Machine Learning for Automated Classification of Patient Cases

Cole Deisseroth, Jon Wang, James Bai

Stanford MEDICINE

Background

Much of high-quality practice and decision-making depends on one physician. Sending patients to the right department is important to save physicians time and help patients find treatment [1]. This is an initial study toward the development of an intelligent patient-allocation system. This serves to save medical personnel valuable time and help patients find the care they need more efficiently by automatically categorizing cases into specific departments. We develop an algorithm which predicts the categories of patient cases from the American Board of Internal Medicine Examinations—a certification that all physicians must go through to practice general medicine.

Ontology

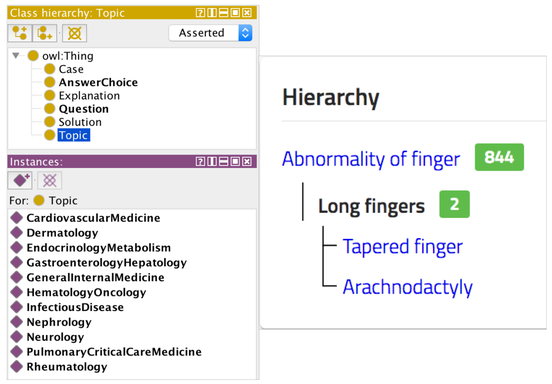

Figure 1 (left): An ontology of our own design that details the components of an ABIM question. Each question has a patient case description, and a Topic (we must predict the latter).
Figure 2 (right): A snapshot of the Human Phenotype Ontology (HPO) [2] which we use to standardize the disease phenotypes mentioned in patient cases

Problem Solving Method

We used ClinPhen[3], a recently published tool that automatically extracts HPO phenotypes from free text, to tokenize the phenotypes for each question.
We built a Naïve Bayes classifier that uses Bayes' theorem to predict the categories:

$$P(X|a,b,c) = P(X)P(a|X)P(b|X)P(c|X) / P(a)P(b)P(c)$$

Assumptions:

1. Categories are mutually exclusive
2. Each observation is independent,
3. Categories are exhaustive.

Probabilistic Graph Models:

1. All words in the cases
2. All Human Phenotype Ontology terms
3. All Human Phenotype Ontology Closures.

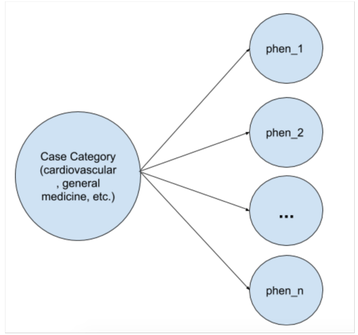

Figure 3: Every case category is associated with how often its patients have each HPO phenotype.

Evaluation

- Objectivistic Summative Evaluation:
  - 3,421 patient cases. 3,081 for training Testing set (340).
  - We then trained the Naïve Bayes classifier using the Training set, and tested its predictions on the testing set.
  - We measured accuracy as the percentage of Test cases for which the classifier predicted the correct category.
- Quantitative metrics of our models included accuracy and macro-averaged precision, macro-averaged recall, and macro-averaged AUROC.
- Subjectivistic Summative Evaluation:
  - We manually analyzed passages that were inaccurately categorized and scanned for any words or phrases that could have confounded the algorithm.

Results

	All Words	Phenotype Terms	Phenotype Closures
Accuracy	80%	56%	56%
Precision	0.81	0.57	0.63
Recall	0.82	0.58	0.58
F1	0.80	0.55	0.53
AUROC	0.98	0.89	0.88

Table 1. All words outperforms on all evaluation metrics.

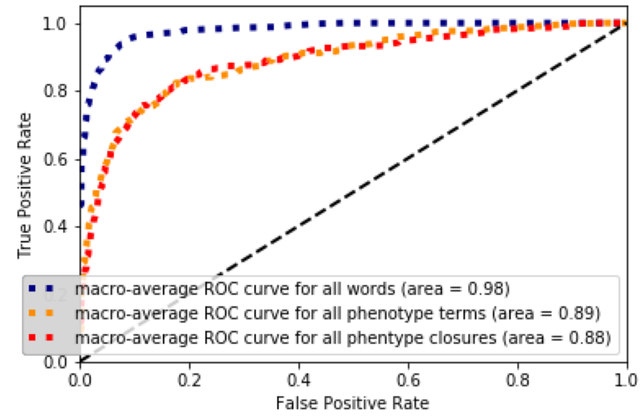

Figure 4. ROC curve demonstrates sensitivity/specificity tradeoff

Qualitative Analysis
Manual inspection revealed two main errors::
1) Generic terms (e.g. "Phenotypic Abnormality") competing with the low frequency of specific terms (e.g. "Seizures")
2) Non-specific phenotypes identified for patient cases.

Discussion/Future Work

- 80% of the time, the classifier put the question into the right category.
- The all-words classifier outperforms the other two classifiers in all metrics
- Advantage: The algorithm uses real descriptions from the medical board exams that reflect hypothetical patient descriptions.
- Drawback: assumes all phenotypes are independent of one another, when in reality they are not.
- Future work:
  - leveraging up-to-date knowledge graphs and building our own graphs to categorize these cases.
  - training on real-world clinical data
  - naive bayes model combining all words and HPO terms
  - neural network implementation using NLP
  - Leveraging embedding space to find similar terms

Division of labor

- Jon: thought of project, curated data, evaluation
- Cole: developed Naïve Bayes classifier, used ClinPhen to parse the patient cases for phenotypes.
- James: literature review, validation/labeling of data, subjectivistic error analysis.

References

References:
[1] J. B. McKinlay, C. L. Link, K. M. Freund, L. D. Marceau, A. B. O'Donnell, and K. L. Lutfey, "Sources of variation in physician adherence with clinical guidelines: results from a factorial experiment," J. Gen. Intern. Med., vol. 22, no. 3, pp. 289–296, 2007.
[2]Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. American journal of human genetics, 83(5), 610-5.
[3]Deisseroth CA, Birgmeier J, Bodle EE, Kohler JN, Matalon DR, Nazarenko Y, Genetti CA, Brownstein CA, Schmitz-Abe K, Schoch K, Cope H, Signer R; Undiagnosed Diseases Network, Martinez-Agosto JA, Shashi V, Beggs AH, Wheeler MT, Bernstein JA, and Bejerano G (2018). ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genetics in Medicine, 2018. DOI: 10.1038/s41436-018-0381-1