Supporting HIV Literature Screening with Data Sampling and Supervised Learning

Hayda Almeida[1], **Marie-Jean Meurs**[1,2], **Leila Kosseim**[3], **Adrian Tsang**[1]

[1] Centre for Structural and Functional Genomics, Concordia University
[2] Dept. of Computer Science, Université du Québec à Montréal
[3] Dept. of Computer Science and Software Engineering, Concordia University
Montréal, QC, Canada

IEEE BIBM 2015

November 10, 2015

Biomedical Literature Screening

- Manual screening: few documents actually kept
- Demanding, time consuming and error-prone
- Not guaranteed to be exhaustive
- Severe bottleneck in manual curation workflow

Machine Learning Approach

Automatic Text Classification

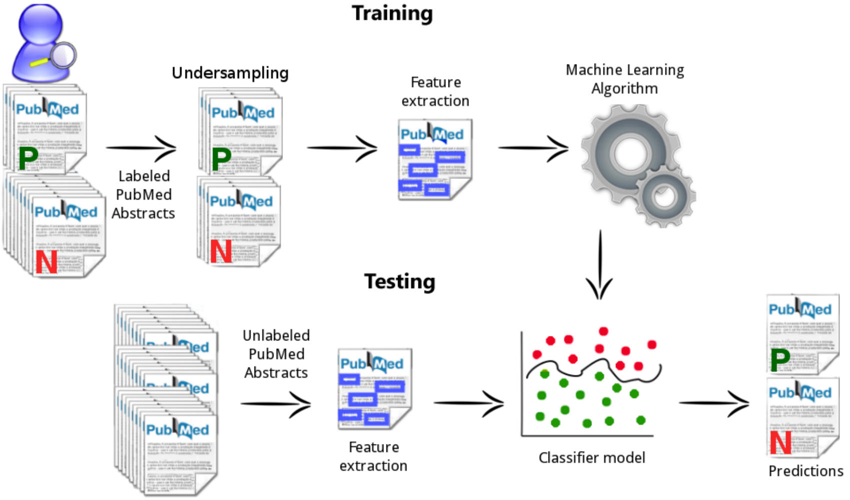
- Release the burden on scientists
- Assess more documents in less time
- Less likely to miss potential information

Goal: reduce effort by flagging candidate documents

Supervised Learning Triage

- Document collection: correct labels
- Training examples: given to classifier
- Model: used to classify new documents
- Test examples: evaluation of model

Supervised Learning Triage



Challenges: Text Classification for Biomedical Triage

1. Imbalanced Class Distribution
   - Large dataset → few relevant documents
   - Non-relevant majority introduces noise
   - Class distribution → affects performance
   - Need to reduce bias
2. Large Feature Space
   - Excessive features causes overfitting
   - Low discriminative value → poor contribution
   - More features → more computational resources
   - Need to identify best subset for the task

Data Sampling

- Selection of a specific subset of the dataset
(Chawla et al., 2002)(Japkowicz, 2000)
- Implementation → pre-processing step
- Less restrictive and less resource-demanding

Dataset Composition

SHARE Database references → http://www.hivevidence.ca

- 27,291 fully reviewed [L1]
- 1,758 included [L3]
- 26,968 unique instances (no duplicates)
- Scientific abstracts retrieved from querying PubMed

→ 18,703 unique instances with PMID

Dataset Balance

- Instances labeled as **excluded**: 17,402 (**93.05%**)

  Negative examples → Majority class

- Instances labeled as **included**: 1,301 (**6.95%**)

  Positive examples → Minority class

- Underlying distribution → real scenario of triage task
- Imbalance affects decision boundary

Dataset Statistics

Attribute	Number	%
Total number of instances	18,703	100%
Negative instances	17,402	93.05%
Positive instances	1,301	6.95%
Unique words in paper abstracts	31,632	-
Unique words in paper titles	6,821	-
Unique MeSH terms in papers	17,971	-

Methodology Overview

- Representative dataset of HIV screening task
- Study of undersampling factors
- Application of different feature settings
- Evaluation of feature selection methods
- Use of off-the-shelf classification algorithms (WEKA)
- Comparison of various supervised learning models

Imbalanced Learning Strategy

Training sets

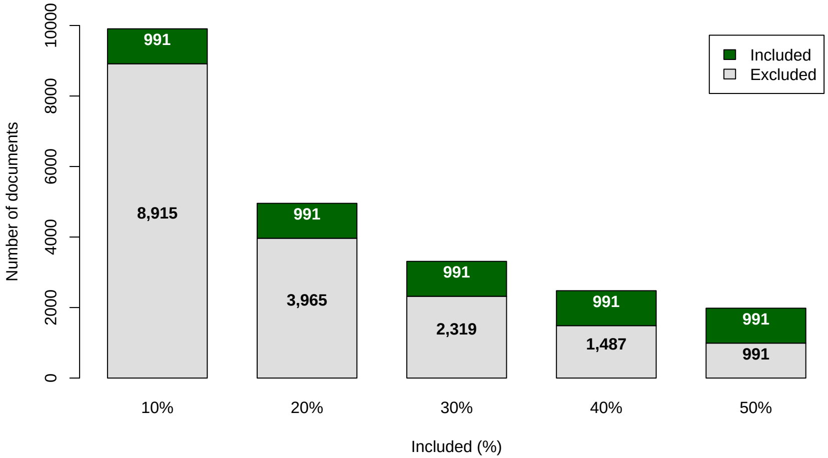
- Variety of class distributions
- Random removal of majority instances
- Progressive undersampling of 10%
- Comparison of different class balances

Test set

- 15% of the document collection
- Random selection of instances
- Real class distribution of triage task

≈7% positive and ≈93% negative instances

Undersampling Factors



Feature Extraction

<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>

- MeSH Terms:

  [adolescent, descriptorname] [hiv infections, descriptorname]

  [etiology, qualifiername] [prevention control, qualifiername]

- Bag-Of-Words:

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

Feature Selection Strategy

Odds Ratio (OR)

- Occurrence of features in positive class
- Confidence interval (CI) of 95% for each score
- Discard features if:
  - $\rightarrow$ CI contains the null hypothesis (1.0)
  - $\rightarrow$ OR score $<=$ null hypothesis (1.0)

Inverse Document Frequency (IDF)

- Occurrence of features in both classes
- Discard features if:
  - $\rightarrow$ IDF score $<= 1.0$ (i.e. Occurrence ratio is $>$ 10:1)

Features: Dataset Representation

- Dataset instances → feature vectors
- Feature occurrence in documents
- Training and Test sets → Feature x Document matrix

MeSH Terms vector

adolescent	*hiv infections*	*etiology*	*prevention control*	...
1	1	1	1	...

Bag-Of-Words vector

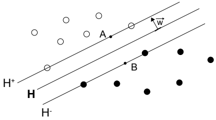
aids	*emerged*	*serious*	*public*	*health*	*threat*	...
1	1	1	1	1	1	...

Classification Algorithms

Naïve Bayes (NB)

- Baseline for triage task
- Naïve method to evaluate our approaches

Support Vector Machine (SVM)



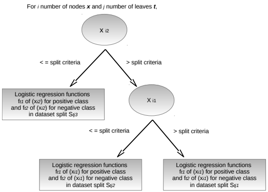
- Commonly applied in tasks with imbalanced data

(Akbani et al., 2004), (Tang et al., 2005),(Mountassir et al., 2012)

Logistic Model Trees (LMT)

- Described as suitable for imbalanced data

(Charton et al., 2013)



Experimental Settings Overview

Sets of Features

S1: Bag-Of-Words (BOW)

S2: Bag-Of-Words + MeSH Terms

S3: Domain Keywords list

Feature selection metrics

Inverse Document Frequency, Odds Ratio

Classification algorithms

NB, SVM, LMT

Undersampling Factors

From 0% USF (93%NEG 7%POS)

to $\approx$40% USF (50%NEG 50%POS)

Evaluation Metrics

- **Precision:** Correct output / all predictions

$$Precision = \frac{TP}{TP+FP}$$

- **Recall:** Correct output / class instances

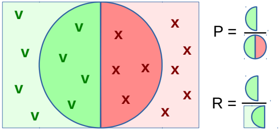
$$Recall = \frac{TP}{TP+FN}$$

- **F-measure:** Harmonic mean of Precision and Recall

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- **F-2 score:** Emphasis on Recall measure

$$\beta = 2, \quad F_\beta = (1+\beta^2) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall}$$



Accuracy: Correct output / all instances

$$Acc = \frac{TP+TN}{(TP+FN)+(FP+TN)}$$

Baseline Classification Model

- NB classifier
- Set of features #1: BOW of abstract and title contents

  Feature space: 22,060
- Training set with 0% USF (7% positive, 93% negative)

  Representative of the real triage task
- Performance

  Precision: 0.231 F-measure: 0.365

  Recall: 0.867 F-2 score: 0.560

Best results: HM1 model

- LMT classifier
- Set of features #2: BOW + Mesh Terms

  Feature space: 14,459
- Training set with ≈30% USF (40% positive, 60% negative)

  More balanced distribution than real scenario
- Performance

Precision: 0.467	F-measure: 0.615
Recall: 0.900	F-2 score: 0.759

Best results: HM2 model

- LMT classifier
- Set of features #2: BOW + Mesh Terms
- Feature selection: Odds Ratio

  Feature space: 2,411

- Training set with ≈30% USF (40% positive, 60% negative)

  More balanced distribution than real scenario

- Performance

Precision: 0.445	F-measure: 0.591
Recall: 0.882	F-2 score: 0.737

Model comparison

	Baseline	**HM1**	**HM2**
Balance	≈ 7% positive	≈ 40% positive	≈ 40% positive
Precision	23.1%	46.7% (+102.16%)	44.5% (+92.64%)
Recall	86.7%	90.0% (+3.81%)	88.2% (+1.73%)
F-measure	36.5%	61.5% (+68.49%)	59.1% (+61.92%)
F-2	56.0%	75.9% (+35.54%)	73.7 (+31.61%)
# features	22,060	14,459 (-34.46%)	2,411 (-89.07%)

Discussion

- Imbalanced learning strategy:

  Valuable to reduce bias effect

  Most fitting training class distribution → 40% **included**

- Set of features → MeSH Terms + BOW
- Odds Ratio → effective to narrow feature space
- Majority of **included** instances were correctly labeled

Observations

- Practical support for literature triage
- Open-source → system toolkit publicly released under MIT license
- Reproducibility:
  - → New triage models: wide-ranging annotation schemas
  - → MeSH, UMLS

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

Partner Organizations

Dataparc, eHealth in Motion Ltd.

Acknowledgment

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.

References

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., Butler G., Tsang A.; *"Data Undersampling for Scientific Literature Triage."* High Performance Computing Symposium (HPCS 2015).

3. **Almeida, H.**, Meurs M.J., Kosseim, L., Butler G., Tsang A.; *"Biomedical Literature Triage using Supervised Learning".* 2nd Workshop on Machine Learning for Clinical Data Analysis, Healthcare and Genomics (NIPS 2014)

4. **Almeida, H.**, Meurs M.J., Kosseim, L., Butler G., Tsang A.; *"A Machine Learning Approach for mycoCLAP Triage."* Biocuration, 2014.