Homework 1: Three Papers about Active Learning
Fatemeh Rahbari

Paper 1: *Active Learning Literature Survey (2009)*

1. What problem does this paper try to solve, i.e., its motivation?

The first objective of this paper is to address the general challenge of minimizing the costs and time associated with labeling data for machine learning models. In the context of numerous machine learning applications, gaining labeled data is hard to mange or expensive, while unlabeled data is often accessible. The paper tries to introduce active learning as a feasible solution to this issue. Active learning empowers machine learning algorithms to improve performance by wisely selecting and querying the most informative data points from the pool of unlabeled data. So, active learning enables the algorithm to force needing fewer labeled sampels, so the labeling costs and time requirements are reduced notably.

2. How does it solve the problem?

The main concept is that a machine learning algorithm can achieve higher accuracy with fewer labeled training samples if it can select the data wisely from which it learns. The proposed solution in this paper has several methods. First is "Active Querying". In active querying, the algorithm selects the most informative samples from the data instead of using randomly labeled data passively. Secons id "Uncertainty Sampling". The algorithm queries samples about which it is least certain to find the most valuable information from a limited number of labeled samples. Third is " Query-By-Committee". In query by committee, multiple models (committee) trained on the current data vote on a label, and the algorithm selects samples where there is the most disagreement among the models. In the end, there is "Expected Model Change". The algorithm queries samples that are expected to cause the greatest change in the model if their labels were known, intending to improve the model's learning efficiency. These techniques reduce the number of labeled samples required to train an effective machine learning model.

3. A list of novelties/contributions.

The paper has a detailed literature survey of active learning, discussing key methods. It introduces various query strategies, such as uncertainty sampling, query-by-committee, expected model change, and density-weighted methods. The paper investigates empirical and theoretical aspects of active learning supporting the efficiency of this method compared to traditional supervised learning approaches. Also, It investigates a broad spectrum of problems: like active learning for structured outputs, batch-mode active learning, and cost-sensitive active learning. In the end, it discuss real-world applications and scenarios where active learning can be beneficial and applicable, including text classification, image retrieval, and speech recognition.

4. What do you think are the downsides of the work?

Active learning raises several challenges. One of the biggest downsides is the high level of computational complexity associated with specific query strategies, mainly those like the Fisher information ratio and estimated error reduction. These methods are computationally expensive and time-consuming, especially when dealing with high-dimensional feature spaces. Also, the desired result of active learning depends on the specific model being used. If the model changes, such as advancements in machine learning, the actively constructed training set may no longer be optimal, which can negatively impact its ability to generalize beyond the training data. Additionally, the paper highlights the possibility of active learning strategies leading to the selection of outliers or samples that do not accurately represent the overall data distribution. This can negatively affect the overall model performance. In the end, a challenge can be a considered task like focusing on simple classification tasks in active learning research. Applying active learning to more complex tasks, such as structured output prediction, remains difficult due to the lack of research and verified methodologies.

Paper 2: *From Theories to Queries: Active Learning in Practice (2011)*

1. What problem does this paper try to solve, i.e., its motivation?

This paper addresses the practical challenges of applying active learning techniques in real-world machine learning tasks and applications. The main motivation of the paper is to make active learning more cost-worthy and efficient by allowing machine learning algorithms to obtain their own training data through queries selectively, which can reduce the number of labeled examples needed (Which is a costly operation). However, the paper also highlights a key problem, while active learning has shown theoretical and experimental success in reducing training set sizes, deploying it in real-time systems has led to unexpected issues. These problems arise because real-world conditions often don't satisfy the simplified assumptions made in earlier research, such as perfect annotators or consistent labeling costs. In summary, the paper tries to address several challenges. The first challenge is about batch querying. It is a challenge that many applications require batch labeling, rather than labeling one sample at a time. The second focuses on human annotators or empirical experiments that may introduce label noise, affecting data quality. Also, labeling costs are not consistent and can vary across samples and tasks. In the end, active learning may bias training data based on the chosen model, which could be problematic if model conditions change.

2. How does it solve the problem?

To address the need for querying multiple samples at once (rather than one by one), the paper suggests several algorithms that can select different and informative batches of data points. Techniques like clustering or optimization approaches (e.g., Bayesian experimental design) are used to ensure the batch includes both mixed and informative examples, reducing redundant labeling and improving efficiency. Also, agnostic active learning is introduced to allow for noisy labels, a solution for errors from the oracle. Another method is querying the same sample multiple times to de-noise the data, averaging out the noise by gathering multiple labels for the same sample. To overcome the uneven costs of labeling, the paper proposes learning a cost model alongside the main learning model.

Additionally, the paper suggests querying for features instead of samples. To overcome the bias introduced by selecting data specific to a certain model, the paper suggests using heterogeneous ensembles. It means by training different types of models, the resulting training data becomes less biased with compared to any single model.

3. A list of novelties/contributions.

In this paper, several innovative ideas are introduced. One of them is batch-mode active learning, which suggests strategies for selecting different and informative samples to be labeled together, thus improving efficiency compared to serial querying. The paper also presents methods for dealing with noisy or unreliable annotations, such as repeated labeling to reduce noise and estimating annotator quality. Furthermore, it proposes methods to consider variable labeling costs when selecting queries, combining both known and unknown costs into active learning. The paper also considers multi-task active learning, which optimizes the information level of queries for multiple learners when a single instance can be labeled for multiple tasks.

4. What do you think are the downsides of the work?

Many of the proposed solutions, such as batch-mode active learning and handling noisy oracles, add computational and algorithmic complexity.

Paper 3: *Active learning for data streams: a survey (2024)*

1. What problem does this paper try to solve, i.e., its motivation?

The paper addresses the challenge of efficiently selecting the most informative data points to label from a “data stream” in real time and real world application. With the increasing availability of streaming data in real-world applications, obtaining labeled data can be costly and time-consuming. Active learning in data streams aims to reduce labeling costs while improving model performance by querying the most valuable data points for labeling. Additionally, as data streams are going to be large, this paper tries to develop active learning methods that make decisions in real-time without storing or processing large amounts of data. Also, data streams often experience changes known as concept drift. The paper is motivated by the need for active learning strategies that can adapt to these changes while selecting the proper samples for labeling

2. How does it solve the problem?

This paper outlines strategies for making real-time decisions and several query strategies that are designed for data streams, which differ from traditional pool-based active learning methods like selecting samples that the model is least confident about, estimateing the expected error reduction from labeling a particular instance and queries those instances that are expected to improve future predictions the most, and using an ensemble of models to select samples where there is disagreement between model predictions.

3. A list of novelties/contributions.

Comprehensive survey of stream-based active learning strategies, bridging the gap between pool-based and stream-based active learning methods. Introduction of real-time querying strategies for stream-based learning, addressing specific challenges like concept drift and label latency. Diversity-based approaches for selecting data points that represent the overall distribution of the data. Handling concept drift by reviewing methods that adapt models in response to changes in the data distribution. Evaluation metrics specifically toward stream-based active learning algorithms, emphasize learning efficiency over time.

4. What do you think are the downsides of the work?

I think some of the proposed methods may not be efficient for most data streams due to the computational complexity of real-time decision-making. ALso, while the paper surveys a broad range of techniques, many methods have not been widely tested or validated in real-world settings, limiting their practical results. In the end, the assumption of immediate label availability may not always be maintained, and dealing with it can additionally complicate the learning process.