mycoCLAP, the Database for Characterized Lignocellulose-Active Proteins of Fungal Origin: Resource and Text Mining Curation Support

Kimchi Strasser, Erin McDonnell, Carol Nyaga, Min Wu, Hayda Almeida, Marie-Jean Meurs, Leila Kosseim, Justin Powlowski, Greg Butler, Adrian Tsang

Centre for Structural and Functional Genomics

and

Computer Science and Software Engineering
Concordia University, Montreal, Canada

gregb@cs.concordia.ca

Biocuration 2015 — 26 April 2015

Outline

What! Only 10 minutes!

I hope you spoke with Kimchi at the poster session for details :-)

... or read the paper in Database: Biocuration Virtual Issue

mycoCLAP

... is database ... ongoing curation ... now over 800 proteins

mycoSORT — *Breaking news: new and improved*

... for triage of articles

mycoMINE

... text mining relevant articles

mycoCLAP Curation

Searchable online database of fungal enzymes

Industrial processes 

Manual curation since 2011

Extensive review of scientific publications

804 enzymes from 226 fungal species | one^{+} reference paper by entry

PubMed  *"fung*"* $\rightarrow$ over 250,000 documents

https://mycoclap.fungalgenomics.ca

Literature Triage

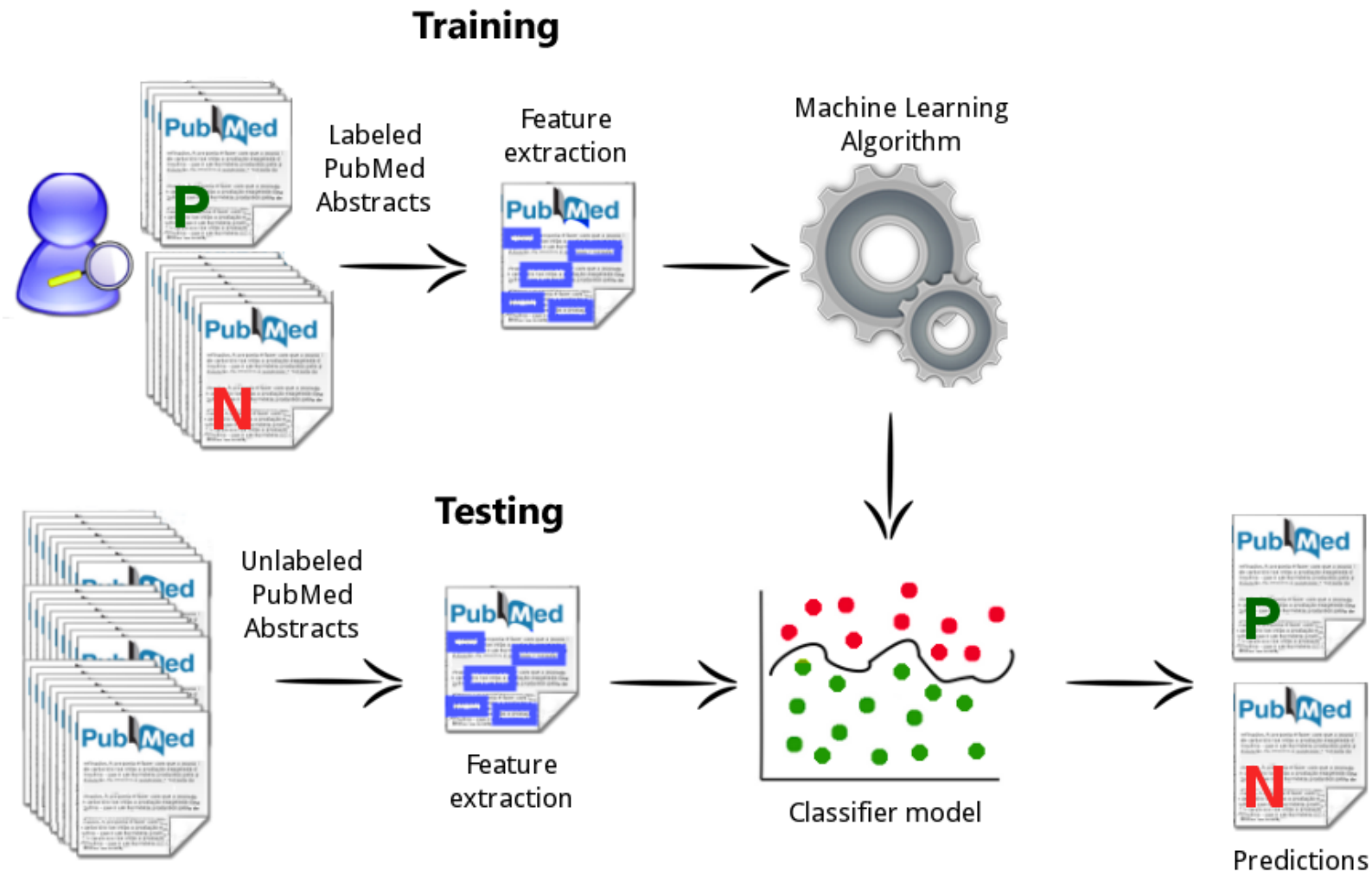
Manual screening: few documents actually kept

Demanding, time consuming and error-prone

Not guaranteed to be exhaustive

Severe bottleneck in manual curation workflow

Supervised Learning Workflow for mycoSORT Triage



Dataset Balance

4 biocurators with inter-annotator agreement $> 80\%$

Instances labeled as **non-relevant**: 6,834 (**90.12%**)

Negative examples $\rightarrow$ Majority class

Instances labeled as **relevant**: 749 (**9.88%**)

Positive examples $\rightarrow$ Minority class

Underlying distribution $\rightarrow$ real scenario of triage task

Imbalance affects decision boundary

Baseline (Naive Bayes) vs. mycoSORT Performance

mycoSORT

- Classifier is Logistic Model Trees
- Under Samplng Factor is 40%
- Feature selection strategy is Odds Ratio

Scores	**Baseline**	**mycoSORT**
Precision	0.307	0.368 (+19.8%)
Recall	0.720	0.860 (+19.4%)
F-measure	0.430	0.515 (+19.7%)
F-2	0.570	0.680 (+19.3%)

In practice, ... for 1000 abstracts ... where 900 irrelevant

Baseline triage keeps $72 + 162 = 234$ and eliminates $738 + 28$

mycoSORT triage keeps $86 + 147 = 233$ and eliminates $753 + 14$

Thank You!

Questions Please?