Active Learning for Phenotyping Tasks

Dmitriy Dligach, Timothy A. Miller, and **Guergana Savova**

Boston Children's Hosptial and Harvard Medical School

June 21, 2016

Introduction

- Phenotyping
  - What's a phenotype?
  - i2b2 and eMERGE
  - Link EHRs to biobanks for genetic analysis
  - Supervised learning for phenotyping
- Manual annotation needed
  - Standard approach: passive learning
  - Alternative: active learning

Active Learning

- Approach for selecting data for annotation
- Data selection delegated to classifier
- Pool-based scenario
  - Lots of unlabeled data
  - Can afford to annotate only a small amount
- Little work in clinical domain

Intuition

Suppose there's a little bit of labeled data

- Classify example $\vec{x}$
  - $p(c_1|\vec{x}) = 0.95$ and $p(c_2|\vec{x}) = 0.05$
  - $p(c_1|\vec{x}) = 0.55$ and $p(c_2|\vec{x}) = 0.45$
- Margin Sampling
  - $PredictionMargin = |P(c_1|\vec{x}) - P(c_2|\vec{x})|$
  - Annotate examples with smallest margin first

How does active learning work?

- Seed classifier
  - Annotate a small amount of data
  - Train a classifier
- Iterative process
  - Apply the classifier to the pool of unlabeled data
  - Select an example and add it to the training set
  - Retrain the classifier
  - Check if we are done
- The learner quickly converges on the decision boundary

Data Representation

- Unit of classification
  - Single patient
- Patient representation
  - Set of CUIs extracted with cTAKES
  - Abstract from lexical variability of medical terminology
  - Filter out non-clinical vocabulary
- Phenotype-specific dictionaries
- Patient vector $\vec{x}$
  - Element x_n is frequency of CUI_n

Naive Bayes

- Need to evaluate $p(c_i|\vec{x})$
- Multinomial Naive Bayes
  - Probabilistic classifier
  - Supports multi-class classification
  - Training and classification speed
- Uncertainty sampling:

$$\textit{prediction margin} = |p(c_1|\vec{x}) - p(c_2|\vec{x})| \tag{1}$$

Counts

Compute posterior probability as follows:

$$p(c_i|\vec{x}) = \frac{1}{Z} p(c_i) \prod_{n=1}^{N} p(CUI_n|c_i)^{x_n} \quad (2)$$

$p(c_i)$ - prior probability of class c_i
N is the number of CUIs in the phenotype-specific dictionary
CUI_n is the n_{th} CUI in that dictionary
x_n is the frequency of CUI_n in $\vec{x}$
Z (evidence) is the scaling factor
Determine $p(c_i)$ and $p(CUI_n|c_i)$ via maximum likelihood estimation

Dataset creation

- Created within the i2b2 initiative
- ICD-9 codes used to form initial cohort
- About 600 patients selected randomly
- Labeled by domain experts

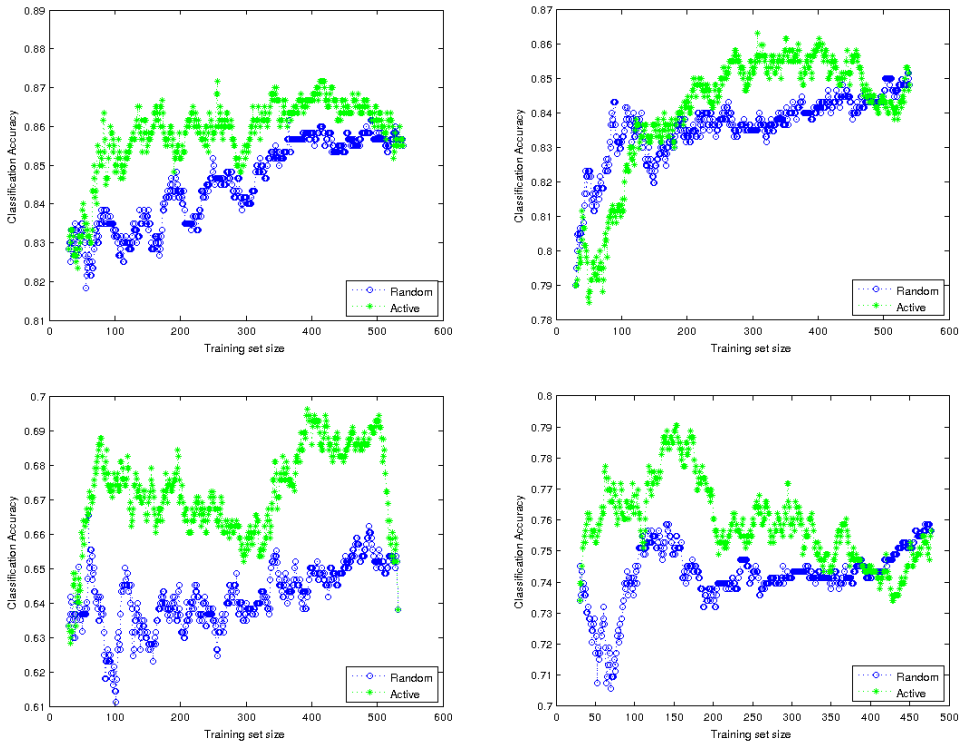
Dataset stats

Phenotype	Total Instances	Number of Classes	Proportion of Predominant Class
Ulcerative Colitis	600	2	0.630
Crohn's Disease	600	2	0.665
Multiple Sclerosis	595	5	0.395
Type II Diabetes	600	3	0.583

Evaluation

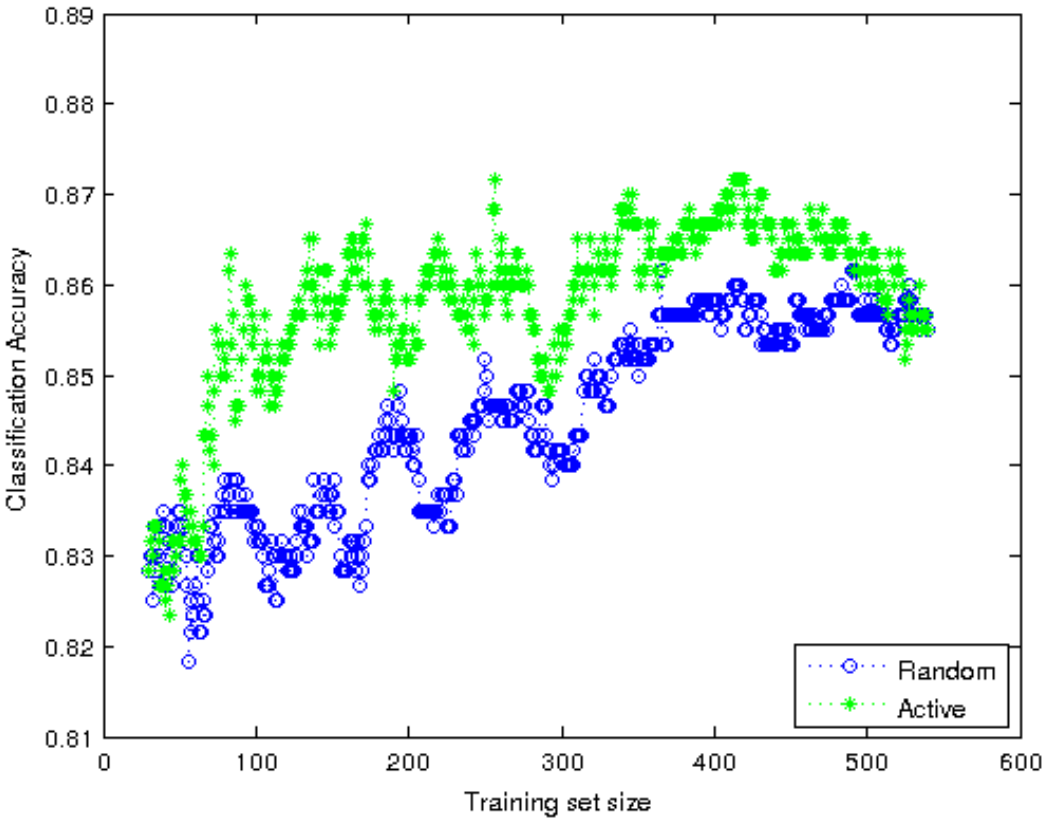
- Learning curve generation
  - Done in the style of 10-fold cross validation
- Within each fold:
  - Training data
  - Pool of "unlabeled" examples
  - Held-out test set
- Various seed sizes
  - Affect of seed size and performance
  - Only showing the plots for seed size $= 30$
  - See the paper for other sizes
- Gold labels in the pool hidden from classifier

Learning Curves



Ulcerative Colitis, Crohn's Disease, Multiple Sclerosis, Type II Diabetes

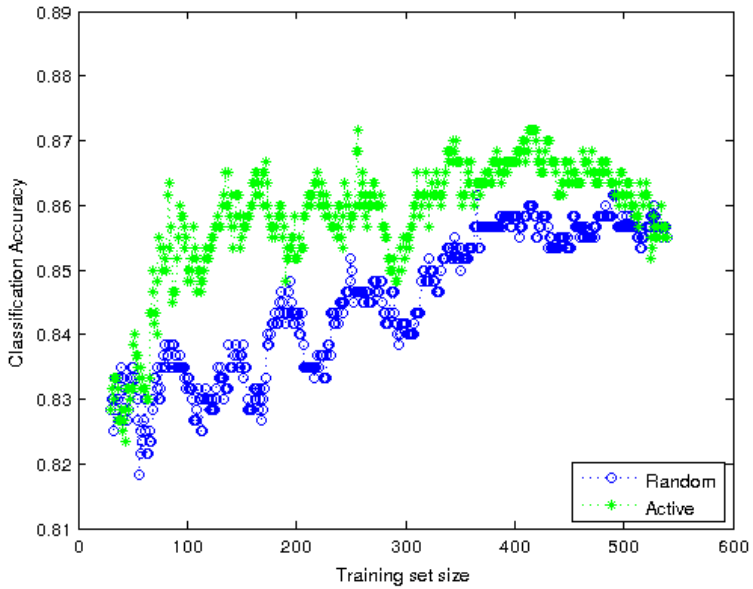
Close-up



Ulcerative Colitis

Sample plot

- Active Learning above passive
- Only need 1/3 of the data
- Best performance higher



Ulcerative Colitis

Difference between areas under the curve (Active - Passive)

Seed Size	Ulcerative Colitis	Crohn's Disease	Multiple Sclerosis	Type II Diabetes
10	6.90	4.17	10.50	11.05
30	6.64	2.21	15.43	7.49
50	8.63	1.75	8.61	8.90

Conclusion

- Annotation effort reduced by 2/3
- Active learning sometimes reaches better accuracy
- Need to know when to stop
- What happens if the base classifier is swapped?

Questions?