Arbre à régression logistique (LMT)

- Très performant en contexte déséquilibré
  (Charton et al., 2013)



Paramètres expérimentaux

Ensembles de caractéristiques

S1: Sac-de-mots (BOW)

S2: Sac-de-mots + termes MeSH

S3: mots clés du domaine

Algorithmes de classification

NB, SVM, LMT

Facteurs de sous-échantillonnage

de 0% (93%NEG 7%POS)

à ≈40% (50%NEG 50%POS)

Métriques pour l'évaluation

- **Précision :** prédictions correctes / toutes les prédictions

$$\text{Précision} = \frac{VP}{VP+FP}$$

- **Rappel :** prédictions correctes / toutes les instances de la classe

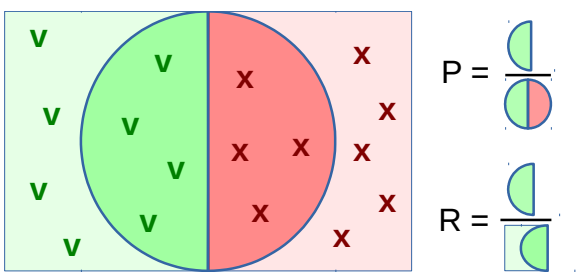
$$\text{Rappel} = \frac{VP}{VP+FN}$$

- **F-mesure:** Moyenne harmonique de la précision et du rappel

$$F = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

- **F-2 score:** Privilégie le rappel

$$\beta = 2, \quad F_\beta = (1+\beta^2) \times \frac{\text{Précision} \times \text{Rappel}}{\beta^2 \times \text{Précision} + \text{Rappel}}$$



Exactitude : prédictions correctes / toutes les instances

$$Accuracy = \frac{VP+VN}{(VP+FN)+(FP+VN)}$$

Modèle de classification de base

- Classifieur bayésien naïf
- Ensemble de caractéristiques #1: BOW résumés + titres

  Espace des caractéristiques : 22 060
- Ensemble d'entraînement sans sous-échantillonnage (7% positives, 93% negatives)

  Représentatif de la tâche de triage
- Performance

Précision: 0.231	F-mesure: 0.365
Rappel: 0.867	F-2 score: 0.560

Meilleurs résultats : modèle 1

- Classifieur LMT
- Ensemble de caractéristiques #2: BOW + termes MeSH

  Espace des caractéristiques : 14 459
- Ensemble d'entraînement avec ≈30% sous-échantillonnage (40% positive, 60% negative)

  Distribution plus équilibrée que dans le scenario original
- Performance

Précision : 0.467	F-mesure: 0.615
Rappel : 0.900	F-2 score: 0.759

Meilleurs résultats : modèle 2

- Classifieur LMT
- Ensemble de caractéristiques #2: BOW + termes MeSH
- Sélection de caractéristiques

  Espace des caractéristiques : 2 411

- Ensemble d'entraînement avec ≈30% sous-échantillonnage (40% positive, 60% negative)

  Distribution plus équilibrée que dans le scenario original

- Performance

Précision : 0.445	F-mesure: 0.591
Rappel : 0.882	F-2 score: 0.737

Comparaison des modèles

	Base	**Modèle 1**	**Modèle 2**
Équilibre	≈ 7% positives	≈ 40% positives	≈ 40% positives
Précision	23.1%	46.7% (+102.16%)	44.5% (+92.64%)
Rappel	86.7%	90.0% (+3.81%)	88.2% (+1.73%)
F-mesure	36.5%	61.5% (+68.49%)	59.1% (+61.92%)
F-2	56.0%	75.9% (+35.54%)	73.7% (+31.61%)
# caract.	22 060	14 459 (-34.46%)	2 411 (-89.07%)

Discussion

- Stratégie d'apprentissage en contexte déséquilibré :
  - Efficace pour réduire le biais
  - Meilleure distribution de classes pour l'entraînement : 40% **positives**
- Ensemble de caractéristiques : termes MeSH + BOW
- Sélection de caractéristiques : réduction de l'espace de recherche
- La majorité des instances **positives** sont correctement étiquetées

Conclusion

- Support pratique pour le triage de la literature
- Système disponible en code ouvert (license MIT)
- Reproductibilité:

  → Nouveaux modèles de triage : nouveaux schémas d'annotation

  → MeSH, UMLS

  `https://github.com/TsangLab/triage`

Remerciements

The authors acknowledge the contributions made by the Ontario HIV Treatment Network (OHTN) and McMaster University, and by the reviewers who have been involved in the development of SHARE. Part of this work was funded by MITACS in partnership with eHealth in Motion Ltd. and Dataparc.

Références

1. Almeida, H., Meurs M.J., Kosseim, L., Butler G., Tsang A.; *"Machine Learning for Biomedical Literature Triage"*. PLoS ONE 9(12), 2014.

2. Almeida, H., Meurs M.J., Kosseim, L., Tsang A.; *"Supporting HIV Literature Screening with Data Sampling and Supervised Learning"* IEEE BIBM 2015, The IEEE International Conference on Bioinformatics and Biomedicine, Washington DC, USA, November 9-12, 2015