Project Proposal: Enhancing Clinical Code Assignment with Active Learning

Group Members: Zhuoqun Li, Yuning Zheng, Grace Du

What are you trying to do?

We aim to investigate the impact of active learning strategies on the efficiency and performance of clinical code assignment using electronic health records (EHR). Specifically, we will explore how different active learning acquisition functions influence model learning rates when applied to the MIMIC-III dataset for ICD-9 code classification.

How is it done today, and what are the limits of current practice?

Current approaches to clinical code assignment rely on supervised machine learning models, particularly deep learning-based multi-label classifiers. These models require extensive labeled training data, which is costly and time-consuming due to the need for domain expertise. Manual annotation remains a bottleneck in scaling such models, and conventional supervised learning techniques often struggle with class imbalance and data scarcity issues. Active learning presents a potential solution by strategically selecting the most informative instances for labeling, reducing annotation costs while maintaining model accuracy. However, limited studies have systematically analyzed the effects of various active learning strategies on ICD-9 classification performance.

What is new in your approach and why do you think it will be successful?

We propose a comparative study of multiple active learning strategies, including uncertainty-based methods, diversity-based methods, Query-by-Committee**,** and other advanced active learning approaches. By systematically evaluating these strategies on ICD-9 code assignment tasks with the passive learning method, we expect to identify techniques that optimize the balance between annotation cost and predictive performance. We hypothesize that the appropriate acquisition function can further reduce data requirements while improving generalization.

Who cares?

Healthcare providers, medical coders, and researchers working on EHR-based clinical decision support systems stand to benefit from more efficient annotation methods. Reducing the burden of manual labeling while maintaining or improving model accuracy can accelerate the adoption of AI-assisted medical coding, ultimately benefiting hospital workflows and patient care.

If you are successful, what difference will it make?

A successful outcome would demonstrate that active learning can significantly decrease annotation effort while preserving or enhancing model performance. This could lead to more

scalable and cost-effective solutions for automated medical coding, benefiting both research and practical clinical applications.

What are the risks?

- **Class Imbalance:** ICD-9 codes exhibit a long-tailed distribution, which may bias model performance toward more common labels.
- **Noise in Annotations:** Clinical notes contain ambiguous language, potentially affecting the reliability of labels.
- **Computational Constraints:** While active learning aims to reduce training data requirements, certain selection strategies may be computationally expensive, requiring careful optimization.

How much will it cost?

The project will leverage publicly available data (MIMIC-III) and open-source machine learning frameworks, minimizing financial costs. Computational expenses will be managed by running experiments on CPU-friendly models with a subset of ICD-9 codes.

How long will it take?

The project will be conducted within the semester time frame, with key milestones aligning with course deliverables:

- **Project Proposal Submission:** March 13
- **Midpoint Progress Report:** April 6
- **Final Report & Presentation:** April 24 – May 2

What are the mid-term and final "exams" to check for success?

- **Mid-term Evaluation:** Performance assessment of preliminary active learning strategies on a subset of MIMIC-III discharge summaries.
- **Final Evaluation:** A full comparative study of active learning methods, measuring accuracy, F1-score, and data efficiency against a supervised learning baseline.

Methodology

Dataset

We will use the **MIMIC-III clinical dataset**, specifically focusing on the discharge summaries for **ICD-9 code assignment**. We will preprocess the text data using TF-IDF and Word2Vec embedding techniques and apply machine learning models for multi-label classification.

Active Learning Strategies

We will experiment with different acquisition functions:

1. **Uncertainty Sampling:**
    - Least Confidence
    - Binary Entropy
    - Smallest Margin
2. **Diversity Sampling:**
    - k-Means Clustering (random, centroid, and border selection)
3. **Query-by-Committee (QBC)**
4. **Expected Model Change**
5. **IWAL (Type I method)**
6. **ZLG, DH, PLAL (Type II methods)**
7. **Hybrid Methods (optional):**
    - Two-Stage Strategy (initial clustering followed by uncertainty selection)
    - Weighted Uncertainty Metric (combining feature similarity and uncertainty)

Evaluation Metrics

1. Micro & Macro F1-Score
2. Data Efficiency (labeled instances required to reach target performance)
3. Computational Overhead (runtime per iteration)

Expected Outcomes

We expect to show that certain active learning strategies, particularly hybrid approaches, can significantly reduce annotation needs while preserving model performance. The insights gained will inform future applications of active learning in clinical NLP tasks.

References

1. Bejan, C. A., Vanderwende, L., & Xia, F. (2013). Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association*, 20(e2), e253–e259.

2. Kaur, R., & Kumar, M. (2022). AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review. *Expert Systems with Applications*, 202, 117173.

3. Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., & Liu, H. (2016). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77, 34–49.

4. Ferreira, M. D., Malyska, M., Sahar, N., Miotto, R., Paulovich, F., & Milios, E. (2021). Active learning for medical code assignment. *arXiv preprint arXiv:2104.05741*.

Feedback

- It would be important to include deep learning based approaches, where you explore different way to measure uncertainty. Most of the query selection methods are already implemented through homework assignment so it would be important to add some new ones.
- It would be interesting to explore selection strategies to take into account ICD-9 code rarity.
- It is unclear if you are planning to evaluate the query selection strategies only one at a time or on a batch setting. I leave it up to you to decide which setting is more realistic in this scenario.
- It would be important to include some downstream analysis on which ICD-9 codes are harder to predict correctly as well as other domain-specific aspects you find in your study.
- Similarity of batches
- It would be really interesting to explore the performance of the query selection strategies under different noise models. That is, the ICD-9 codes are not exactly correct.