ADs: Active Data-sharing for Data Quality Assurance in Advanced Manufacturing Systems

Yue Zhao[1], Yuxuan Li[2], Chenang Liu[2], and Yinan Wang[1]

[1]Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute
[2]School of Industrial Engineering and Management, Oklahoma State University

1 Research Overview

My research focuses on engineering-driven machine learning for advanced manufacturing systems. This research overview will present one of my works on ensuring the quality of the shared data in advanced manufacturing systems.

2 introduction

Machine learning (ML) methods are widely used in manufacturing applications, which usually require a large amount of training data. However, data collection needs extensive time costs and investments in the manufacturing system, and data scarcity commonly exists. With the development of the industrial internet of things (IIoT), data-sharing is widely enabled among multiple machines with similar functionality to augment the dataset for building ML models. Although these machines may have similar functionality, the mismatch of distribution inevitably exists in their data due to different working conditions, process parameters, measurement noise, etc. However, the effective application of ML methods is built upon the assumption that the training and testing data are sampled from the same distribution. Thus, an intelligent data-sharing framework is urgently needed to ensure the quality of the shared data such that only beneficial information is shared to improve the performance of ML methods. There still exist two main challenges in tackling this problem: (1) the labeling information is usually unavailable to distinguish data points from different distributions; (2) it is not straightforward to evaluate the "quality of information" in each data point.

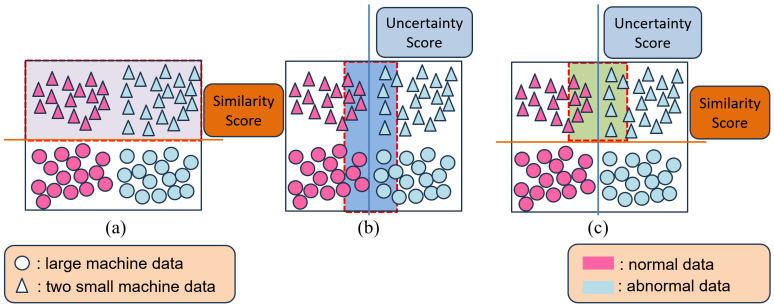

Figure 1: Visualizing the sample space by applying the ADS framework to the unlabeled data pool.

To tackle these two challenges, we proposed an Active Data-sharing (ADs) framework to ensure the quality of the shared data among multiple machines. It is designed as a self-supervised learning framework by integrating the architecture of contrastive learning and active learning. Contrastive learning is adapted to measure the similarity of semantic features among data points such that data from the same underlying distribution will share a high similarity score. A novel acquisition function is then developed for active learning by integrating the information measure and the similarity score. In our setting, we consider that machines S_1 and S_2 are similar, providing data that follows a similar distribution, while machine L_1 is different and the monitoring data from a different distribution. Both types of machines have normal or abnormal working conditions. The downstream task is to conduct anomaly detection for a specific type of machine. The objective is to identify the most informative subset of data points that (1) benefit anomaly detection and (2) from the target distribution, which are highlighted in the green area in Figure 1 (c).

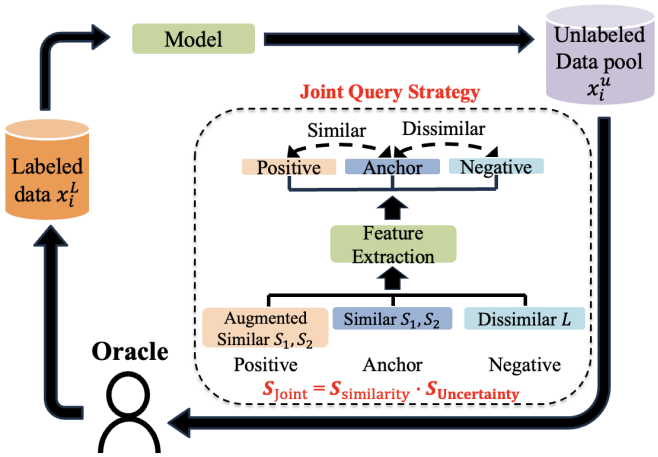

Figure 2: Architecture of active data sharing (ADs).

The architecture for our proposed ADs framework is shown in Figure 2. The steps are summarized as follows: (1) In the AL (active learning) phase, the classifier is trained using the initially labeled data. (2) Leveraging existing classifiers and pre-trained CL (contrastive learning), we extract the similarity feature and uncertainty feature for both labeled and unlabeled data. (3) The features are used to calculate the integrated score, which is a combination of the similarity score and uncertainty score, on the entire unlabeled dataset. (4) The data samples with the highest integrated scores are considered from the green highlighted region in Figure 1 (c) and selected for annotation. (5) The labeled dataset undergoes an update by incorporating the newly queried and labeled data. Signifying the inception of a cyclic process that iteratively repeats steps 1-5. Notably, the classifier model evolves with each cycle, updating with new labeled data from the oracle.

The major contributions of the work are: (1) A novel Active Data-sharing (ADs) framework is proposed to ensure the quality of industrial data-sharing when subject to data scarcity, distribution mismatch, and low annotation budget. (2) A novel acquisition function is developed for AL under distribution mismatch in the input space by integrating the informativeness and distribution similarity scores. (3) The effectiveness of the framework is evaluated on real-world *in-situ* monitoring data from additive manufacturing (AM) processes. The codes and dataset for this paper are available in this link https://github.com/zhaoy23/ADs. The full version of this work can refer to https://arxiv.org/abs/2404.00572.