Abstract

In several industrial settings, the process and reliability experts hold the responsibility of generating precise ground truth labels for the generated dataset. These ground truth labels are then utilized for carrying out supervised learning tasks depending on the use cases. However, to achieve better performance, many labeled datasets have to be made available, for which domain experts spend a considerable amount of time. As a result, ground truth generation becomes complex, and sometimes it may be inaccurate depending on the complexity of the data. Therefore, the expert's accurate annotation of the dataset holds the key for many AI applications when thses tasks are carried out.

Active Learning (AL) can be used to address such problems. It is a powerful tool to handle modern machine learning problems with significantly fewer labeled training instances. An active learner starts with few labeled instances and builds a model by utilizing a proportion of unlabelled data. The building process is assisted with an oracle in the loop, which assists in determining the class for the queried instance. However, the implementation of traditional AL methodologies in practical scenarios is accompanied by multiple challenges due to the inherent assumptions. There are several hindrances, viz. unavailability of labels for the AL algorithm to start with; unreliable external source of labels during the querying process may provide inaccurate information for the ground truth,

or incompatible mechanism to evaluate the performance of Active Learner.

This work presents an AL framework that simultaneously addresses three practical challenges: cold-start problem, oracle uncertainty, and evaluating the performance of Active Learner in the absence of ground truth. While a pre-clustering approach is employed to address the cold-start problem, the uncertainty surrounding the expertise of the labeler and confidence in the given labels is incorporated to handle oracle uncertainty. The heuristics obtained during the querying process serve as the fundamental premise for accessing the performance of Active Learner. The robustness of the proposed AL framework is evaluated across different environments and industrial settings.