# CS5339 Project – Active Learning

Uncertainty Sampling Query Strategies Literature Review

Ma Yuan E0674520

March 22, 2021

## 1 Introduction

Data labeling requires expensive labor or various costs. In the face of massive unstructured data, how to label it economically and accurately is a thorny issue. As a method of constructing an effective training set, the active learning algorithm with a small amount of labeled data let the machine Learning model effectively interactive with annotation expert(oracle), thereby choosing the most informative samples and effectively reducing the amount of labeled data needed to model Learning. To overcome the labeling bottleneck, active learning is widely used such as Speech recognition, Information extraction, Classification and filtering [3].

In this report, firstly, we will discuss about the overview of active learning. This includes 2 essential building blocks of active learning which are the *Scenarios* and *Query strategies* [1].Then we will mainly focus on the most commonly used query strategies *Uncertainty Sampling* which proposed by Lewis and J. Catlett in 1994 [2, 3]. Under this topic, we will also review the proposal suggested by J. Zhu, H. Wang in how to encountered the outliers problem [4].

## 2 Description of Active learning

In this section, we will brief the active learning process. In the conventional modeling process, it usually includes several steps: sample selection, model training, model prediction and model update. In the field of active learning, two more steps are introduced, extraction of label candidate set and annotation by labeling "experts"(oracle) [1].

## 2.1 The Active Learning Process

To perform an active learning, we define the active learning model as

$$A = (C, Q, S, L, U)$$

where $C$ is the classifier model, $L$ represents the labeled sample set, $S$ stands for annotation "experts"(oracle), $Q$ is the query strategy in use and $U$ represents the unlabeled datasets. The flowchart [3] can be interpreted as the following steps(take classification as an example):

1. Select the appropriate classifier as $C$ and query strategy as $Q$.
2. Split labeled sample $L$ datasets into *train_sample* used to train model and *validation_sample* used to verify the current model performance. And prepare unlabeled dateset *active_sample* $U$.
3. Initialization: Random initialization or source domain initialization. If there are labeled samples of the target domain, train the model through these labeled samples;
4. Use the current model $C$ to predict the samples in *active_sample* one by one (the prediction does not require labels), and get the prediction result of each sample.
5. Here, we choose *Uncertainty Sampling* strategy as illustration to measure the labeled value of the sample. The sample with the predicted result closer to 0.5 indicates that **the current model has higher uncertainty, that is, the higher the value of the sample that needs to be labeled**.
6. The expert annotates the selected samples and append the annotated samples into *train_sapmle*.
7. Retrain and update the model $C$ use the new *train_sapmle*.
8. Use current model $C$ to verify the *validation_sample*. If current model $C$ reaches the target or no longer continue to label new samples (no experts or no samples), the iterative process ends. Otherwise, repeat steps (4)-(7)

For following discussion and experiment, a fully labeled datasets is prepared which, in other words, the oracle has labeled all samples although, in practice, the oracle annotates the sample in runs. So in this report our major discussion will focus on the strategy of how to extract the most informative sample to be annotated from the unlabeled datasets $U$.

In order to extract most valuable sample in each run, we need two components, a *scenario* to be pose the quires from and a *query strategy* which chooses the sample based on the scoring metric.

### 2.2 Scenario

In this section, three active learning scenarios are introduced

### 2.3 Query Strategy Framework

### 2.4 The Algorithm

### 2.5 Experimental Examples

## 3 Mathematical Analysis

## 4 Extensions and Further Results

In this section, we briefly outline some more advanced algorithms and theory building on the previous sections. Due to space limitations, we only provide a short discussion on each of these.

### 4.1 Base Learners Beyond Decision Stumps

### 4.2 Multi-Class Boosting

### 4.3 Characterization of the Test Error

# References

[1] Active learning (machine learning). https://en.wikipedia.org/wiki/Active_learning_(machine_learning).

[2] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12. ACM Springer.

[3] Burr Settles. Active learning literature survey. *Computer Sciences Technical Report 1648*.

[4] J. Zhu, H. Wang, B. K. Tsou, and M. Ma. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1323–1331, 2010.

# Appendix

## A Figures

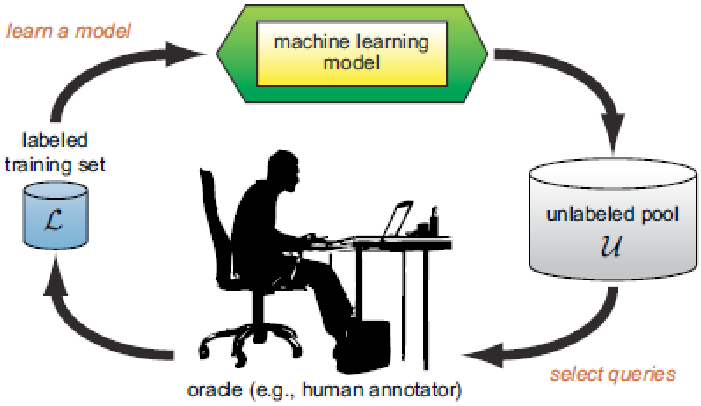

Figure 1: The pool-based active learning cycle.