Ensemble Active Learning by Contextual Bandits for AI Incubation in Manufacturing

Journal:	*Transactions on Intelligent Systems and Technology*
Manuscript ID	Draft
Manuscript Type:	Regular Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Zeng, Yingyan; Virginia Polytechnic Institute and State University, Grado Department of Industrial and Systems Engineering Chen, Xiaoyu; University of Louisville JB Speed School of Engineering, Department of Industrial Engineering Jin, Ran; Virginia Polytechnic Institute and State University, Grado Department of Industrial and Systems Engineering
Keyword:	Active learning, Reinforcement learning, Ensemble methods, AI Incubation, Contextual Bandits, Industrial Cyber-physical System, Online data annotation

Ensemble Active Learning by Contextual Bandits for AI Incubation in Manufacturing

YINGYAN ZENG, Grado Department of Industrial and Systems Engineering, Virginia Tech, USA
XIAOYU CHEN, Department of Industrial Engineering, University of Louisville, USA
RAN JIN*, Grado Department of Industrial and Systems Engineering, Virginia Tech, Virginia, USA

The online sensing techniques and computational resources in an Industrial Cyber-physical System (ICPS) provide a digital foundation for data-driven decision making by artificial intelligence (AI) models. However, the poor data quality (e.g., inconsistent distribution, imbalanced classes) of high-speed, large-volume data streams poses significant challenges to the online deployment of the offline trained AI models. As an alternative, updating AI models online based on streaming data enables continuous improvement and resilient modeling performance. However, for a supervised learning model (i.e., a base learner), it is labor-intensive to continuously annotate all streaming samples and it is also challenging to select a subset with good quality to update the model. Hence, a data acquisition method is needed to select the data for annotation from streaming data to ensure data quality while saving annotation efforts. In the literature, active learning methods have been proposed to acquire informative samples. Different acquisition criteria were developed for exploration of under-represented regions in the input variable space or exploitation of the well-represented regions for optimal estimation of base learners. However, it remains a challenge to balance the exploration-exploitation trade-off under different online annotation scenarios. On the other hand, an acquisition criterion learned by AI (e.g., by reinforcement learning) adapts itself to a scenario dynamically, but the ambiguous consideration of the trade-off limits its performance in frequently changing manufacturing contexts. To overcome these limitations, we propose an ensemble active learning method by contextual bandits (CbeAL). CbeAL incorporates a set of active learning agents (i.e., acquisition criteria) explicitly designed for exploration or exploitation by a weighted combination of their acquisition decisions. The weight of each agent will be dynamically adjusted based on the usefulness of its decisions to improve the performance of the base learner. With adaptive and explicit consideration of both objectives, CbeAL efficiently guides the data acquisition process through selecting informative samples to reduce the human annotation efforts. Furthermore, we characterize the exploration and exploitation capability of the proposed agents theoretically. The evaluation results in a numerical simulation study and a real case study demonstrate the effectiveness and efficiency of CbeAL in manufacturing process modeling of the ICPS.

CCS Concepts: • **Computing methodologies → Active learning settings**; *Ensemble methods*; *Reinforcement learning*.

Additional Key Words and Phrases: AI Incubation, Contextual Bandits, Industrial Cyber-physical System, Online Data Annotation

Authors' addresses: Yingyan Zeng, yingyanzeng@vt.edu, Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia, USA, 24061; Xiaoyu Chen, xiaoyu.chen@louisville.edu, Department of Industrial Engineering, University of Louisville, Kentucky, Louisville, USA, 40292; Ran Jin, jran5@vt.edu, Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia, USA, 24061.

2157-6904/2022/3-ART $15.00
https://doi.org/10.1145/1122445.1122456

ACM Reference Format:
Yingyan Zeng, Xiaoyu Chen, and Ran Jin. 2022. Ensemble Active Learning by Contextual Bandits for AI Incubation in Manufacturing. *ACM Trans. Intell. Syst. Technol.* 1, 1 (March 2022), 28 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Industrial Cyber-physical Systems (ICPSs) integrate the cyber and physical worlds, which serve as the backbone of the Fourth Industrial Revolution [17]. By embracing the Internet of Things (IoT), an ICPS interconnects manufacturing equipment with ubiquitous sensors, actuators and computing units, forming a low-cost, high-availability and high-accessibility network [74]. The large-volume and high-speed sensing data collected from such a network have advanced many data-driven decision-making methods to support the manufacturing efficiency and quality improvement and cost reduction. For example, artificial intelligence (AI) models such as support vector machine (SVM) and deep neural networks for supervised learning have been employed for quality modeling and process monitoring of fused deposition modeling (FDM) processes [24, 67], Aerosol® Jet Printing processes [59], etc. However, most AI models are proposed following an offline-training-online-deployment (OTOD) strategy, which is only considered to be effective when the quality of the training data set is guaranteed (e.g., training data can provide adequate estimations of the underlying true model of the variable relationships in supervised learning). In practice, various factors can change such underlying model in manufacturing processes (e.g., due to the degradation of manufacturing equipment or the change of product design), which results in an erratic performance of AI models during their online deployment.

To improve the modeling performance, one can either build a dynamic model [35] or to investigate an online model training mechanism to adapt existing models to online data streams. However, constructing a dynamic model highly depends on the prior knowledge of the distribution changing patterns and root causes. Instead of focusing on creating a better model, data-centric AI has been proposed as a more general approach to engineer the data needed to successfully build an AI model [57]. From a holistic view, we envision a resilient AI system to identify and mitigate the performance fluctuation of AI models caused by abrupt changes (i.e., data distribution, learning algorithms, and computational resources), which jointly consider managing the data quality as well as adapting the existing models during the online deployment. Therefore, in this paper we focus on online model training by actively acquiring samples to ensure data quality such that the resilient AI performance can be achieved.

Here we focus on the supervised learning model as the base learner, where the data quality is considered from the aspect of representativeness (i.e., consistent distribution between the training and testing data sets) and class imbalance [26]. In the offline-training step, the high-quality data set can be defined when there are sufficient representative samples for training, such as sufficient samples for multimodal distributions [28], a distribution of the training samples close to that of the testing samples, and when there is balanced class distributions [9, 62]. However, OTOD cannot support the AI modeling since in the context of high-speed, large-volume streaming data, the data quality of training data sets needs to be evaluated continuously. As a motivation example, Fig. 1 demonstrates the multimodal distribution and the imbalanced class of samples collected from a highly personalized FDM process, where the samples are projected to the principle component directions of the input variable space by Principle Component Analysis (PCA). The two clusters are generated from two layers in FDM due to different product geometric designs. The objective of the data analysis in Fig. 1 is to use *in situ* process variables to predict the layer-to-layer binary quality variable, which indicates the surface roughness as a classification problem. Assume that the samples collected until time t will be used to train the model, where the samples from only one

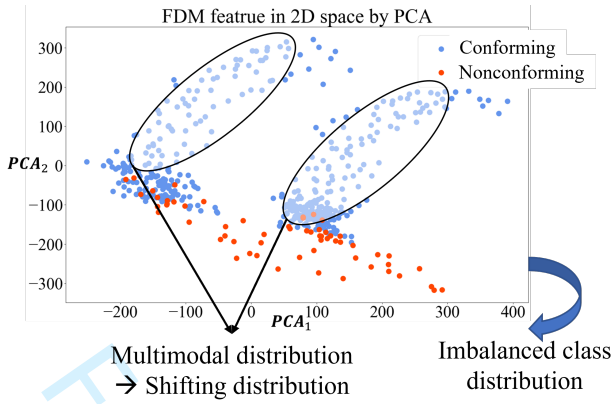


Fig. 1. Distribution of FDM input data with reduced dimensions by Principle Component Analysis (PCA)

cluster have been observed. After time $t + 1$, the streaming samples are from the other cluster, i.e., the FDM goes to the next layer with another design in that slice. The collected training data set is not representative due to the shift of distributions, thus resulting in a sudden decrease of the prediction performance of the pre-trained model.

Motivated by Fig. 1, it is important to actively select the streaming samples for annotation which ensures data quality. In earlier studies, Design of Experiments (DoE) was proposed to improve the supervised learning models by generating samples to identify significant variables [21, 34]. Recent efforts in data filtering, either model-based [42] or model-free methods [43, 63] aim at accurately modeling the underlying system by sampling a subset with good representativeness of the population distribution. However, DoE focus on actively generating the data while data filtering methods require a completely collected data set before the selection. Neither method can be directly applied to acquire the streaming data in the ICPS.

To obtain high-quality data, humans also play an indispensable role in the data annotation with their domain knowledge. In particular, the online data annotation requires real-time experimentation and human-machine interface for domain experts to interact with [64], which is costly, time-consuming, and labor-intensive. While automatic annotation methods employ semi-supervised learning methods to annotate the samples by the most confident predictions [38], one potential disadvantage is that the mislabelled cases may impair the performance of AI models [11]. Recognizing the significance of human-in-the-loop, an AI incubation framework [16] was proposed to allow domain experts to interact with an AI agent for the training, validation, and deployment of supervised learning models, which is considered as a critical part for creating the resilient AI methods. Under this framework, an Interpretable Neural Network (INN) has been proposed for human decision rule-based model [16]; and a Visual Language Processing (VLP) modeling [14] has been proposed to generate features for AI models based on human visual searching patterns. In this work, we focus on the training data selection and annotation as a part of the AI incubation framework, aiming at reducing human efforts while efficiently improving the learning performance of the base learner. Without loss of generality, we focus on classification models as base learners. Note that the quality of human annotation is out of the scope of this research and will be considered in the future work.

To create an online data acquisition method, an acquisition criterion needs to be designed to determine whether a sample should be selected for human annotation in order to acquire only informative samples and to provide high-quality training data sets for the base learner. This

acquisition decision can be viewed as a dilemma between the exploration and exploitation of the input variable space [8]. Here, exploitation is defined as acquiring a sample around the conceptual boundary for boundary learning, whereas exploration is defined as acquiring a sample located in the under-represented region for the input variable space discovery. Exploitation-oriented criteria work well when the base learner can easily detect the important regions [46]. Otherwise, exploration is required for more complex scenarios such as the exclusive XOR problem. For the scenario of motivation example shown in Fig. 1, if we concentrate exclusively on the samples near the decision boundary for exploitation, samples from another cluster may be overlooked throughout the streaming process, leading to an inaccurate estimate of the decision boundary and poor performance in under-represented regions. On the other hand, the exclusive exploration will easily lead to a base learner with high uncertainty. Therefore, a well-balanced exploration-exploitation trade-off is essential for guiding the online annotation of AI models with complex distribution in the input variable space [44].

In this paper, we propose an ensemble active learning method by contextual bandits (CbeAL) to improve the exploration and exploitation trade-off under various scenarios, thus guiding an efficient and effective human annotation process. In CbeAL, a set of active learning agents with human-designed criteria is incorporated by contextual bandits [7], where a joint acquisition decision is made by the weighted combination of individual decisions. Here, we use "agents" as an umbrella term to refer the acquisition criteria in active learning methods. The incorporated candidate active learning agents are designed to pursue an explicit objective of exploration or exploitation with theoretical justification, respectively. Thus, during the annotation process, the weight (i.e., decision power) of each agent indicates the current tendency for exploration or exploitation, which will be updated dynamically by the bandits solver subject to the historical reward. To improve the learning performance of the base learner, the reward is defined as the usefulness of the acquiring behaviour of CbeAL where acquiring a sample which would be wrongly predicted by the base learner is considered useful. Hence, the online data acquisition problem can be effectively addressed by CbeAL, which pursues the exploration and exploitation trade-off via the ensemble of a set of active learning agents. In this respect, CbeAL is a generic active learning framework which reduces the manual adjustment of active learning agents under frequently changed manufacturing data distributions. Apart from improving the learning performance, CbeAL also increases the interpretability of AI models, not from the modeling perspective [54, 65] but from the data perspective [75], since the weight of each agent in each acquisition step explains whether the sample is annotated for exploration or exploitation.

The remainder of this paper is organized as follows. Section 2 summarizes the related work. Section 3 introduces the proposed CbeAL method and provides the theoretical justification. Section 4 evaluates the performance of the proposed CbeAL by simulation studies. Section 5 validates CbeAL via a real case study of the online quality modeling of FDM in the ICPS. We conclude this work with some discussions of the future work in Section 6.

2 RELATED WORK

2.1 Online Model Updating in Industrial Cyber-physical Systems

In the past decades, the ICPS integrates the physical manufacturing equipment with sensing and actuation networks as well as ubiquitous computational resources, which provides the digital foundation for the online updating of AI models [49, 69]. With the streaming observational data and online computational resources, the online updating techniques of AI models have been investigated to enable the close modeling of manufacturing processes and facilitates the efficient decision-making in ICPSs. For example, Bastani et al. [6] proposed an online classification model

for real-time monitoring in additive and semiconducting manufacturing processes. Wang et al. [66] developed a large-scale online multitask learning model to coordinate machine actions in the ICPS online. Besides, Zhang et al. [74] proposed and deployed the distributed family learning algorithm in the computation network to support the online joint modeling of similar-but-non-identical products in SLM. Online model updating strategies have also been developed for model calibration and predictive maintenance [41, 72]. However, the aforementioned studies focus on developing the online updating algorithm of the AI model via Bayesian methods or distributed optimization methods to reduce the computational burden with large-volume streaming data, which are effective for unsupervised learning problems or supervised learning scenarios with easily collected responses. Yet for many supervised learning scenarios in the ICPS, the passively collected data need to be annotated via real-time experimentation by domain experts, e.g., the inspection of a batch of 400 wafers may take more than 8 hours [33]. The lack of consideration of human annotation efforts renders these online updating methods inefficient for supervised AI models when the annotation process is labor-intensive and time-consuming, especially in highly personalized manufacturing environments with rapid product and process changes [2].

2.2 Data Quality and Data Acquisition Methods

Compared to the accuracy and efficiency of learning algorithms, validating and monitoring the quality of data fed to AI models is an equally important problem [12]. Metrics for assessing the data quality for classification tasks include outlier detection, boundary complexity, label noise, shifting distribution, class imbalance, etc [26]. In the context of streaming data, Caveness et al. [12] developed a data analysis and validation system to monitor the significant changes between successive batches of the training data by summary statistics (i.e., mean, variance, etc.) with human investigation for a machine learning pipeline. However, without considering the informativeness of the data pertaining to the AI model, the data collected by such a system cannot improve the modeling performance effectively.

On the other hand, to improve the performance of supervised learning models and to reduce the labor efforts of human annotation, methods have been developed to facilitate effective data acquisition for informative and high-quality data. These methods include providing acquisition recommendations for human annotation (e.g., sequential design and active learning) and automatic annotation (e.g., semi-supervised learning) [18], where limited approach suits the online streaming data. Sequential design focuses on selecting the samples in a sequential manner to achieve certain optimality criteria such as maximum entropy, maxmin distance [39, 56]. For example, Yan et al. [73] proposed an adaptive sequential sampling method to balance the sampling efforts between the exploration and exploitation of anomalous regions for anomaly detection in the ICPS. However, these methods provide active recommendations that require experiments to be conducted at selected points in the input variable space, which cannot address the passively collected data. Semi-supervised learning has been employed to automate the annotating process such that the base learner can learn from both labelled and unlabelled data. [76]. However, adding mislabelled cases by a semi-supervised learner to the training set may hamper the base learner's learning performance [11]. Due to the aforementioned limitations, we focus on active learning methods which provide acquisition recommendations for the passively collected data. Active learning has been leveraged for minimizing the human effort as well as improving the modeling performance in various applications including human activity recognition [1], threatening surveillance event detection [45], wearable sensing platforms annotation [55], etc.

2.3 Exploration and Exploitation in Active Learning

Active learning reduces the annotation efforts for supervised learning models by evaluating the informativeness of samples and acquiring the most informative ones [51]. The decision on whether one sample should be labelled can be viewed as a dilemma between the exploration and exploitation of the input variable space. In the earlier work, most efforts exploit samples with large amount of information about the base learner as an exploitation-oriented strategy. Metrics such as classification uncertainty[40], margin[4], and entropy [22] of the base learner have been adopted to measure the informativeness and compared with corresponding thresholds to make the acquisition decision. To achieve exploration for online streaming data, Ienco et al. [32] modeled the local density of an sample to acquire the sample lying in a dense region with a small classification margin. For trivial scenarios where only parts of the input variable space have to be known in order to perform optimally, exploitation-oriented acquisition criteria can be more effective to avoid exploring regions that are irrelevant for the decision boundary estimation, especially for a high-dimensional input variable space [50, 60]. However, in nontrivial scenarios, exploration plays a more important role to explore the relevant but unknown regions if the base learner has not made accurate estimation on the decision boundary. Thus, the exploration-exploitation trade-off becomes essential for the online model updating under scenarios with exclusive XOR problem, clusterwise structure, imbalanced class distribution, etc [44, 46]. Neither exploration or exploitation can achieve promising learning performance for classification models under varying online annotation scenarios, due to the lack of compatibility or the suboptimality in specific scenarios.

To achieve a compromise, a common acquisition strategy is to conduct exploration and exploitation simultaneously. Considering two acquisition criteria which are dedicated to exploration (e.g., random sampling) and exploitation (e.g., uncertainty sampling) respectively, the compromise can be achieved by selecting one criterion with a certain probability for each streaming sample. One typical example is the ϵ-greedy policy which enforces the input variable space exploration with probability ϵ in each round [60]. Representative sampling methods have also been designed as the combination of exploitation-oriented and exploration-oriented criteria [23, 70]. However, the ambiguity of the weight of each objective requires further fine-tuning for each learning scenario. To avoid ambiguity, Loy et al. [44] extended the Query-by-Committee (QBC) [52] paradigm to a nonparametric Bayesian model to address unknown class discovery and imbalanced class distribution for the online annotation. However, without taking into account the informativeness (i.e., uncertainty) of a sample about the base learner, the proposed QBC-PYP cannot adjust the exploration-exploitation to the learning performance of the base learner. In brief, the simple combination of active learning criterion cannot handle challenging online data acquisition scenarios even with fine-tuning, which is due to the lack of compatibility with the data stream and the lack of adaptiveness to the learning performance of the base learner [20]. Therefore, an ensemble of multiple criteria is desired to guide the dynamic exploration-exploitation trade-off in an adaptive and data-dependent manner.

As a promising approach, active learning has been recently formulated in the framework of reinforcement learning (RL) and multi-armed bandits where the objective is to learn the optimal acquisition criterion as a policy to maximize the cumulative reward [19]. However, these methods can be also lack of systematic and explicit considerations for both exploration and exploitation objectives. Wassermann et al. [71] proposed Reinforced Active Learning (RAL) which modeled the stream-based active learning as a contextual bandits problem. In RAL, a set of base learners were gathered as the committee to provide acquisition advice based on the certainty degree of the sample to each learner. The acquisition criterion can be viewed as the weighted combination of different uncertainty sampling policies. In spite of its adaptiveness to the data stream, RAL is highly exploitation-oriented since it mainly focuses on decision boundary learning for each

learner. Baram et al. [5] first proposed COMB to blend multiple acquisition criteria as experts and consider samples as arms in multi-armed bandits. Later, Hsu and Lin [29] developed ALBL also with the bandits analogy but refined COMB by taking candidate acquisition criteria as experts (i.e. candidate policies). The bandits framework enables these approaches to adjust the weight of candidate criteria to the exploration and exploitation status of the active learning process, thus dynamically controlling the trade-off. However, neither COMB or ALBL answered the question of how to select candidate acquisition criteria, nor considered the explicit objective of exploration and exploitation. A random selection of general acquisition criteria with a small size may not well address different online annotation scenarios, while a large size of experts may cause problematic performance of the bandits solver.

3 METHODOLOGY

To develop the active learning agents for CbeAL and derive the theoretical characterization of the agents, we make the following assumptions: (i) The sample size of the initial training set $\mathcal{D}_0$ is not large enough to guarantee satisfactory modeling performance and the samples in $\mathcal{D}_0$ are not uniformly distributed in the input variable space. (ii) The streaming data have highly imbalanced class distribution. (iii) There are multiple clusters in the input data distribution. One common example is that the input data follow a Gaussian mixture distribution. This assumption is validated by the simulation setup and validated in the case study. Note that the proposed CbeAL framework is designed for general online annotation scenarios and does not require the assumptions on the input data distribution.

3.1 Overview of the Proposed Methodology

Consider the online data annotation scenario with a sample $\boldsymbol{x}_t$ collected at time $t, t = 1, 2, ..., T$, where $\boldsymbol{x}_t \in \mathbb{R}^p$ is the input for the base learner (i.e.,the classification model) f_t. We assume that the classification problem has c classes, and $y_t \in C = \{1, 2, \ldots, c\}$ is the label of the sample $\boldsymbol{x}_t$, which can only be obtained by human annotation if an acquisition decision is made to acquire $\boldsymbol{x}_t$. Denote the labelled data pool at time t as $\mathcal{D}_t = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_{n_t}, y_{n_t})\}$ with $|\mathcal{D}_t| = n_t$. The base learner f_0 is pretrained by an initial $\mathcal{D}_0$, which contains a limited number of annotated samples. We have a budget of B samples for annotation during the streaming process.

As an overview (Fig. 2), our key idea is to ensemble the acquisition decisions made by the exploration- and exploitation-oriented agents and adaptively balance the two aspects based on the context of incoming samples, the learning performance of the base learner, and the historical performance of the agents. During the online annotation process, at each time point t, (i) we receive a sample $\boldsymbol{x}_t$. (ii) Afterwards, one can obtain the predicted label $\hat{y}_t$ and the side information, such as the predicted probability $\boldsymbol{P}^f(\hat{y}_t|\mathcal{D}_t)$ of each class from f_t and take $\boldsymbol{x}_t$ and other information as the context input for the proposed contextual bandits solver Exp4.P-EWMA. And (iii) we make the acquisition decision as a weighted majority of the decisions obtained from the candidate agent set $\{AG1, AG2, ...\}$. If the decision is to acquire the sample, we acquire y_t from human annotation and obtain the reward r_t, update CbeAL with r_t, and retrain the classifier with $(\boldsymbol{x}_t, y_t)$; otherwise, we pass this sample without annotation. The advantage of the proposed framework lies in three aspects: (i) CbeAL explicitly pursues the input variable space discovery and the decision boundary learning via incorporating exploration- and exploitation-oriented agents while it balances the overall exploration-exploitation trade-off adaptive to the data stream and the learning performance of the base learner by contextual bandits. (ii) The systematic ensemble of multiple pairs of agents save the efforts for agents selection under various learning scenarios with different input data distribution, feature dimension, signal to noise ratio, etc. Therefore, CbeAL is a developed as a generic active learning framework to achieve an adaptive and well-balanced exploration-exploitation trade-off

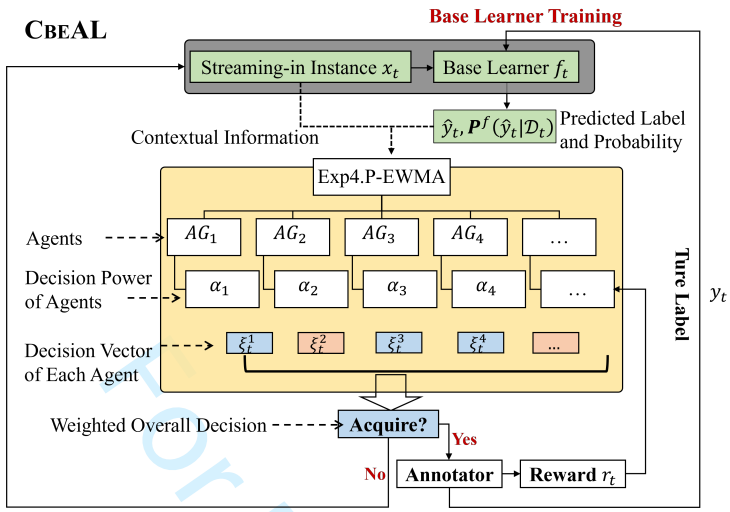


Fig. 2. Overview of the proposed CBEAL framework

for the incubation of classification models with streaming data. (iii) CBEAL is scalable in terms of the number of active learning agents to enhance the exploration or exploitation.

3.2 The Ensemble Active Learning by Contextual Bandits

During the online annotation process, the needs for exploration and exploitation is changing over time, which depends on the observed samples, the incoming sample, and the updated learning performance of the base learner. Since there does not exist a consistent optimization criteria to adjust the trade-off, the shift between the two aspects is nontrivial.

To address the challenge of achieving a good exploration-exploitation trade-off adaptive to various online annotation scenarios, we formulate the shift between two objectives as a contextual multi-armed bandits problem. In the bandits problem, the bandits solver needs to make the decision of pulling one of the K arms as the action for time point t based on the received contextual information. With each arm characterized by an unknown reward distribution, the objective of the solver is to gain the highest cumulative reward $R = \sum_{t=1}^{\infty} r_t$. Under the framework of CBEAL, we consider the decision to acquire or not acquire a sample as the arms and propose to ensemble exploration- and exploitation-oriented agents as candidate policies (i.e., experts). The acquisition decision (i.e., decision of pulling one arm) is jointly made by a weighted combination of the individual decisions from each agent. Thus, the adjustment of exploration-exploitation trade-off is converted to the selection among different types of agents with the goal of gaining a higher reward, which can be solved by some well-developed bandits solvers.

In CBEAL, the arm pulled at time t (i.e., $a_t \in A = \{1, 2\}, |A| = K = 2$) represents the overall acquisition decision, where $a_t = 1$ refers to acquiring the sample $\boldsymbol{x}_t$ and otherwise $a_t = 2$. At each time point t, the incoming sample $\boldsymbol{x}_t$, prediction $\hat{y}_t$, and the predicted probability $\boldsymbol{P}^f(\hat{y}|\boldsymbol{x}_t) \in \mathbb{R}^c$ obtained by the base learner f_t are considered as the observed contextual information that affects the exploration-exploitation trade-off. Each incorporated active learning agent ($AG_i, i \in \{1, ..., N\}$) makes its own decision $\boldsymbol{\xi}_t^i \in \mathbb{R}^K$ based on these contextual information $\{\boldsymbol{x}_t, \hat{y}_t, \boldsymbol{P}^f(\hat{y}|\boldsymbol{x}_t)\}$. Here $\xi_{a,t}^i$ represents the probability of the i-th agent taking the a-th action. Specifically, we have the decision vector $\boldsymbol{\xi}_t^i = [p_t^i, 1 - p_t^i]$, where p_t^i is the acquisition probability for sample $\boldsymbol{x}_t$. Simultaneously, each agent is assigned with a decision power $\alpha_{i,t}$ at time t. The overall decision is a majority voting of

all agents' decisions weighted by their decision power, which leads to a decision vector $\boldsymbol{P}_t \in \mathbb{R}^K$. If the overall decision asks for the ground-truth label, a reward r_t is received after the execution. Afterwards, the proposed bandits solver Exp4.P-EWMA updates the decision power of each agent based on its decision $\boldsymbol{\xi}_t^i$ in this iteration and the reward r_t. With the objective of gaining a high cumulative reward, the ensemble of agents is the same as to combine the decisions made by each agent such that the reward gained in each iteration is close to the highest we can get from the best agent in the agent set. The execution of CbeAL is summarized as Algorithm 1.

Algorithm 1: CbeAL

Input: set of agents $\{AG_1, \ldots, AG_N\}$, $\mathcal{D}_0, f_0, B$
Initialize: $t = 0$, $budget\_used = 0$
while $budget\_used < B$ **do**
  Receive sample $\boldsymbol{x}_t$ and contextual information $\{\boldsymbol{x}_t, \hat{y}_t, \boldsymbol{P}^f(\hat{y}|\boldsymbol{x}_t)\}$
  Obtain the decision vector $\boldsymbol{\xi}_t^1, \ldots, \boldsymbol{\xi}_t^N$ from agents
  Execute Exp4.P-EWMA solver for one iteration and obtain the action a_t
  if $a_t = 1$ **then**
    Acquire y_t, $\mathcal{D}_{t+1} = \mathcal{D}_t \cup (\boldsymbol{x}_t, y_t)$
    Train the base learner $f_{t+1} \leftarrow D_{t+1}$
    $budget\_used = budget\_used + 1$
  else
    $D_{t+1} = D_t, f_{t+1} = f_t$
  Update $\{AG_1, \ldots, AG_N\}$
  $t = t + 1$
Output: $\mathcal{D}_{t+1}, f_{t+1}$

Notably, the online updating of the base learner is not the focus of this study and we simply retrain the base learner based on all annotated samples from $\mathcal{D}_{t+1}$. For a more efficient updating, online learning algorithms, such as first-order algorithms [77] and Bayesian-based approaches [13], can be adopted depending on the base learner.

To integrate the goals of active learning and multi-armed bandits in order to provide informative acquisition, the design and characterization of the reward are critical. We define the reward r_t as suggested in [71]:

$$r_t = \begin{cases} \rho^+, & \text{if } \hat{y}_t \neq y_t \\ \rho^-, & \text{if } \hat{y}_t = y_t. \end{cases} \tag{3.1}$$

The reward r_t can only be obtained if the overall decision made by CbeAL acquires the ground-truth label. Otherwise, it is zero. Intuitively, the acquisition action will be rewarded if the base learner would have made a wrong prediction, otherwise it will be penalized since this acquisition is considered unnecessary. Therefore, it measures both the informativeness and usefulness of an acquisition decision. Based on this design, the reward of acquiring a sample is determined by the performance of the current base learner f_t, the incoming sample $(\boldsymbol{x}_t, y_t)$, and also the performance of the bandits learner CbeAL. Thus, the sequence of reward $\{r_1, r_2, ..., r_t\}$ is autocorrelated. This characteristic is another reason that we adopt the setting of adversarial bandits [3], where one active learning agent (AG_i) is considered as one expert and decisions from each expert are simultaneously considered to make a joint decision, since no statistical assumption is made on the reward generation in this setting. Additionally, instead of considering one agent as one arm, incorporating agents

as experts also makes the number of agents scalable. In summary, with the designed setting of contextual information and the reward, the updated decision power of each agent adjusts the exploration-exploitation trade-off to improve the learning performance of the base learner.

To solve the formulated contextual bandits problem in CbeAL, we propose the Exp4.P-EWMA solver, where we embed a control chart-based flipping mechanism to Exp4.P[7]. To balance the overall exploration-exploitation behaviour, the exploration- and exploitation-oriented agents are incorporated in CbeAL by pairs, which can be easily adjusted for a specific online annotation scenario. However, with the pair ensemble, the direct application of Exp4.P can easily lead to a dominant agent (i.e., an agent consistently has the highest decision power) from an early stage, which makes CbeAL act no difference from a single active learning agent dedicated to exploration or exploitation. This can be expected since the pure exploration strategy in the early stage may cause the acquisition of samples with low uncertainty, so that the decision power of exploration-oriented agents keeps decreasing until a level too low to contribute to the overall acquisition decision any longer. To avoid the early convergence in Exp4.P, a control chart based flipping mechanism is integrated into the solver. Denote the standardized weight (i.e., standardized decision power) of the i-th agent at time t as $\alpha^s_{i,t}$, where $\alpha^s_{i,t} = \frac{\alpha_{i,t}}{\sum_{i=1}^{N} \alpha_{i,t}}$. We monitor each standardized weight by an EWMA chart [31], which detects drifting of weights over time. The intuition is that if the decision power of one agent keeps decreasing or increasing from the beginning, the decision power of all pairs of agents will be flipped so that the agents with lower power have more chances to lead the decision in the following period. Note this forced-exploration phase will only happen in a short period during the whole process, which is controlled by a hyperparameter γ. Denote the weighting factor for EWMA as λ, the size factor of shift to detect as h, and the estimated variance of $\alpha^s_{i,t}$ as $s^2_{i,t}$. With the flipping mechanism, the proposed Exp4.P-EWMA solver is detailed as follows:

As listed in Algorithm 2, at each time point t, the solver will first execute Exp4.P to make the acquisition decision a_t and update the decision power of each agent $\alpha_{i,t+1}$ based on its decision vector $\boldsymbol{\xi}^i_t$, the final decision probability $\boldsymbol{P}_t$, and the reward r_t. In the second step, the flipping will be triggered if the standardized weight of any agent is out of the updated control limits.

3.3 The Exploration and Exploitation Agents

With the proposed ensemble framework, active learning agents with existing acquisition criteria can be incorporated as experts to perform the online acquisition task. To save the tuning and balance the exploration-exploitation trade-off under different online annotation scenarios, we design distinguished active learning agents with exploration or exploitation objectives to be incorporated into CbeAL so that a systematic approach is developed without ambiguous selection. Another advantage of this design is the tendency for exploration or exploitation can be directly implied by the decision power of different types of agents.

3.3.1 Low-density Based Exploration Agent (LD-Agent). The objective of exploration is to identify the structure of the input data distribution during the learning process. Two types of agents are proposed to encourage the exploration of the input variable space. The first type adopts a density-based criterion. With multiple clusters, it is important to encourage the labelling efforts around each cluster boundary to discover new clusters. A sample is more likely to be located around the boundary of a cluster if it lies in a sparse region. Therefore, we propose to acquire the samples lying in an region with low density and we adopt the idea in [32] to model the density of a sample.

Denote set $\mathcal{W}$ as a sliding window of L previously observed samples, $d(\cdot,\cdot)$ as the distance between two samples. Denote $MaxDist$ as an function, where $MaxDist(\boldsymbol{x}_i, \mathcal{W})$ returns the maximum distance between $\boldsymbol{x}_i$ and other samples in the sliding window $\mathcal{W}$. To approximate the local density for a new coming sample, we define local sparsity (i.e., low-density) factor of a sample $\boldsymbol{x_i}$ as the

Algorithm 2: EXP4.P-EWMA SOLVER

Parameters: $\delta, \gamma, h, p_{\min} \in [0, 1/K]$
Initialization: Set $\alpha_{i,1} = 1, \alpha_{i,1}^{EWMA} = \frac{1}{N}$ for $i = 1, \ldots, N, \mu = \frac{1}{N}$.
for *t=1,2,...* **do**
  Input: decision vector of each agent $\boldsymbol{\xi}_t^1, \ldots, \boldsymbol{\xi}_t^N$
  Step 1: EXP4.P
    For $a = 1, \ldots, K$ get the final decision probability $\boldsymbol{P}_t$:
    $P_{t,a} = (1 - Kp_{\min}) \sum_{i=1}^{N} \frac{\alpha_{i,t}\xi_{a,t}^i}{\sum_{i=1}^{N}\alpha_{i,t}} + p_{\min}$
    Draw the action a_t based on $\boldsymbol{P}_t$ and receive the reward r_t
    for $a = 1, \ldots, K$ set $\hat{q}_{a,t} = \begin{cases} r_t/P_{t,a}, & \text{if } a = a_t \\ 0, & \text{otherwise} \end{cases}, \hat{\boldsymbol{q}}_t = [\hat{q}_{1,t}, ..., \hat{q}_{K,t}] \in \mathbb{R}^K$
    for $i = 1, \ldots, N$ set $\hat{g}_{i,t} = \boldsymbol{\xi}_t^i \cdot \hat{\boldsymbol{q}}_t^T, \hat{v}_{i,t} = \sum_a \xi_j^i / P_{t,a}$
    Update the decision power: $\alpha_{i,t+1} = \alpha_{i,t} \cdot \exp\left(\frac{p_{\min}}{2}(\hat{g}_{i,t} + \hat{v}_{i,t}\sqrt{\frac{\ln N}{KT}})\right)$
  Step 2: EWMA-based flipping mechanism
    for *i=1,...,N* **do**
      $\alpha_{i,t+1}^s = \frac{\alpha_{i,t+1}}{\sum_{i=1}^{N}\alpha_{i,t+1}}$
      $\alpha_{i,t+1}^{ewma} = \lambda \cdot \alpha_{i,t+1}^s + (1-\lambda)\alpha_{i,t}^{ewma}$
      $LCL = \mu - h \cdot \frac{\lambda}{2-\lambda} \cdot s_{i,t}^2, UCL = \mu + h \cdot \frac{\lambda}{2-\lambda} \cdot s_{i,t}^2$
      Set $\alpha_{i,t+1}^s = \begin{cases} 2\mu - \alpha_{i,t+1}^s, & \alpha_{i,t+1}^{ewma} > UCL \text{ or } \alpha_{i,t+1}^{ewma} < LCL \\ \alpha_{i,t+1}^s, & \text{otherwise} \end{cases},$
      $\alpha_{i,t+1} = (\sum_{i=1}^{N}\alpha_{i,t+1}) \cdot \alpha_{i,t+1}^s$
    $h := h \cdot \exp\gamma$
  Output: a_t

number of times $\boldsymbol{x}_i$ is the furthest away from other samples in $\mathcal{W}$ as follows:

$$lsf(\boldsymbol{x}_i) = \sum_{\boldsymbol{x}_j \in \mathcal{W}} \mathbb{I}\{MaxDist(\boldsymbol{x}_j) < d(\boldsymbol{x}_i, \boldsymbol{x}_j)\}. \tag{3.2}$$

Algorithm 3 provides the pseudocode to acquire samples with lower local density. Given a streaming sample $\boldsymbol{x}_t$ at time t, low-density based exploration agent first calculates the local sparsity factor $lsf(\boldsymbol{x}_t)$ to determine the acquisition probability p_t as the output. Then, the sliding window $\mathcal{W}$ and the maximum pairwise distance between each sample in $\mathcal{W}$ will be updated. The sliding window mechanism is adopted to adjust the approximated density based on the most recent data stream. Note the window length L, and the sparsity fraction δ_L are hyperparameters that affects the acquisition probability, which can be tuned to best suit the scenario.

3.3.2 Space-filling Based Exploration Agent (SPF-Agent). The second type of exploration-oriented agents is based on a space-filling criterion. In the DoE literature, space-filling designs are applied to fully explore the response surface of computer experiments [53]. Therefore, as an alternative strategy to explore the input variable space, a space-filling based exploration-oriented agent is developed to acquire samples uniformly distributed in the space. We adopt the idea of minimum pairwise distance criterion in the sequential space-filling design literature [36] and propose a

Algorithm 3: Low-density Based Exploration Agent (LD-Agent)

Input: $\boldsymbol{x}_t, \mathcal{W}, L, \mathcal{D}_0, \delta_L$
Calculate $lsf(\boldsymbol{x}_t)$
for j=1,2,..., L **do**
  if $d(\boldsymbol{x}_i, \boldsymbol{x}_j) > MaxDist(\boldsymbol{x}_j)$ **then**
    $MaxDist(\boldsymbol{x}_j) = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$
Output: Acquisition probability $p_t = \frac{lsf(\boldsymbol{x}_t)}{L\delta_L}$
if $|\mathcal{W}| > L$ **then**
  $\mathcal{W} := \mathcal{W} \setminus \boldsymbol{x}_{t-L}$
$\mathcal{W} := \mathcal{W} \cup \boldsymbol{x}_t$

corresponding criterion to minimize the pairwise distance between acquired samples during the online data acquisition.

Similarly, a sliding window $\mathcal{W}$ keeps the most recent L samples. Denote $MinDist$ as a function where $MinDist(\boldsymbol{x}_i, \mathcal{W})$ returns the minimum distance between $\boldsymbol{x}_i$ and all samples in $\mathcal{W}$.

Algorithm 4: Space-filling Based Exploration Agent (SPF-Agent)

Input: $\boldsymbol{x}_t, \mathcal{W}, L, \mathcal{D}_0$
for i=1,2,...,L **do**
  $\min(d_i) = \min d(\boldsymbol{x}_i, \boldsymbol{x}_j), \forall j \in \mathcal{W}$
Calculate $MinDist(\boldsymbol{x}_t, \mathcal{W})$
Output: Acquisition probability $p_t = \frac{MinDist(\boldsymbol{x}_t, \mathcal{W})}{\max_{i \in \mathcal{W}} \min(d_i)}$
if $|\mathcal{W}| > L$ **then**
  $\mathcal{W} := \mathcal{W} \setminus \boldsymbol{x}_{t-L}$
$\mathcal{W} := \mathcal{W} \cup \boldsymbol{x}_t$

In Algorithm 4, with a coming sample $\boldsymbol{x}_t$ at time t, its minimum distance from the samples in $\mathcal{W}$ is compared with the largest minimum pairwise distance of samples in $\mathcal{W}$ to obtain the acquisition probability, leaving a higher probability for samples distant from the observed ones in $\mathcal{W}$.

Intuitively, density-based criterion will explore the boundary of the input variable space faster at an early stage whereas space-filling criterion allows for a more uniform exploration during the process. The combination of two exploration criteria will enhance the compatibility and adaptiveness of CbeAL to various learning scenarios. In practical, an ϵ-greedy policy can also be embedded into CbeAL which forces a sample to be acquired with probability ϵ to further encourage the exploration.

3.3.3 Reinforced Exploitation Agent (RAL-Agent). The goal of exploitation in active learning is to capture the decision boundary, which is generally accomplished through acquiring samples with ambiguous class membership. However, the basic uncertainty sampling method with a fixed threshold can either easily stop the acquisition, leaving the base learner unfitted or keep frequent acquisition with an already well-trained base learner. To enable the exploitation-oriented agent to intelligently identify the acquisition demand, we formulate it as a reinforcement learning problem which aims at learning an adaptive threshold as the optimal policy to maximize the cumulative reward. As suggested by [71], a reinforcement learning based controller is designed to adjust the certainty threshold θ based on the contribution of historical acquisition decisions. In detail,

upon receiving a sample $\boldsymbol{x}_t$ at time t, the prediction certainty $ct(\boldsymbol{x}_t) = \max \boldsymbol{P}^f(\hat{y}|\boldsymbol{x}_t)$ obtained by the base learner f_t is compared with the current certainty threshold θ_t to make the acquisition decision. The reward r_t will be received if $\boldsymbol{x}_t$ is acquired, which follows a consistent definition (i.e., $r_t \in \{0, \rho^+, \rho^-\}$) as defined in (3.1). Afterwards, the certainty threshold will be updated as:

$$\theta_{t+1} = \min\left\{\theta_t(1 + \eta \cdot (1 - 2^{\frac{r_t}{\rho^-}})), 1\right\}. \tag{3.3}$$

Note that the threshold will increase slightly with a positive reward and vice versa, which enables a policy adaptive to the decision boundary learned by the base learner. The algorithm of the reinforced exploitation agent is detailed in Algorithm 5.

Algorithm 5: Reinforced Exploitation Agent (RAL-Agent)

Input: $\boldsymbol{x}_t, \theta_0, \eta, \rho^+, \rho^-$
if $ct(\boldsymbol{x}_t) < \theta_t$ **then**
  $p_t = 1$, obtain the reward r_t
  Update the certainty threshold $\theta_{t+1} = \min\left\{\theta_t(1 + \eta \cdot (1 - 2^{\frac{r_t}{\rho^-}})), 1\right\}$
else
  $p_t = 0$
Output: Acquisition probability p_t

3.4 Characterization of Agents

To validate the exploration and exploitation capability of the proposed agents, theoretical justification is provided for the designed acquisition criteria. The variance of the acquired samples $\mathcal{D}_t$ by one agent serves as an appropriate metric for assessing its exploration and exploitation activity during the online annotation process. A higher variance suggests a learner's ability to explore the input variable space via acquiring samples in a larger region, whereas a lower variance implies a high frequency of acquisition in a small region for exploitation. To compare the variance of $\mathcal{D}_t$, we examine the probability of a single sample being acquired by the proposed agents.

We assume the streaming data belongs to a mixture of Gaussian distributions. Denote the previously observed samples stored in the sliding window $\mathcal{W}$ before time t as the set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_L\}$, where $\boldsymbol{x}_i$ belongs to the i-th Gaussian distribution (i.e., $\boldsymbol{x}_i \sim \mathcal{N}_q(\boldsymbol{\mu}_i, \Sigma^{(i)})$). Given a streaming sample $\boldsymbol{x}_t$ at time t which follows another Gaussian distribution (i.e., $\boldsymbol{x}_t \sim \mathcal{N}_q(\boldsymbol{\mu}_k, \Sigma^{(k)})$), with the Euclidean distance $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}$, we have the following result for the expectation of the probability that the LD-Agent acquires $\boldsymbol{x}_t$:

Proposition 1. *If the streaming samples follow an independent multivariate Gaussian distribution (i.e., $\Sigma^{(i)} = \sigma_i^2 \boldsymbol{I}$), then we have the expectation of the acquisition probability of a LD-Agent p_t as*

$$\mathbb{E}_{\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_t}[p_t] = \sum_{i=1}^{L}\left\{\int_0^\infty \left[\Pi_{j=1,j\neq i}^{L-1}\Phi\left(\frac{y_i - \left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|_2^2 - (\sigma_i^2 + \sigma_j^2)q}{(\sqrt{4(\sigma_i^2 + \sigma_j^2)\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|_2^2 + 2q(\sigma_i^2 + \sigma_j^2)^2}}\right)\right]\cdot \right.$$
$$\left.\phi\left(\frac{y_i - \left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|_2^2 - (\sigma_i^2 + \sigma_k^2)q}{\sqrt{4(\sigma_i^2 + \sigma_k^2)\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|_2^2 + 2q(\sigma_i^2 + \sigma_k^2)^2}}\right)dy_i\right\}/(L \cdot \delta_L), \tag{3.4}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) for the standard normal distribution and $\phi(\cdot)$ is the probability density function (PDF) for the standard normal distribution.

The expectation of the acquisition probability has the following property:

Theorem 1. *If the streaming samples follow an independent multivariate Gaussian distribution (i.e., $\Sigma^{(i)} = \sigma_i^2 \boldsymbol{I}$), then there exist $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^L$ such that if $\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2 > M_{2,i}, \forall i \in \{1, ..., L\}$, then the expected acquisition probability of a LD-Agent $\mathbb{E}_{\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k}[p_t]$ will exceed 1, where $\boldsymbol{M}_1, \boldsymbol{M}_2$ satisfies:*

$$M_{2,i} + \mathrm{erf}^{-1}(1 - 2\delta_L) \cdot \sqrt{2 \cdot (4(\sigma_i^2 + \sigma_k^2))M_2 + 2q(\sigma_i^2 + \sigma_k^2))} + (\sigma_i^2 + \sigma_k^2)q - M_{1,i} = 0$$

$$M_{1,i} > \left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2 + (\sigma_i^2 + \sigma_j^2)q + \sqrt{4(\sigma_i^2 + \sigma_j^2))\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2 + 2q(\sigma_i^2 + \sigma_j^2)}$$
$$\cdot \left(\Phi^{-1}(1 - \frac{1}{L-1}) + \gamma\left[\Phi^{-1}(1 - \frac{1}{L-1} \cdot e^{-1}) - \Phi^{-1}(1 - \frac{1}{L-1})\right]\right), \forall i \in \{1, ..., L\}. \tag{3.5}$$

This result illustrates that with the increasing of the distance between the center of the distribution of the observed samples and that of the incoming sample, the expected acquisition probability approaches and exceeds 1. This ensures the acquisition of samples from a remote cluster, resulting in an increased variance and thus, the exploration of the input variable space.

For the reinforced exploitation agent, assume a logistic regression model is selected as the base learner, and at time t the base learner f_t is parameterized by $\boldsymbol{\beta}_t$. Given the labelled data pool $\mathcal{D}_t$ at time t, we have the following result for the expectation of the probability that the RAL-Agent acquires $\boldsymbol{x}_t$:

Proposition 2. *If $\boldsymbol{\beta}_t = \mathrm{argmax}\, P(\mathcal{D}_t|\boldsymbol{\beta}) = \mathrm{argmax}\, \Pi_{i=1}^{n_t} P(\boldsymbol{x}_i, y_i|\boldsymbol{\beta})$, then:*

$$\mathbb{E}_{\boldsymbol{x}_t}[p_t] = \int_{-\ln\frac{\theta_t}{1-\theta_t}}^{\ln\frac{\theta_t}{1-\theta_t}} \frac{1}{\sigma_H\sqrt{2\pi}} \exp\left(-\frac{1}{2} \cdot \frac{h - \mu_H}{\sigma_H}\right)^2 dh \tag{3.6}$$

$$\mu_H = \mathbb{E}[H] = \sum_{i=1}^{q} \beta_q \cdot \mu_{k,i} \tag{3.7}$$

$$\sigma_H^2 = \mathbb{V}[H] = \sum_{i=1}^{q} \beta_i^2 \cdot \Sigma_{i,i}^{(k)} + 2\sum_{i=1}^{q}\sum_{j:j>i}^{q} \beta_i\beta_j\Sigma_{i,j}^{(k)}. \tag{3.8}$$

where $\Sigma^{(k)} i, j$ is the $(i, j)-th$ entry of the covariance matrix $\Sigma^{(k)}$.

And the following result is found for the convergence of the expected acquisition probability:

Theorem 2. *Given the labelled data pool $\mathcal{D}_t$ at time t, assume the center of the labelled samples in $\mathcal{D}_t$ is $\boldsymbol{\mu}_i \in \mathbb{R}^q$ and the incoming sample $\boldsymbol{x}_t \sim \mathcal{N}_q(\boldsymbol{\mu}_k, \sigma_k^2\boldsymbol{I})$. With the increase of the distance between two centers $\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2$, there does not exist $M_3 \in \mathbb{R}$ such that $P\{|\mathbb{E}_{\boldsymbol{x}_t}[p_t] - M_3| \geq \epsilon\} = 0, \forall \epsilon \in \mathbb{R}$.*

Since p_t belongs to $[0, 1]$ for a RAL-Agent, the result implies that the acquisition probability of the incoming $\boldsymbol{x}_t$ will not converge with the increase of the distance between the center of the distribution of $\mathcal{D}_t$ and that of $\boldsymbol{x}_t$. Hence, for one sample from a remote cluster, the acquisition decision made by a RAL-Agent is determined by the base learner's uncertainty, which does not necessarily lead to an increasing variance.

In summary, the theoretical analysis justifies the exploration and exploitation capability of the proposed agents. Therefore, with the ensemble of two types of agents, the exploration and exploitation can be dynamically adjusted to balance the trade-off during the human annotation

process. To this end, in the proposed CbeAL, one RAL-agent will be paired with one SPF-Agent or LD-Agent to balance the effort spent on exploration and exploitation. The selection of agents in the ensemble will be further discussed in the Section 4. The derivation and numerical study can be found in the supplemental material due to the page limit.

4 NUMERICAL SIMULATION

4.1 Simulation Setup

Suppose we have a binary classifier as the base learner which requires online updating. Recall the third assumption that multiple clusters exist in the input variable space. Therefore, we adopt a cluster-based classification data set generation method [27, 47] to generate the input $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and the corresponding label $\boldsymbol{y} \in \mathbb{R}^n$, where n is the sample size and p is the input variable dimension. We assume there are two clusters in each class, thus we have $2 \times 2 = 4$ clusters in total. In brief, the centroids of 4 Gaussian clusters are first generated as the vertices of one polytope. Then, the input variables are independently drawn from each Gaussian cluster with unit variance and then multiplied by a random matrix to introduce the random covariance. As the last step, the samples in two of the four clusters will be assigned with the same label as $\boldsymbol{y}$.

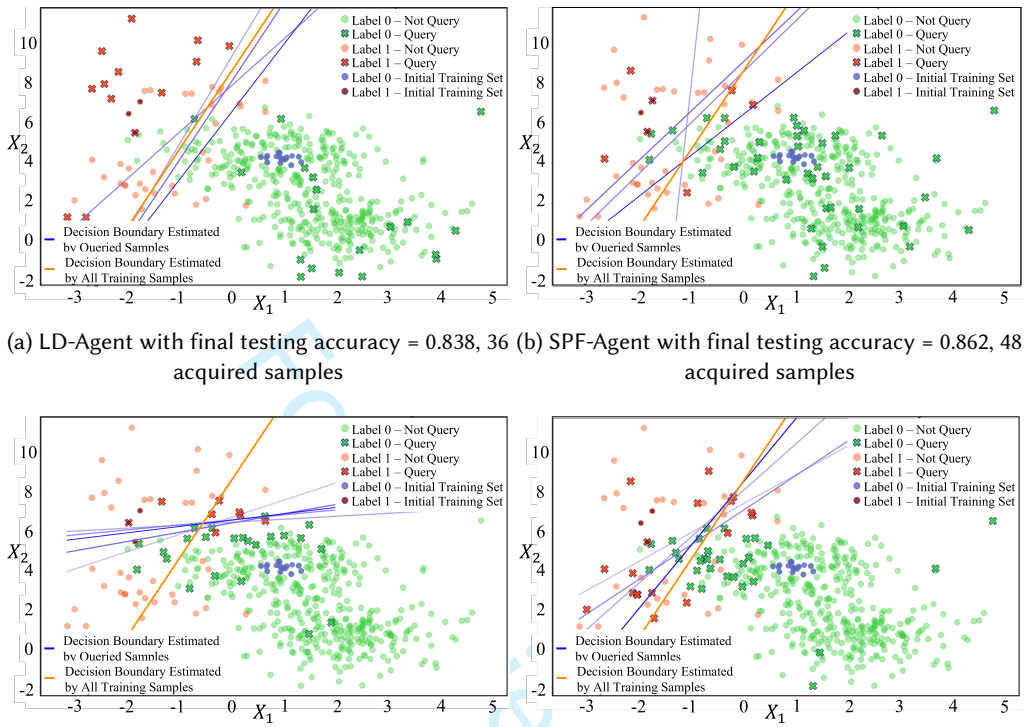
To evaluate CbeAL comprehensively, four settings are varied to generate different online annotation scenarios: (i) training sample size n, which includes both the initial training set and the streaming training set; (ii) the percentage of samples in the positive class pc, which determines the balanceness of the two classes; (iii) the percentage of disturbance ds; and (iv) the percentage of sparsity sp, which is defined as the percentage of insignificant input variables among total p input variables. Note that disturbances are added through flipping the labels of randomly selected samples. Additionally, to control the sparsity level, insignificant variables are randomly generated and concatenated to the informative ones.

The data set generated for each online annotation scenario is subdivided into three subsets: the initial training set, the streaming training set, and the testing set. The initial training set has a constant size of 20 and the testing set has a size of 500. The number of samples in each class is balanced to be equal in the testing set to better illustrate the classification performance of the base learner. For all simulation scenarios, the budget is set to be 10% which gives the number of samples available to be labelled as $B = 10\% \cdot (n - 20)$ during the streaming process. All scenarios are replicated for 10 times with randomly generated data set in each replication.

Based on the suggestion in [32, 71] and grid search in simulation experiments, we set the following values for the hyperparameters in CbeAL: $p_{\min} = \sqrt{\frac{\ln N}{KT}}, T = 2000, \delta = 0.1, \lambda = 0.3, h = 5, \gamma = t/T$, reward $\rho^+ = 1$, penalty $\rho^- = 0.5$. Meanwhile, three pairs of agents with recommended hyperparameter values are incorporated into CbeAL, forming the set of six agents in Table 1. Note that the hyperparmaters can be further tuned for different learning scenarios.

Table 1. Agent Set Adopted in CbeAL

Pair Index	Agent Index	Agent	Hyperparameters
1	AG_1	LD_1	$L = 100, \delta_L = 0.01$
	AG_2	RAL_1	$\theta_0 = 0.95, \eta = 0.005$
2	AG_3	LD_2	$L = 150, \delta_L = 0.005$
	AG_4	RAL_2	$\theta_0 = 0.95, \eta = 0.01$
3	AG_5	SPF_1	$L = 60$
	AG_6	RAL_3	$\theta_0 = 0.90, \eta = 0.01$



(a) LD-Agent with final testing accuracy = 0.838, 36 acquired samples

(b) SPF-Agent with final testing accuracy = 0.862, 48 acquired samples

(c) RAL-Agent with final testing accuracy = 0.786, 31 acquired samples

(d) CBEAL-2 with final testing accuracy = 0.886, 46 acquired samples

Fig. 3. Evolution of the base learner's decision boundary in the toy example: The set of blue lines represents the decision boundaries learned by the base learner every 100 time points where the color depth of the line is proportional to time t; The oragne line is the ground-truth decision boundary.

To demonstrate the simulation setup, Fig. 3 visualizes the generated imbalanced and clusterwise input data of a toy example with 2 input variables (i.e., $p = 2, n = 500, pc = 10\%, sp = 0\%, ds = 0\%, B = 48$). The logistic regression model with default hyperparameters is selected as the base learner [37, 47]. The set of blue lines in Fig. 3 shows the evolution of the decision boundary learned by the base learner with samples acquired by the candidate agents and CBEAL. The color depth of the line is proportional to time t and the darkest blue line is the final decision boundary which will be examined on the testing set to obtain the final testing accuracy. The decision boundary estimated with all training data is considered as the underground-truth shown as orange lines. Here we present the results of the agents with the best performance in the agent set (i.e., Table 1). CBEAL ensembles one exploration agent and one exploitation agent with the best performance in pairs, marked as CBEAL-2.

It is clearly shown that the LD-Agent actively seeks the samples around the boundary of the input variable space, whereas the samples acquired by SPF-Agent are more evenly distributed. Both exploration-oriented agents successfully acquire the samples in both clusters of each class, but very limited samples around the ground-truth decision boundary are selected. In regard to the RAL-Agent, although the uncertainty threshold is supposed to be updated adaptively, Fig. 3 reveals that it is stuck in one cluster of the positive class while the agent keeps acquiring around the

wrongly estimated decision boundary due to the lack of explicit exploration. Combining the two strategies together with the proposed ensemble mechanism, CbeAL-2 acquires samples from both clusters in each class with a focus around the ground-truth decision boundary. This also validates that a well-balanced dynamic trade-off between exploration and exploitation is key to the active learning process with imbalanced and clusterwise input data distribution.

4.2 A Comprehensive Simulation Study

In the comprehensive simulation study, the settings are varied with the following levels: $n \in \{500, 1000, 1500\}$; $pc \in \{10\%, 5\%\}$; $ds \in \{0\%, 3\%\}$; $sp \in \{30\%, 70\%\}$. The input variable dimension is set as $p = 15$. Different from the toy example, SVM is selected as the base learner with default parameters [25, 47], which validates the effectiveness of CbeAL as a generic framework for classification models. Note the prediction probability $\boldsymbol{P}^f(\hat{y}_t|x_t)$ of SVM is estimated and calibrated by Platt scaling [48]. Denote CbeAL-2 as the ensemble of the first pair of agents in Table 1 (i.e., AG_1 and AG_2), CbeAL-4 as the ensemble of the first two pairs and so forth for CbeAL-6. Specifically, CbeAL-6 is proposed as the recommended configuration due to its superior performance enhanced by the ensemble of multiple distinguished agents, which will be detailed in the scalability study.

In this study, CbeAL is firstly compared with the incorparated agents (see Section 3) and their variants to study the effectiveness of the ensemble mechanism. Then, we investigate the impact of the number of agents in the ensemble to study the scalability of CbeAL. Finally, as the recommended configuration, the performance of CbeAL-6 is compared with other four benchmark methods from literature (i.e., uncertainty sampling [40] and random sampling, DBALStream [32] and QBC-PYP [44]) to test the general performance.

In summary, nine benchmark methods are compared with CbeAL-6 where the first three are the candidate agents (i.e., LD, SPF and RAL-Agent), the middle two are the ensemble models with different number of agents (i.e., CbeAL-2 and CbeAL-4), and the last four are methods from literature. Among the benchmarks, RAL-Agent employs an acquisition criterion learned by multi-armed bandits as a cutting-edge AI-guided active learning method; CbeAL-2 and CbeAL-4 adopt the proposed ensemble framework; LD-Agent, SPF-Agent and Random Sampling (RS) focus on the exploration of the input variable space while Uncertainty Sampling (US) [40] caters to exploitation; DBALStream [32] and QBC-PYP [44] are two state-of-the-art composite active learning methods which integrate the objective of exploration and exploitation in their design of acquisition criteria.

CbeAL is evaluated to demonstrate its effectiveness in achieving: (i) high learning performance of the base learner with limited budgets; (ii) a balanced exploration-exploitation trade-off adaptive to each online annotation scenario. As the first evaluation metric, we investigate the classification accuracy of the base learner trained by the samples acquired by each method on the testing set. Then, we compare the percentage of positive samples acquired during the learning process as another metric to illustrate the exploration-exploitation trade-off.

Additionally, we measure the computational time of the proposed method in one replication of all simulation scenarios. The simulation is implemented in Python 3.7.6 on a workstation with 3.70 GHz AMD Ryzen 5 5600X 6-Core Processor, 16.0 GB RAM and Windows 10. It takes on average of 0.086 seconds to make an acquisition decision and train the CbeAL-6 model. This guarantees the practical implementation of the proposed method in manufacturing processes.

4.2.1 Compared with Individual Agents. Firstly, the comparison of base learners' classification accuracy between CbeAL, the incorporated agents, and their variants is shown in Table 2. The performance of the agents that achieve the highest accuracy on average among the exploration- and exploitation-oriented agents in the agent set is selected to be reported as "Opt. Explor." and "Opt. Exploit." Results of other individual agents are omitted here for better readability. It is observed

that in the toy example, some of the agents do not use up the budget B. To validate that it is a fair comparison with different number of acquired samples, we create "Opt. Explor. (Full)" and "Opt. Exploit. (Full)" as two variants where random sampling is used to artificially acquire from the unselected samples after the agents finish their acquisition of the streaming data, until the budget is used up. Besides, the ϵ-greedy policy with $\epsilon = 0.01$ is applied to RAL-Agents when they are used individually, which will effectively improve their learning performance to be a more competitive benchmark [71]. They are also applied to CBEAL methods for a fair comparison.

Table 2. The average values and standard errors (in parenthesis) of classification accuracy in the simulation study. Significant best results are highlighted in **bold**.

Level		Method	Sparsity = 30%			Sparsity = 70%		
			Training Sample Size			Training Sample Size		
Disturbance	Percentage of Positive Samples		500	1000	1500	500	1000	1500
0%	10%	Opt. Explor.	60.1% (0.02)	61.5% (0.02)	63.3% (0.03)	69.8% (0.04)	72.2% (0.03)	69.7% (0.03)
		Opt. Explor. (Full)	60.1% (0.02)	61.6% (0.02)	63.3% (0.03)	69.8% (0.04)	72.2% (0.03)	69.7% (0.03)
		Opt. Exploit.	58.1% (0.02)	70.2% (0.02)	70.4% (0.02)	67.4% (0.03)	72.6% (0.03)	73.7% (0.01)
		Opt. Exploit. (Full)	58.5% (0.03)	70.1% (0.03)	70.4% (0.02)	68.7% (0.02)	72.9% (0.03)	74.3% (0.01)
		CBEAL-2	58.1% (0.02)	70.3% (0.03)*	71.8% (0.03)*	67.5% (0.03)	**75.9**% (0.02)*	74.3% (0.03)*
		CBEAL-6	**61.2% (0.02)**	**73.9% (0.03)**	**72.5% (0.02)**	**69.9% (0.03)**	73.9% (0.03)	**76.7% (0.02)**
	5%	Opt. Explor.	52.6% (0.01)	60.8% (0.03)	60.5% (0.03)	59.8% (0.02)	64.3% (0.03)	66.3% (0.03)
		Opt. Explor. (Full)	52.6% (0.01)	60.8% (0.03)	60.5% (0.03)	59.8% (0.02)	64.3% (0.03)	66.2% (0.03)
		Opt. Exploit.	61.0% (0.03)	70.2% (0.03)	64.6% (0.03)	64.5% (0.03)	68.4% (0.04)	70.5% (0.03)
		Opt. Exploit. (Full)	61.1% (0.03)	70.2% (0.03)	64.6% (0.03)	65.2% (0.03)	68.8% (0.04)	**70.6% (0.03)**
		CBEAL-2	60.6% (0.02)	70.4% (0.03)*	65.2% (0.02)*	63.9% (0.03)	68.7% (0.04)*	67.4% (0.03)
		CBEAL-6	**65.3% (0.02)**	**72.0% (0.02)**	**66.7% (0.02)**	**68.9% (0.03)**	**69.4% (0.03)**	69.0% (0.03)
3%	10%	Opt. Explor.	60.4% (0.02)	60.8% (0.02)	62.1% (0.03)	62.0% (0.02)	62.1% (0.02)	71.7% (0.03)
		Opt. Explor. (Full)	60.5% (0.02)	60.6% (0.03)	62.1% (0.03)	62.0% (0.03)	62.4% (0.02)	71.7% (0.03)
		Opt. Exploit.	68.1% (0.03)	69.3% (0.03)	72.9% (0.03)	**68.9% (0.03)**	69.2% (0.02)	75.0% (0.04)
		Opt. Exploit. (Full)	60.5% (0.03)	69.8% (0.03)	73.3% (0.02)	**68.9% (0.03)**	69.3% (0.02)	75.3% (0.04)
		CBEAL-2	67.7% (0.03)	**72.2% (0.03)***	73.1% (0.03)*	65.1% (0.04)	69.3% (0.02)*	77.5% (0.01)*
		CBEAL-6	**69.0% (0.04)**	71.6% (0.03)	**73.8% (0.02)**	68.0% (0.03)	**70.8% (0.02)**	**79.4% (0.02)**
	5%	Opt. Explor.	53.5% (0.01)	57.6% (0.02)	60.8% (0.03)	54.8% (0.02)	63.9% (0.03)	65.8% (0.02)
		Opt. Explor. (Full)	53.5% (0.01)	57.5% (0.02)	60.8% (0.03)	54.8% (0.02)	63.9% (0.03)	65.8% (0.02)
		Opt. Exploit.	61.0% (0.03)	67.4% (0.03)	**69.2% (0.02)**	**62.5% (0.03)**	70.7% (0.03)	77.1% (0.02)
		Opt. Exploit. (Full)	62.4% (0.03)	67.4% (0.03)	69.0% (0.02)	62.1% (0.03)	71.3% (0.04)	78.3% (0.02)
		CBEAL-2	61.5% (0.02)*	65.2% (0.04)	65.6% (0.03)	60% (0.03)	72.3% (0.03)*	77.9% (0.02)*
		CBEAL-6	**61.9% (0.02)**	**68.3% (0.03)**	64.5% (0.03)	58.8% (0.03)	**74.6% (0.03)**	**80.1% (0.03)**

Table 2 summarizes the averages of the classification accuracy and standard errors over 10 replications of the base learner trained by $\mathcal{D}_t$. It can be observed that the proposed CBEAL-6 outperforms the best individual agent under 20 among 24 scenarios, which verifies that the ensemble of multiple agents with explicit consideration for both exploration and exploitation can effectively enhance the learning performance under highly imbalanced class distribution. The advantage on learning performance compared to benchmarks is more significant when there is no disturbance and the class proportion is more balanced (i.e., $ds = 0\%, pc = 10\%$). However, with a more severe imbalance (i.e., $pc = 5\%$), "Opt. Exploit." and its variants sometimes achieve slightly higher accuracy. One possible reason is that under such scenarios, it will be more efficient to only focus on decision boundary learning since the number of positive samples is limited. Besides, the inferior performance of CBEAL-6 under the scenario $ds = 3\%, pc = 10\%, n = 500, sp = 70\%$ can be caused by the disturbance. It can be found that, in general, the learning performance of the base learner is improved with a data stream with a large size and a higher sparsity. However, when the sparsity is low, the accuracy sometimes decreases as the training sample size increases, which can be caused by the high imbalance and the lack of degree of freedom.

Another finding is CBEAL-2 achieves comparable performance with the better of the two incorporated agents, which implies that the proposed ensemble framework enables the intelligent selection

among candidate agents in an adaptive manner. Under 14 out of 24 scenarios, it outperforms both "Opt. Explor." and "Opt. Exploit.", where the results are marked with *.

Comparing the "Opt. Explor." with "Opt. Explor. (Full)" and "Opt. Exploit." with "Opt. Exploit. (Full)", we find that consuming the remaining budget by random acquisition will not make a significant improvement on the learning performance under most scenarios. And sometimes it will select less contributive samples, which leads to a lower accuracy due to the highly imbalanced distribution. Therefore, it validates that the agents have acquired the most informative samples based on their criteria. Thus, this variant will not be considered in the following analysis.

Table 3. The average values and standard errors (in parenthesis) of the percentage of positive samples in the labelled data pool $\mathcal{D}_t$ in the simulation study. Highest values are highlighted in **bold**.

Level		Method	Sparsity = 30%			Sparsity = 70%		
			Training Sample Size			Training Sample Size		
Disturbance	Percentage of Positive Samples		500	1000	1500	500	1000	1500
0%	10%	Opt. Explor.	19.1% (0.02)	16.8% (0.01)	15.9% (0.02)	23.3% (0.03)	23.7% (0.02)	20.5% (0.03)
		Opt. Exploit.	27.3% (0.02)	37.9% (0.01)	42.5% (0.01)	28.0% (0.02)	38.1% (0.02)	46.9% (0.01)
		CBEAL-2	24.9% (0.02)	42.8% (0.02)*	43.9% (0.04)*	33.7% (0.02)*	**45.1% (0.01)***	43.4% (0.04)
		CBEAL-6	**31.3% (0.01)**	**49.9% (0.02)**	**47.0% (0.02)**	**36.1% (0.01)**	42.8% (0.02)	**48.5% (0.03)**
	5%	Opt. Explor.	10.2% (0.01)	9.20% (0.01)	11.0% (0.01)	12.8% (0.01)	12.9% (0.01)	13.2% (0.01)
		Opt. Exploit.	23.6% (0.01)	29.7% (0.01)	31.7% (0.01)	23.7% (0.01)	27.4% (0.01)	29.9% (0.03)
		CBEAL-2	22.8% (0.02)	30.3% (0.02)*	31.6% (0.03)	23.7% (0.02)*	26.1% (0.02)	30.8% (0.03)*
		CBEAL-6	**27.2% (0.01)**	**34.4% (0.01)**	**32.4% (0.01)**	**27.1% (0.01)**	**30.9% (0.02)**	**33.6% (0.02)**
3%	10%	Opt. Explor.	16.4% (0.01)	16.8% (0.01)	16.4% (0.02)	20.7% (0.02)	19.9% (0.01)	19.0% (0.01)
		Opt. Exploit.	28.6% (0.02)	37.9% (0.03)	48.0% (0.02)	28.1% (0.01)	41.6% (0.02)	48.9% (0.02)
		CBEAL-2	34.0% (0.03)*	**44.6% (0.03)***	45.5% (0.04)	32.4% (0.03)*	44.4% (0.02)*	50.9% (0.03)*
		CBEAL-6	**37.9% (0.03)**	44.2% (0.04)	**53.7% (0.02)**	**35.0% (0.02)**	**45.4% (0.02)**	**56.2% (0.03)**
	5%	Opt. Explor.	12.0% (0.01)	10.8% (0.01)	11.0% (0.01)	14.3% (0.01)	12.5% (0.01)	12.6% (0.01)
		Opt. Exploit.	**23.7% (0.01)**	**33.6% (0.03)**	33.1% (0.03)	20.9% (0.02)	**35.6% (0.01)**	41.9% (0.02)
		CBEAL-2	23.3% (0.02)	27.9% (0.04)	32.0% (0.03)	20.7% (0.02)	28.4% (0.03)	42.6% (0.02)*
		CBEAL-6	22.3% (0.02)	32.7% (0.03)	**33.6% (0.04)**	**22.0% (0.02)**	35.3% (0.03)	**44.7% (0.03)**

To further investigate the exploration-exploitation trade-off achieved in the active learning process, we calculate and summarize the percentage of positive samples in the labelled data pool $\mathcal{D}_t$ in Table 3. Based on the results, we found the proposed CbeAL-6 obtains a significant better balanced labelled data set $\mathcal{D}_t$ compared to the benchmarks under most scenarios. The percentage of positive samples is close to or even higher than 50% under some scenarios, which shows a well-balanced exploration and exploitation trade-off. Moreover, the classification accuracy of a method in Table 2 is higher if it acquires a higher percentage of positive samples in Table 3, which indicates the percentage of positive samples in $\mathcal{D}_t$ is closely related to the base learner's learning performance.

4.2.2 Scalability Study. It has been observed in the previous results (i.e., Tables 2, 3) that CbeAL-6 achieves better performance than CbeAL-2 in general. Here, we further investigate the following two questions: what will be the impact of the ensemble of varying numbers of agent pairs and how should the agents be selected. The classification accuracy of the base learners of CbeAL-2, CbeAL-4 and CbeAL-6 are compared to study the scalability of CbeAL.

From the results summarized in Table 4, We observe that CbeAL-6 demonstrates a dominate superiority in the learning performance, which indicates the advantage brought by multiple agents. However, comparing the result of CbeAL-4 with CbeAL-2, CbeAL-4 achieves better performance in fewer than half of all scenarios. The counterintuitive result indicates that the ensemble of more agents may not make improvement on the performance. To identify the reason, we investigate the acquisition decision made by each agent in the agent set and their standardized weights in

Table 4. The average values and standard errors (in parenthesis) of the classification accuracy in the simulation study. Significant best results are highlighted in **bold**.

Level		Method	Sparsity = 30%			Sparsity = 70%		
			Training Sample Size			Training Sample Size		
Disturbance	Percentage of Positive Samples		500	1000	1500	500	1000	1500
0%	10%	CBEAL-2	58.1% (0.02)	70.3% (0.03)	71.8% (0.03)	67.5% (0.03)	75.9% (0.02)	74.3% (0.03)
		CBEAL-4	**61.5% (0.03)**	67.3% (0.03)	**74.3% (0.02)**	69.6% (0.03)	**77.4% (0.02)**	73.1% (0.03)
		CBEAL-6	61.2% (0.02)	**73.9% (0.03)**	72.5% (0.02)	**69.9% (0.03)**	73.9% (0.03)	**76.7% (0.02)**
	5%	CBEAL-2	60.6% (0.02)	70.4% (0.03)	65.2% (0.03)	63.9% (0.03)	68.7% (0.04)	67.4% (0.03)
		CBEAL-4	60.7% (0.03)	67.9% (0.03)	64.8% (0.02)	63.5% (0.03)	67.6% (0.03)	66.7% (0.03)
		CBEAL-6	**65.3% (0.02)**	**72.0% (0.02)**	**66.7% (0.02)**	**68.9% (0.03)**	**69.4% (0.03)**	**69.0% (0.03)**
3%	10%	CBEAL-2	67.7% (0.03)	**72.2% (0.03)**	73.1% (0.03)	65.1% (0.04)	69.3% (0.02)	77.5% (0.01)
		CBEAL-4	68.2% (0.03)	67.0% (0.03)	73.2% (0.03)	66.7% (0.03)	66.4% (0.02)	74.5% (0.03)
		CBEAL-6	**69.0% (0.04)**	71.6% (0.03)	**73.8% (0.02)**	**68.0% (0.03)**	**70.8% (0.02)**	**79.4% (0.02)**
	5%	CBEAL-2	61.5% (0.02)	65.2% (0.04)	**65.6% (0.03)**	**60.0% (0.03)**	72.3% (0.03)	77.9% (0.02)
		CBEAL-4	57.7% (0.02)	63.2% (0.03)	65.5% (0.03)	58.7% (0.04)	72.7% (0.02)	73.5% (0.03)
		CBEAL-6	**61.9% (0.02)**	**68.3% (0.03)**	64.5% (0.03)	58.8% (0.03)	**74.6% (0.03)**	**80.1% (0.03)**

CbeAL-6 under one scenario in Fig. 4 where CbeAL-4 shows inferior performance than CbeAL-2 but CbeAL-6 shows better performance.

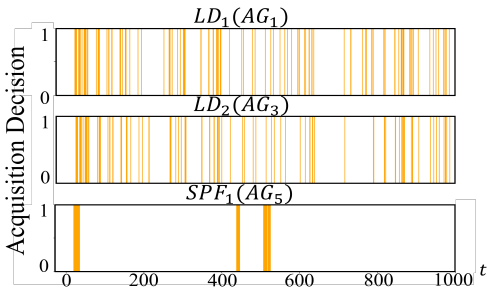
It can be observed from the bar charts (Fig. 4(a)-(c)) that in CbeAL-6, the first two pairs of the agents (LD_1 and LD_2, RAL_1 and RAL_2) make similar acquisition decisions while the third pair behaves differently. As a direct outcome of the homogeneity, the standardized weights of the first two pairs of exploration- and exploration-oriented agents will be close and change in a similar pattern in CbeAL-6 (i.e., Fig. 4(d)), which also causes the comparable performance of CbeAL-4 and CbeAL-2. This as well explains that the superior performance of CbeAL-6 lies in the heterogeneous decisions brought by the third pair of agents (SPF_1 and RAL_3). Besides, the weights of agents in CbeAL-6 indicate that at the beginning, the exploration dominates the active learning process. Later, the proposed CbeAL switch its tendency to exploitation, and the exploration capability still remains adaptive to the data stream, which contributes to the effective and efficient online annotation.

In summary, the ensemble of distinguished agents provides comprehensive criteria to evaluate the informativeness of each streaming sample in terms of exploration and exploitation, thus achieving a well-balanced trade-off. Since the acquisition behaviour of one active learning agent varies under different scenarios and there does not exist one overall winner, CbeAL-6 is recommended as a default configuration to solve these challenging learning tasks.

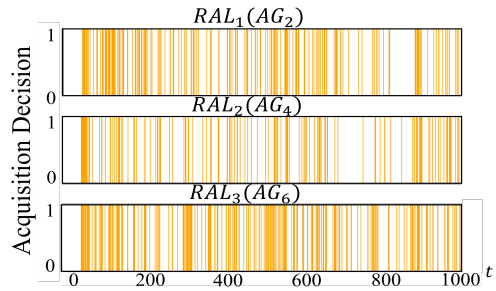
4.2.3 Compared with Benchmark Methods. Finally, CbeAL is compared with other four benchmark methods (i.e., RS, US, DBALStream and QBC-PYP) and Table 5 summarizes the classification accuracy of the base learner trained by samples acquired by each method.

By investigating the results in Table 5, it is concluded that CbeAL-6 achieves significantly better performance compared to the benchmarks under most scenarios. With a limited budget, CbeAL-6 can achieve high classification accuracy close to that of using all the training data with a data stream of larger size and higher sparsity (i.e., $n = 1500, sp = 70\%$).

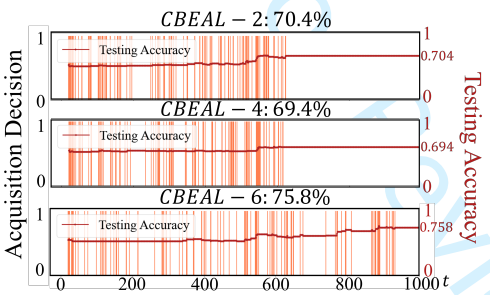
Considering the benchmark methods, US demonstrates its competitive performance compared to other benchmarks but with higher standard errors. This indicates the importance of exploitation for the online annotation scenarios, and this also explains the superiority of RAL-Agents in Table 2. However, the lack of adaptiveness to the data stream causes its inferior performance compared to CbeAL-6. The inferior performance of DBALStream might attribute to its concentration on samples with both high local density and large margin, which does not perform effective input variable space exploration. On the contrary, the proposed LD-Agent is able to complete this exploration



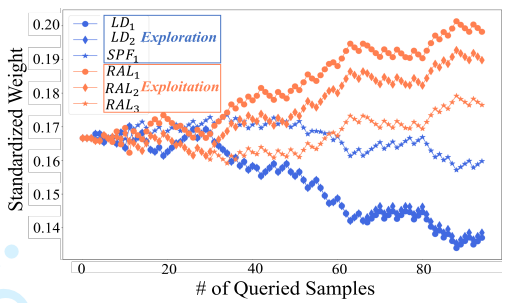
(a) Acquisition decisions made by LD_1, LD_2, SPF_1 agents incorporated in CBEAL-6



(b) Acquisition decisions made by RAL_1, RAL_2, RAL_3 agents incorporated in CBEAL-6



(c) Acquisition decisions made by CBEAL-2, CBEAL-4, CBEAL-6 and their testing accuracy



(d) Standardized weight of each agents in CBEAL-6

Fig. 4. Testing accuracy of CBEAL methods under the learning scenario $n = 1000$, $ds = 0\%$, $sp = 30\%$, $pc = 10\%$. (a)-(c): Bar charts of the acquisition decisions made by all candidate agents and CBEAL methods; (d) standardized weights of each candidate agent in CBEAL-6.

task. The accuracy of QBC-PYP demonstrates its capability of joint exploration and exploitation under a highly imbalanced class distribution. However, it takes a mixture of Gaussians to quantify the ambiguity of the class membership of one sample, which fails to adapt to the performance of the base learner and thus yields inferior performance.

Overall, the results validate that the proposed method can effectively and efficiently acquire samples in an adaptive manner under various circumstances, thus confirming the benefits of the ensemble of designed exploration-oriented and exploitation-oriented agents.

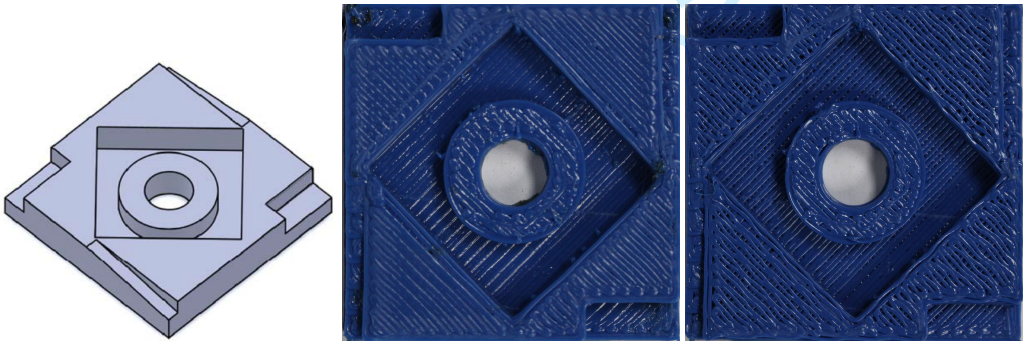
5 CASE STUDY

The proposed CBEAL method is applied to a FDM process for online quality modeling and inspection [15], which is introduced as the motivation example in Section 1. During the printing process, various *in situ* process variables (i.e., vibration, nozzle temperature, etc.) are collected in the ICPS to monitor the process and predict the quality of the FDM part [58]. Here we focus on the layerwise surface roughness as a binary quality indicator, which is judged and annotated as conforming/nonconforming by domain experts. Fig. 5 shows the example normal and rough surfaces of the printed FDM part. To enable real-time quality prediction and online modeling, the *in situ* measurements are registered and divided into 10-second windows as samples. The online updating of the quality model requires experts to consistently observe and examine the surface

Table 5. The average values and standard errors (in parenthesis) of the classification accuracy in the simulation study. Significant best results are highlighted in **bold**.

Level		Method	Sparsity = 30%			Sparsity = 70%		
			Training Sample Size			Training Sample Size		
Disturbance	Percentage of Positive Samples		500	1000	1500	500	1000	1500
0%	10%	Initial	51.4% (0.00)	52.3% (0.01)	52.2% (0.01)	51.0% (0.00)	51.1% (0.01)	50.9% (0.00)
		RS	52.0% (0.01)	57.3% (0.02)	55.6% (0.03)	57.4% (0.02)	60.6% (0.02)	62.1% (0.02)
		US	60.6% (0.03)	70.3% (0.04)	67.9% (0.04)	65.6% (0.04)	72.8% (0.04)	72.4% (0.04)
		DBALStream	54.6% (0.01)	60.6% (0.02)	58.1% (0.02)	56.7% (0.01)	60.7% (0.02)	59.4% (0.02)
		QBC-PYP	51.2% (0.01)	67.8% (0.04)	68.1% (0.03)	63.5% (0.04)	70.3% (0.04)	66.1% (0.04)
		CBEAL-6	**61.2% (0.02)**	**73.9% (0.03)**	**72.5% (0.02)**	**69.9% (0.03)**	**73.9% (0.03)**	**76.7% (0.02)**
		All Training Data	76.7% (0.01)	78.9% (0.02)	81.5% (0.01)	77.7% (0.01)	81.5% (0.01)	80.7% (0.01)
	5%	Initial	52.5% (0.01)	52.9% (0.01)	51.1% (0.00)	51.7% (0.01)	52.1% (0.01)	51.5% (0.01)
		RS	51.9% (0.01)	53.7% (0.01)	55.1% (0.02)	54.9% (0.02)	55.6% (0.03)	53.8% (0.01)
		US	62.2% (0.04)	70.3% (0.04)	61.6% (0.04)	**69.8% (0.03)**	66.6% (0.04)	68.3% (0.04)
		DBALStream	54.7% (0.01)	56.3% (0.00)	54.2% (0.01)	54.4% (0.01)	55.2% (0.01)	54.6% (0.01)
		QBC-PYP	53.4% (0.03)	61.9% (0.03)	60.2% (0.03)	58.0% (0.03)	64.0% (0.06)	63.8% (0.05)
		CBEAL-6	**65.3% (0.02)**	**72.0% (0.02)**	**66.7% (0.02)**	68.9% (0.03)	**69.4% (0.03)**	**69.0% (0.03)**
		All Training Data	75.1% (0.02)	78.7% (0.01)	78.4% (0.01)	77.7% (0.01)	80.2% (0.01)	80.7% (0.01)
3%	10%	Initial	51.9% (0.01)	51.3% (0.01)	50.7% (0.00)	51.1% (0.00)	50.9% (0.00)	51.2% (0.00)
		RS	53.0% (0.01)	56.2% (0.02)	57.7% (0.01)	54.3% (0.01)	53.8% (0.01)	60.6% (0.03)
		US	60.7% (0.04)	68.0% (0.04)	70.5% (0.04)	67.6% (0.04)	66.2% (0.04)	71.0% (0.05)
		DBALStream	56.9% (0.02)	60.1% (0.03)	60.7% (0.02)	56.1% (0.01)	57.1% (0.02)	61.3% (0.02)
		QBC-PYP	61.2% (0.03)	56.8% (0.03)	68.8% (0.04)	64.5% (0.04)	63.6% (0.04)	66.4% (0.04)
		CBEAL-6	**69.0% (0.04)**	**71.6% (0.03)**	**73.8% (0.02)**	**68.0% (0.03)**	**70.8% (0.02)**	**79.4% (0.02)**
		All Training Data	79.2% (0.02)	78.8% (0.01)	81.0% (0.01)	80.9% (0.01)	76.4% (0.01)	81.1% (0.01)
	5%	Initial	52.2% (0.01)	53.2% (0.01)	54.1% (0.01)	51.5% (0.01)	51.2% (0.01)	51.5% (0.00)
		RS	52.9% (0.01)	54.4% (0.01)	52.8% (0.02)	53.2% (0.02)	54.5% (0.02)	52.6% (0.01)
		US	59.4% (0.03)	67.9% (0.04)	**68.0% (0.03)**	59.9% (0.04)	62.6% (0.04)	73.9% (0.04)
		DBALStream	54.3% (0.01)	56.0% (0.01)	54.7% (0.01)	55.6% (0.01)	55.1% (0.01)	57.6% (0.02)
		QBC-PYP	**62.3% (0.04)**	58.1% (0.04)	62.5% (0.04)	**60.2% (0.04)**	64.2% (0.04)	61.3% (0.04)
		CBEAL-6	61.9% (0.02)	**68.3% (0.03)**	64.5% (0.03)	58.8% (0.03)	**74.6% (0.03)**	**80.1% (0.03)**
		All Training Data	73.4% (0.02)	78.0% (0.01)	76.6% (0.01)	73.1% (0.02)	78.1% (0.01)	80.5% (0.01)

roughness during the printing process for the window-wise annotation, which is labor-intensive and time-consuming. Therefore, CBEAL is employed to develop an accurate quality model with less labeling efforts and high-quality training data through wisely selecting the samples for annotation.

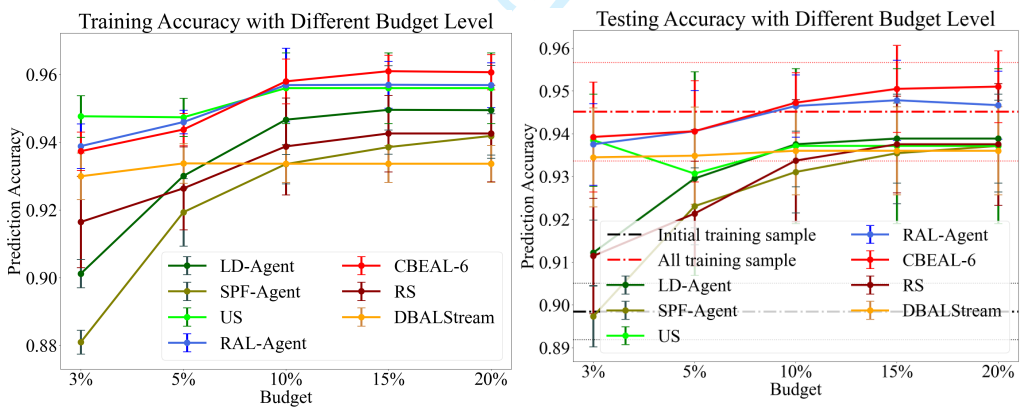


(a) Printed FDM part modified from NAS 979 part (b) Normal surface (c) Rough surface

Fig. 5. Printed part in the case study and the example surfaces, (a) is modified from Huang et al. [30] with authors' permission

We refer [49, 58] for the details of the experiment sensor network and we refer [15] for the details of the data collection. During the process, the *in situ* extruder vibration, table vibration, nozzle temperature and table temperature are measured and collected in a functional data format. Considering the wavelet analysis applied to the funcational measures, the process setting variables (i.e., feed/flow ratio and layer thickness) and summary statistics for each functional measurement (i.e., mean, standard deviation, skewness, and kurtosis), 519 features are obtained in total as the model input. 48 FDM parts are printed successfully in total with 1588 samples (i.e., windows of measurements). Label the nonconforming samples by 1 and conforming samples by 0. As a common scenario in quality inspection, we find there are 180 nonconforming samples in total, which implies a highly imbalanced class distribution. We also notice that the positive labels appear in succession during the process (i.e., $\ldots 0000011100 \ldots$) because the malfunction of the printer in a period of time will affect the quality of several consecutive layers. With this in mind, we maintain the original order of samples in sequence when we separate the training and testing set. A logistic regression model with $L1$ penalty is adopted as the base learner for the quality online modeling. In brief, the highly imbalanced data stream with patterns in sequence and an underlying multimodal distribution (i.e., Fig.1) brings a challenging online annotation scenario.

We evaluate the classification accuracy of the base learner of the proposed CBEAL-6 under different level of budgets (i.e., $\{3\%, 5\%, 10\%, 15\%, 20\%\}$) with 10 replications. In each replication, 1/3 samples are extracted with a random starting point from the whole data set in time order as the testing data set, with the remaining for training. The first 10 samples in the training set will be used for the model pretraining as D_0. The best performed individual agents in Table 1 and the other four benchmarks in Section 4 are employed for the comparison.



(a) Averaged training classification accuracy (b) Averaged testing classification accuracy

Fig. 6. The average values of training and testing classification accuracy of CBEAL and benchmark methods for the case study

Fig. 6 demonstrates the averages of the training and testing classification accuracy of the base learners trained by samples acquired by different methods, where the error bars represent the standard errors over 10 replication. The dash-dotted red line represents the testing accuracy of the base learner trained by all training data and the dotted red line is the standard error. Correspondingly, the testing accuracy of the base learner trained by the initial training data is marked with black lines in a similar way. Here, the result of QBC-PYP is not included since it consistently refuses to

acquire samples during the streaming process, which might be caused by the pattern of positive labels in the sequence that requires the method to adapt its exploration-exploitation tendency to the data stream.

By investigating the results, it can be observed that the proposed method consistently outperforms its incorporated agents and the rest of benchmark methods in testing accuracy under different levels of budgets. The testing classification accuracy of most benchmark methods has an increasing trend as the budget increases from 3% to 10% and then the trend goes smoother, which indicates the samples acquired in the 10% budget make the most of the contribution to the quality modeling. However, the testing accuracy of CbeAL-6 keeps increasing with a higher budget, which validates the continuous acquisition of informative samples. Notably, the testing accuracy of CbeAL-6 exceeds the accuracy of the base learner trained by all available samples when the budget is higher than 10%. This implies, despite the high dimension, a training set with a smaller sample size but better balanced samples has better quality, thus improving the modeling accuracy. Therefore, the proposed method not only reduces the labelling efforts but also contributes to the online modeling performance via acquiring high-quality samples.

In conclusion, the case study verifies that the proposed CbeAL method achieves a well-balanced exploration-exploitation trade-off during the streaming process in an adaptive manner, which enables the highly accurate online quality modeling with limited labelling efforts for the FDM quality inspection.

6 CONCLUSION

While the high-speed, high-volume streaming data brought by the ICPS enhance the data-driven decision making by AI models, the quality of online data may hamper the AI modeling performance for manufacturing. To provide resilient AI modeling performance, informative samples need to be acquired from the streaming data to provide a high-quality training data set as well as reducing the human annotation efforts required for AI incubation in an online manner. Existing active learning methods cannot well balance the exploration-exploitation trade-off in the challenging online annotation scenario. In this work, we propose an ensemble of exploration- and exploitation-oriented active learning agents as CbeAL to balance the exploration-exploitation trade-off. With the pairwise ensemble of agents considering each objective explicitly and the modified Exp4.P-EWMA solver, CbeAL adjusts the weight of each incorporated agent to control the exploration and exploitation tendency adaptive to the streaming data and the learning performance of the base learner, thus making intelligent acquisition decisions. A comprehensive simulation study and the case study in FDM processes demonstrates the advantage of CbeAL over benchmark methods in learning accuracy and the balanced trade-off with limited budget for the sample size.

We notice some limitations of the proposed method. First, the proposed CbeAL shows its superiority under learning scenarios with complex input data distribution. Under other generic scenarios without demanding exploration-exploitation trade-off, CbeAL may lose its advantage with the current ensemble setting due to the encoded explicit exploration objective. In this case, instead of a pairwise ensemble, CbeAL can enhance its concentration on exploitation through adding exploitation-oriented agents or removing exploration-oriented agents. Second, the hyperparameters for the agents are optimized and selected with the grid search in the study. We will formulate the hyperparmater tuning as a meta-learning problem for different base learners or data sets as the future work [61].

The work leaves us several future research directions. First, the reward function in CbeAL can be modified to quantify the informativeness of a sample in terms of the exploitation in regression problems [10], thus extending the framework to generic supervised models. Second, we will study to employ the CbeAL framework for active data space partition to identify the appropriate data

collection method (i.e., DoE or observational data) for each partitioned region. This will help to balance the quality and cost of data collection for an ever-changing underlying model. Furthermore, we consider the ensemble of multiple modalities as the second level actions under the framework of CbeAL. With a hierarchical decision of actions, the learner can decide not only whether to acquire the sample but also from which data resource it should acquire [68].

ACKNOWLEDGMENTS

The authors acknowledge Dr. Xinwei Deng, Ruojun Wang, and Rou Wen for their comments and suggestions to improve the paper.

REFERENCES

[1] Rebecca Adaimi and Edison Thomaz. 2019. Leveraging active learning and conditional mutual information to minimize data annotation in human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–23.

[2] Kosmas Alexopoulos, Nikolaos Nikolakis, and George Chryssolouris. 2020. Digital twin-driven supervised machine learning for the development of artificial intelligence applications in manufacturing. *International Journal of Computer Integrated Manufacturing* 33, 5 (2020), 429–439.

[3] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing* 32, 1 (2002), 48–77.

[4] Maria-Florina Balcan, Andrei Broder, and Tong Zhang. 2007. Margin based active learning. In *International Conference on Computational Learning Theory*. Springer, 35–50.

[5] Yoram Baram, Ran El Yaniv, and Kobi Luz. 2004. Online choice of active learning algorithms. *Journal of Machine Learning Research* 5, Mar (2004), 255–291.

[6] Kaveh Bastani, Prahalad K Rao, and Zhenyu Kong. 2016. An online sparse estimation-based classification approach for real-time monitoring in advanced manufacturing processes from heterogeneous sensor data. *IIE Transactions* 48, 7 (2016), 579–598.

[7] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. 2011. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 19–26.

[8] Alexis Bondu, Vincent Lemaire, and Marc Boullé. 2010. Exploration vs. exploitation in active learning: A bayesian approach. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–7.

[9] Paula Branco, Luís Torgo, and Rita P Ribeiro. 2016. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)* 49, 2 (2016), 1–50.

[10] Wenbin Cai, Ya Zhang, and Jun Zhou. 2013. Maximizing expected model change for active learning in regression. In *2013 IEEE 13th International Conference on Data Mining*. IEEE, 51–60.

[11] Nicholas Carlini. 2021. Poisoning the Unlabeled Dataset of Semi-Supervised Learning. *arXiv preprint arXiv:2105.01622* (2021).

[12] Emily Caveness, Paul Suganthan GC, Zhuo Peng, Neoklis Polyzotis, Sudip Roy, and Martin Zinkevich. 2020. Tensorflow data validation: Data analysis and validation in continuous ml pipelines. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2793–2796.

[13] Kian Ming Adam Chai, Hai Leong Chieu, and Hwee Tou Ng. 2002. Bayesian online classifiers for text classification and filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval*. 97–104.

[14] Xiaoyu Chen. 2021. *Multiscale Quantitative Analytics of Human Visual Searching Tasks*. Ph.D. Dissertation. Virginia Tech. http://hdl.handle.net/10919/104200

[15] Xiaoyu Chen, Hongyue Sun, and Ran Jin. 2016. Variation analysis and visualization of manufacturing processes via augmented reality. In *presented at the IIE Annual Conference. Proceedings*.

[16] Xiaoyu Chen, Yingyan Zeng, Sungku Kang, and Ran Jin. 2022. INN: An Interpretable Neural Network for AI Incubation in Manufacturing. *ACM Transactions on Intelligent Systems and Technology (TIST)* Accepted (2022).

[17] Armando W Colombo, Stamatis Karnouskos, Okyay Kaynak, Yang Shi, and Shen Yin. 2017. Industrial cyberphysical systems: A backbone of the fourth industrial revolution. *IEEE Industrial Electronics Magazine* 11, 1 (2017), 6–16.

[18] Alexander Diete, Timo Sztyler, and Heiner Stuckenschmidt. 2017. A smart data annotation tool for multi-sensor activity recognition. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 111–116.

[19] Sandra Ebert, Mario Fritz, and Bernt Schiele. 2012. Ralf: A reinforced active learning formulation for object class recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3626–3633.
[20] Dina Elreedy, Amir F Atiya, and Samir I Shaheen. 2019. A novel active learning regression framework for balancing the exploration-exploitation trade-off. *Entropy* 21, 7 (2019), 651.
[21] Ronald Aylmer Fisher. 1936. Design of experiments. *Br Med J* 1, 3923 (1936), 554–554.
[22] Yifan Fu, Xingquan Zhu, and Bin Li. 2013. A survey on instance selection for active learning. *Knowledge and information systems* 35, 2 (2013), 249–283.
[23] Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. 1999. Selective sampling for example-based word sense disambiguation. *arXiv preprint cs/9910020* (1999).
[24] Christian Gobert, Edward W Reutzel, Jan Petrich, Abdalla R Nassar, and Shashi Phoha. 2018. Application of supervised machine learning for defect detection during metallic powder bed fusion additive manufacturing using high resolution imaging. *Additive Manufacturing* 21 (2018), 517–528.
[25] Steve R Gunn et al. 1998. Support vector machines for classification and regression. *ISIS technical report* 14, 1 (1998), 5–16.
[26] Nitin Gupta, Shashank Mujumdar, Hima Patel, Satoshi Masuda, Naveen Panwar, Sambaran Bandyopadhyay, Sameep Mehta, Shanmukha Guttula, Shazia Afzal, Ruhi Sharma Mittal, et al. 2021. Data quality for machine learning tasks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 4040–4041.
[27] I. Guyon. 2003. Design of experiments for the NIPS 2003 variable selection benchmark.
[28] Q Peter He and Jin Wang. 2007. Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes. *IEEE transactions on semiconductor manufacturing* 20, 4 (2007), 345–354.
[29] Wei-Ning Hsu and Hsuan-Tien Lin. 2015. Active learning by learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.
[30] Wenyan Huang, Xiaoyu Chen, Ran Jin, and Nathan Lau. 2020. Detecting cognitive hacking in visual inspection with physiological measurements. *Applied ergonomics* 84 (2020), 103022.
[31] J Stuart Hunter. 1986. The exponentially weighted moving average. *Journal of quality technology* 18, 4 (1986), 203–210.
[32] Dino Ienco, Indrė Žliobaitė, and Bernhard Pfahringer. 2014. High density-focused uncertainty sampling for active learning over evolving stream data. In *Proceedings of the 3rd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*. PMLR, 133–148.
[33] Ran Jin, Chia-Jung Chang, and Jianjun Shi. 2012. Sequential measurement strategy for wafer geometric profile estimation. *Iie transactions* 44, 1 (2012), 1–12.
[34] Ran Jin and Xinwei Deng. 2015. Ensemble modeling for data fusion in manufacturing process scale-up. *IIE Transactions* 47, 3 (2015), 203–214.
[35] Ran Jin, Xinwei Deng, Xiaoyu Chen, Liang Zhu, and Jun Zhang. 2019. Dynamic quality-process model in consideration of equipment degradation. *Journal of Quality Technology* 51, 3 (2019), 217–229.
[36] Ronald W Kennard and Larry A Stone. 1969. Computer aided design of experiments. *Technometrics* 11, 1 (1969), 137–148.
[37] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. 2002. *Logistic regression*. Springer.
[38] Georgios Kostopoulos, Stamatis Karlos, Sotiris Kotsiantis, and Omiros Ragos. 2018. Semi-supervised regression: A recent review. *Journal of Intelligent & Fuzzy Systems* 35, 2 (2018), 1483–1500.
[39] Chen Quin Lam. 2008. *Sequential adaptive designs in computer experiments for response surface model fit*. Ph.D. Dissertation. The Ohio State University.
[40] David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*. Springer, 3–12.
[41] Jingran Li, Ran Jin, and Z Yu Hang. 2018. Integration of physically-based and data-driven approaches for thermal field prediction in additive manufacturing. *Materials & Design* 139 (2018), 473–485.
[42] Yifu Li, Xinwei Deng, Shan Ba, William R Myers, William A Brenneman, Steve J Lange, Ron Zink, and Ran Jin. 2021. Cluster-based data filtering for manufacturing big data systems. *Journal of Quality Technology* (2021), 1–13.
[43] Edo Liberty, Kevin Lang, and Konstantin Shmakov. 2016. Stratified sampling meets machine learning. In *International conference on machine learning*. PMLR, 2320–2329.
[44] Chen Change Loy, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. 2012. Stream-based joint exploration-exploitation active learning. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1560–1567.
[45] Chen Change Loy, Tao Xiang, and Shaogang Gong. 2010. Stream-based active unusual event detection. In *Asian Conference on Computer Vision*. Springer, 161–175.
[46] Thomas Osugi, Deng Kim, and Stephen Scott. 2005. Balancing exploration and exploitation: A new algorithm for active machine learning. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*. IEEE, 8–pp.
[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine

Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[48] John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3 (1999), 61–74.

[49] Prahalad K Rao, Jia Peter Liu, David Roberson, Zhenyu James Kong, and Christopher Williams. 2015. Online real-time quality monitoring in additive manufacturing processes using heterogeneous sensors. *Journal of Manufacturing Science and Engineering* 137, 6 (2015).

[50] Thomas J Santner, Brian J Williams, William I Notz, and Brain J Williams. 2003. *The design and analysis of computer experiments*. Vol. 1. Springer.

[51] Burr Settles. 2012. Active learning. *Synthesis lectures on artificial intelligence and machine learning* 6, 1 (2012), 1–114.

[52] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*. 287–294.

[53] Boyang Shang and Daniel W Apley. 2021. Fully-sequential space-filling design algorithms for computer experiments. *Journal of Quality Technology* 53, 2 (2021), 173–196.

[54] Qiaomu Shen, Yanhong Wu, Yuzhe Jiang, Wei Zeng, KH Alexis, Anna Vianova, and Huamin Qu. 2020. Visual interpretation of recurrent neural network on multi-dimensional time-series forecast. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 61–70.

[55] Roger Solis, Arash Pakbin, Ali Akbari, Bobak J Mortazavi, and Roozbeh Jafari. 2019. A human-centered wearable sensing platform with intelligent automated data annotation capabilities. In *Proceedings of the International Conference on Internet of Things Design and Implementation*. 255–260.

[56] Erwin Stinstra, Dick den Hertog, Peter Stehouwer, and Arjen Vestjens. 2003. Constrained maximin designs for computer experiments. *Technometrics* 45, 4 (2003), 340–346.

[57] Eliza Strickland. 2022. *Andrew Ng: Unbiggen AI.* https://spectrum.ieee.org/andrew-ng-data-centric-ai [Accessed: 2022-02-18].

[58] Hongyue Sun, Xinwei Deng, Kaibo Wang, and Ran Jin. 2016. Logistic regression for crystal growth process modeling through hierarchical nonnegative garrote-based variable selection. *IIE Transactions* 48, 8 (2016), 787–796.

[59] Hongyue Sun, Kan Wang, Yifu Li, Chuck Zhang, and Ran Jin. 2017. Quality modeling of printed electronics in aerosol jet printing based on microscopic images. *Journal of Manufacturing Science and Engineering* 139, 7 (2017).

[60] Sebastian Thrun. 1995. Exploration in active learning. *Handbook of Brain Science and Neural Networks* (1995), 381–384.

[61] Sebastian Thrun and Lorien Pratt. 2012. *Learning to learn*. Springer Science & Business Media.

[62] Luís Torgo, Rita P Ribeiro, Bernhard Pfahringer, and Paula Branco. 2013. Smote for regression. In *Portuguese conference on artificial intelligence*. Springer, 378–389.

[63] Jan E Trost. 1986. Statistically nonrepresentative stratified sampling: A sampling technique for qualitative studies. *Qualitative sociology* 9, 1 (1986), 54–57.

[64] Huy Tu, Zhe Yu, and Tim Menzies. 2020. Better data labelling with emblem (and how that impacts defect prediction). *IEEE Transactions on Software Engineering* (2020).

[65] Junpeng Wang, Liang Gou, Wei Zhang, Hao Yang, and Han-Wei Shen. 2019. Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. *IEEE transactions on visualization and computer graphics* 25, 6 (2019), 2168–2180.

[66] JunPing Wang, YunChuan Sun, WenSheng Zhang, Ian Thomas, ShiHui Duan, and YouKang Shi. 2016. Large-scale online multitask learning and decision making for flexible manufacturing. *IEEE Transactions on Industrial Informatics* 12, 6 (2016), 2139–2147.

[67] Lening Wang, Xiaoyu Chen, Daniel Henkel, and Ran Jin. 2021. Pyramid Ensemble Convolutional Neural Network for Virtual Computed Tomography Image Prediction in a Selective Laser Melting Process. *Journal of Manufacturing Science and Engineering* 143, 12 (2021), 121003.

[68] Lening Wang, Pang Du, and Ran Jin. 2021. MOSS—Multi-Modal Best Subset Modeling in Smart Manufacturing. *Sensors* 21, 1 (2021), 243.

[69] Lening Wang, Yutong Zhang, Xiaoyu Chen, and Ran Jin. 2020. Online Computation Performance Analysis for Distributed Machine Learning Pipelines in Fog Manufacturing. In *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*. IEEE, 1628–1633.

[70] Meng Wang and Xian-Sheng Hua. 2011. Active learning in multimedia annotation and retrieval: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2, 2 (2011), 1–21.

[71] Sarah Wassermann, Thibaut Cuvelier, and Pedro Casas. 2019. RAL-Improving Stream-Based Active Learning by Reinforcement Learning. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) Workshop on Interactive Adaptive Learning (IAL)*.

[72] Tangbin Xia, Yifan Dong, Lei Xiao, Shichang Du, Ershun Pan, and Lifeng Xi. 2018. Recent advances in prognostics and health management for advanced manufacturing paradigms. *Reliability Engineering & System Safety* 178 (2018), 255–268.

[73] Hao Yan, Kamran Paynabar, and Jianjun Shi. 2020. AKM2D: An adaptive framework for online sensing and anomaly quantification. *IISE Transactions* 52, 9 (2020), 1032–1046.
[74] Yutong Zhang, Lening Wang, Xiaoyu Chen, and Ran Jin. 2019. Fog computing for distributed family learning in cyber-manufacturing modeling. In *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*. IEEE, 88–93.
[75] Xun Zhao, Weiwei Cui, Yanhong Wu, Haidong Zhang, Huamin Qu, and Dongmei Zhang. 2019. Oui! Outlier Interpretation on Multi-dimensional Data via Visual Analytics. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 213–224.
[76] Xiaojin Zhu and Andrew B Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning* 3, 1 (2009), 1–130.
[77] Martin Zinkevich. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*. 928–936.

Ensemble Active Learning by Contextual Bandits for AI Incubation in Manufacturing: Supplementary Materials

YINGYAN ZENG, Grado Department of Industrial and Systems Engineering, Virginia Tech, USA
XIAOYU CHEN, Department of Industrial Engineering, University of Louisville, USA
RAN JIN*, Grado Department of Industrial and Systems Engineering, Virginia Tech, Virginia, USA

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Yingyan Zeng, Xiaoyu Chen, and Ran Jin. 2021. Ensemble Active Learning by Contextual Bandits for AI Incubation in Manufacturing: Supplementary Materials. *ACM Trans. Intell. Syst. Technol.* 1, 1 (March 2021), 9 pages. https://doi.org/10.1145/1122445.1122456

1 JUSTIFICATION OF CHARACTERISTICS OF EXPLORATION AND EXPLOITATION AGENTS

To quantify the extent of exploration and exploitation of an acquisition criteria, the variance of the acquired samples by one agent is selected as a metric. A higher variance suggests a learner's ability to explore the input variable space, whereas a lower variance implies a high frequency of queries in a single region, implying exploitation. To illustrate the variance of the acquired samples, the expectation of the acquisition probability of an incoming sample is investigated to validate the properties of our proposed acquisition criteria.

Assume all streaming samples are generated from a mixture of a finite number of Gaussian distributions.

1.1 Exploration-oriented Agent

At first, we consider the acquisition decision made by an exploration-oriented agent, LD-Agent. Denote the previously observed samples stored in the sliding window $\mathcal{W}$ as the set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_L\}$, where $\boldsymbol{x}_i$ belongs to the i-th Gaussian distribution (i.e., $\boldsymbol{x}_i \sim \mathcal{N}_q(\boldsymbol{\mu}_i, \Sigma^{(i)})$).

Given a streaming sample $\boldsymbol{x}_t$ at time t which follows another Gaussian distribution (i.e., $\boldsymbol{x}_t \sim \mathcal{N}_q(\boldsymbol{\mu}_k, \Sigma^{(k)})$), the acquisition probability of the proposed low-density based exploration agent is:

$$p_t = \frac{lsf(\boldsymbol{x}_t))}{L\delta_L} = \frac{\sum_{i=1}^{L} \mathbb{I}\{\max d(\boldsymbol{x}_i, \boldsymbol{x}_j), j \neq i, j \in \{1, ..., L\} < d(\boldsymbol{x}_i, \boldsymbol{x}_t)\}}{L\delta_L}. \tag{1.1}$$

Authors' addresses: Yingyan Zeng, yingyanzeng@vt.edu, Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia, USA, 24061; Xiaoyu Chen, xiaoyu.chen@louisville.edu, Department of Industrial Engineering, University of Louisville, Kentucky, Louisville, USA, 40292; Ran Jin, jran5@vt.edu, Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia, USA, 24061.

2157-6904/2021/3-ART $15.00
https://doi.org/10.1145/1122445.1122456

With the distance $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2}$, the acquisition probability can be rewritten as:

$$p_t = \sum_{i=1}^{L} P\{\max_{j \neq i, j \in \{1,\ldots,L\}} \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 < \|\boldsymbol{x}_i - \boldsymbol{x}_t\|_2^2\}/(L \cdot \delta_L). \tag{1.2}$$

To evaluate the acquisition decision, we calculate the expectation of the acquisition probability:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k}[p_t] &= \mathbb{E}\left[\sum_{i=1}^{L} P\{\max_{j \neq i, j \in \{1,\ldots,L\}} \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 < \|\boldsymbol{x}_i - \boldsymbol{x}_t\|_2^2\}/(L \cdot \delta_L)\right] \\ &= \sum_{i=1}^{L} \mathbb{E}[P\{\max_{j \neq i, j \in \{1,\ldots,L\}} \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 < \|\boldsymbol{x}_i - \boldsymbol{x}_t\|_2^2\}]/(L \cdot \delta_L).\end{aligned} \tag{1.3}$$

Denote $X_{i,j}$ and $Y_{i,t}$ as the squared Euclidean distance between $\boldsymbol{x}_i, \boldsymbol{x}_j$ and $\boldsymbol{x}_i, \boldsymbol{x}_t$, (i.e., $X_{i,j} = \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 \in \mathbb{R}$ where $j \neq i, i, j \in \{1, \ldots, L\}$, and $Y_{i,t} = \|\boldsymbol{x}_i - \boldsymbol{x}_t\|_2^2 \in \mathbb{R}, i \in \{1, \ldots, L\}$). Let $Z_i = \max_{j \neq i, j \in \{1,\ldots,L\}} \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2$, $F_Z(z) = P(X_{i,1} < z, X_{i,2} < z, \ldots, X_{i,L} < z) = \Pi_{j=1}^{L} F_{X_{i,j}}(z)$.

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k}[p_t] &= \sum_{i=1}^{L} \mathbb{E}[\mathbb{E}[P\{Z_i < Y_{i,t}\}|Y_{i,t}]]/(L \cdot \delta_L) \\ &= \sum_{i=1}^{L} \mathbb{E}[F_Z(Y_{i,t})]/(L \cdot \delta_L) \\ &= \sum_{i=1}^{L} \int_0^{\infty} \left[\Pi_{j=1, j \neq i}^{L-1} F_{X_{i,j}}(y))\right] f_{Y_{i,t}}(y) dy/(L \cdot \delta_L),\end{aligned} \tag{1.4}$$

where $F_{X_{i,j}}(\cdot)$ is the cumulative density function for $X_{i,j}$ and $f_{Y_{i,t}}$ is the probability density function for $Y_{i,t}$.

Since the samples in $\mathcal{W}$ and the incoming sample $\boldsymbol{x}_t$ can be considered as independent draws from different Gaussian distributions, we have $\boldsymbol{x}_i - \boldsymbol{x}_j \sim \mathcal{N}(\boldsymbol{x}_i - \boldsymbol{x}_j, \Sigma^{(i)} + \Sigma^{(j)})$, and $\boldsymbol{x}_i - \boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_i - \boldsymbol{x}_t, \Sigma^{(i)} + \Sigma^{(k)})$. Therefore, $X_{i,j}$ and $Y_{i,t}$ are quadratic forms of random normal variables and follow a generalized chi-square distribution. The probability density function for $X_{i,j}$ is [3]:

$$f_{X_{i,j}=y} = \sum_{k=1}^{\infty} (-1)^k c_k \frac{y^{\frac{q}{2}+k-1}}{\Gamma(\frac{q}{2}+k)}, 0 \leq y \leq \infty. \tag{1.5}$$

The cumulative density function is:

$$F(X_{i,j} < y) = \sum_{k=0}^{\infty} (-1)^k c_k \frac{y^{\frac{q}{2}+k}}{\Gamma(\frac{q}{2}+k+1)}, 0 < y < \infty, \tag{1.6}$$

where c_0 and c_k are defined by:

$$H^T(\Sigma^{(i)} + \Sigma^{(j)})H = \text{diag}(\lambda_1, ..., \lambda_q) \tag{1.7}$$

$$H^T H = I, \tag{1.8}$$

$$b = H^T(\Sigma^{(i)} + \Sigma^{(j)})^{-\frac{1}{2}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{1.9}$$

$$c_0 = \exp\left(-\frac{1}{2}\sum_{j=1}^{q} b_j^2\right)\Pi_{j=1}^{q}(2\lambda_j)^{-\frac{1}{2}} \tag{1.10}$$

$$d_k = \frac{1}{2}\sum_{j=1}^{q}(1 - kb_j^2)(2\lambda_j)^{-k}, k \geq 1 \tag{1.11}$$

$$c_k = \frac{1}{k}\sum_{r=0}^{k-1} d_{k-r}c_r, k \geq 1 \tag{1.12}$$

For tractability of the computation, we assume the input variables in the Gaussian distributions to be independent (i.e., $\Sigma^{(i)} = \sigma_i^2 \boldsymbol{I}$), then the distance can be simplified and approximated by a normal distribution when q is large based on the central limit theorem:

$$X_{i,j} \sim \mathcal{N}_q\left(\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2 + (\sigma_i^2 + \sigma_j^2)q, 4(\sigma_i^2 + \sigma_j^2))\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2 + 2q(\sigma_i^2 + \sigma_j^2)\right) \tag{1.13}$$

$$Y_{i,t} \sim \mathcal{N}_q\left(\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2 + (\sigma_i^2 + \sigma_k^2)q, 4(\sigma_i^2 + \sigma_k^2))\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2 + 2q(\sigma_i^2 + \sigma_k^2)\right). \tag{1.14}$$

This gives us the expectation of acquisition probability as:

$$\mathbb{E}_{\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k}[p_t] = \sum_{i=1}^{L}\left\{\int_0^\infty \left[\Pi_{j=1,j\neq i}^{L-1}\Phi\left(\frac{y_i - \left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|_2^2 - (\sigma_i^2 + \sigma_j^2)q}{(\sqrt{4(\sigma_i^2 + \sigma_j^2)\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|_2^2 + 2q(\sigma_i^2 + \sigma_j^2)^2}}\right)\right]\cdot\right.$$
$$\left.\phi\left(\frac{y_i - \left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|_2^2 - (\sigma_i^2 + \sigma_k^2)q}{\sqrt{4(\sigma_i^2 + \sigma_k^2)\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|_2^2 + 2q(\sigma_i^2 + \sigma_k^2)^2}}\right)dy_i\right\}/(L \cdot \delta_L), \tag{1.15}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative density function and probability density function of the standard normal distribution.

Theorem 1. *If the streaming samples follow an independent multivariate Gaussian distribution (i.e., $\Sigma_i = \sigma_i^2 \boldsymbol{I}$)), then there exist $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^L$ such that if $\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2 > M_{2,i}\ \forall i \in \{1, ..., L\}$, the expectation acquisition probability of LD-Agent $\mathbb{E}_{\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k}[p_t] \geq 1$, where $\boldsymbol{M}_1, \boldsymbol{M}_2$ satisfies:*

$$M_{2,i} + \text{erf}^{-1}(1 - 2\delta_L) \cdot \sqrt{2 \cdot (4(\sigma_i^2 + \sigma_k^2))M_2 + 2q(\sigma_i^2 + \sigma_k^2))} + (\sigma_i^2 + \sigma_k^2)q - M_{1,i} = 0$$
$$M_{1,i} > \left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2 + (\sigma_i^2 + \sigma_j^2)q + \sqrt{4(\sigma_i^2 + \sigma_j^2))\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2 + 2q(\sigma_i^2 + \sigma_j^2)}$$
$$\cdot\left(\Phi^{-1}(1 - \frac{1}{L-1}) + \gamma\left[\Phi^{-1}(1 - \frac{1}{L-1} \cdot e^{-1}) - \Phi^{-1}(1 - \frac{1}{L-1})\right]\right),$$
$$\forall i \in \{1, ..., L\}. \tag{1.16}$$

where $\Phi(\cdot)$ is the cdf function for the standard normal distribution.

Proof. Based on (1.3), if

$$[1 - F_{Y_{i,t}}(\max_{j\neq i, j\in\{1,\ldots,L\}} \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2)]/\delta_L \geq 1 \; \forall i \in \{1, \ldots, L\},$$

the expectation of the acquisition probability

$$\mathbb{E}_{\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k}[p_t] = \sum_{i=1}^{L}[1 - F_{Y_{i,t}}(\max_{j\neq i, j\in\{1,\ldots,L\}} \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2)]/(\delta_L \cdot L) \geq 1.$$

Thus, to ensure $\mathbb{E}_{\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k}[p_t] \geq 1$, we have:

$$F_{Y_{i,t}}^{-1}(1-\delta_L) \geq \max_{j\neq i j\in\{1,\ldots,L\}} \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 = Z_i, \forall i \in \{1, \ldots, L\} \tag{1.17}$$

$$\frac{Z_i - [\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2 + (\sigma_i^2 + \sigma_k^2)q]}{\sqrt{2 \cdot (4(\sigma_i^2 + \sigma_k^2))\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2 + 2q(\sigma_i^2 + \sigma_k^2))}} \leq \mathrm{erf}^{-1}(1 - 2\delta_L)$$

$$\rightarrow g(\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2) = \left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2 + \mathrm{erf}^{-1}(1 - 2\delta_L) \cdot \sqrt{2} \cdot$$

$$(4(\sigma_i^2 + \sigma_k^2))\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2 + 2q(\sigma_i^2 + \sigma_k^2)) + (\sigma_i^2 + \sigma_k^2)q - Z_i \geq 0. \tag{1.18}$$

By Fisher–Tippett–Gnedenko theorem [2] and the independent assumption on the pairwise distances, Z_i can be approximated by generalized extreme value (GEV) distribution where the expectation of Z_i can be estimated with:

$$\mathbb{E}[Z_i] \approx \left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2 + (\sigma_i^2 + \sigma_j^2)q + \sqrt{4(\sigma_i^2 + \sigma_j^2))\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2 + 2q(\sigma_i^2 + \sigma_j^2)}$$
$$\cdot \left(\Phi^{-1}(1 - \frac{1}{L-1}) + \gamma\left[\Phi^{-1}(1 - \frac{1}{L-1} \cdot e^{-1}) - \Phi^{-1}(1 - \frac{1}{L-1})\right]\right),$$
$$\forall i \in \{1, \ldots, L\}, \tag{1.19}$$

where γ is the Euler–Mascheroni constant. Accordingly, the variance is:

$$\mathbb{V}[Z_i] \approx \sqrt{4(\sigma_i^2 + \sigma_j^2))\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2 + 2q(\sigma_i^2 + \sigma_j^2)} \cdot$$
$$\left[\Phi^{-1}(1 - \frac{1}{L-1} \cdot e^{-1}) - \Phi^{-1}(1 - \frac{1}{L-1})\right], \forall i \in \{1, \ldots, L\}. \tag{1.20}$$

By Chebyshev's inequality, we have:

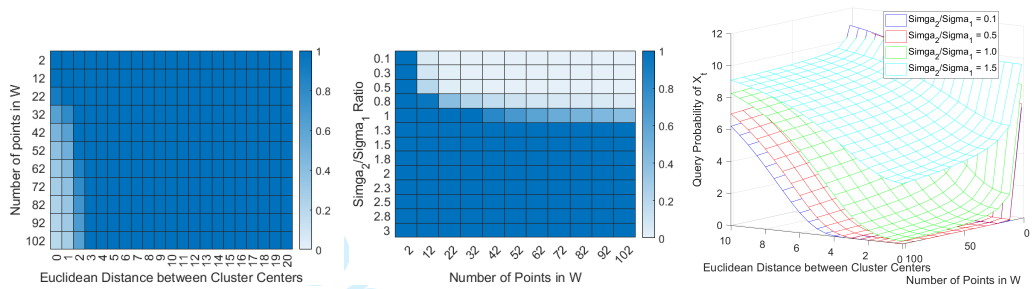
$$P(|Z_{i,n} - \mathbb{E}[Z_i]| \geq \epsilon) \leq \frac{\mathbb{V}[Z_i]}{n\epsilon^2}. \tag{1.21}$$

Therefore, there exist $M_{1,i} > \mathbb{E}[Z_i]$ such that $Z_{i,n} < M_{1,i}, \forall n \in \mathbb{Z}$. Since $0 < \delta_L < 1$, $\mathrm{erf}^{-1}(1 - 2\delta_L) > 0$, $g(\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2)$ is monotonic with $\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2$. Hence, there exist $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^L$ such that if $\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2 > M_{2,i} \; \forall i \in \{1, \ldots, L\}$, $\mathbb{E}_{\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k}[p_t] \geq 1$, where $\boldsymbol{M}_1, \boldsymbol{M}_2$ satisfies:

$$M_{2,i} + \mathrm{erf}^{-1}(1 - 2\delta_L) \cdot \sqrt{2 \cdot (4(\sigma_i^2 + \sigma_k^2))M_2 + 2q(\sigma_i^2 + \sigma_k^2))} + (\sigma_i^2 + \sigma_k^2)q - M_{1,i} = 0$$
$$M_{1,i} > \mathbb{E}[Z_i], \forall i \in \{1, \ldots, L\}. \tag{1.22}$$

□

Numerical analysis is further investigated for the LD-Agent. To be more intuitive, we set the samples in $\mathcal{W}$ all belonging to the same Gaussian distribution while $\boldsymbol{x}_t$ belongs to another (i.e., $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j = \boldsymbol{\mu}_1$, $\sigma_i = \sigma_j = \sigma_1$, $\sigma_k = \sigma_2$). Set $q = 15$ and $\delta_L = 0.05$, the expectation of the acquisition probability calculated by (1.15) is shown as follows:



(a) Set $\sigma_1 = \sigma_2 = 1$, change number of points in $\mathcal{W}$ and $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$.

(b) Set $\sigma_1 = 1$, $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2 = 4$, change number of points in $\mathcal{W}$ and σ_2/σ_1 ratio.

(c) Set $\sigma_1 = 1$, change number of points in $\mathcal{W}$ and $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$.

Fig. 1. Expectation of the probability of acquiring $\boldsymbol{x}_t$ by LD-Agent with $\sigma_1 = 1, q = 15$ with different $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2$ and σ_2/σ_1 ratio.

As shown in Figure 1, with a larger distance between two centers of the clusters, with less samples in the sliding window $\mathcal{W}$ and with higher $\frac{\sigma_2}{\sigma_1}$ ratio, the expectation of the acquisition probability will increase and approach to 1. This ensures the acquisition decision of samples from a remote cluster, resulting in an increased variance and thus, exploration of the input space.

1.2 Exploitation-oriented Agent

Assume the logistic regression is selected as the base learner, and at time t it is parameterized by $\boldsymbol{\beta}_t$. Given the labelled data pool $\mathcal{D}_t = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_{n_t}, y_{n_t})\}$ at time t, we have:

$$P(y_i = 1) = \frac{1}{1 + \exp(-\boldsymbol{x}_i^T \boldsymbol{\beta}_i)}, P(y_i = 0) = \frac{\exp(-\boldsymbol{x}_i^T \boldsymbol{\beta}_i)}{1 + \exp(-\boldsymbol{x}_i^T \boldsymbol{\beta}_i)}$$

$$\boldsymbol{\beta}_t = \operatorname{argmax} P(D_t|\boldsymbol{\beta}) = \operatorname{argmax} \Pi_{i=1}^{n_t} P(\boldsymbol{x}_i, y_i|\boldsymbol{\beta}) = \operatorname{argmax} \Pi_{i=1}^{n_t} P(y_i = 0)^{y_i} \cdot P(y_i = 1)^{1-y_i}$$

The acquisition probability p_t by the RAL-Agent for $\boldsymbol{x}_t$ is:

$$\begin{aligned} p_t &= \mathbb{I}\left\{ \max\left(P(\hat{y} = 1), P(\hat{y} = 0)\right) < \theta_t \right\} \\ &= \mathbb{I}\left\{ \max\left(\frac{1}{1 + \exp(-\boldsymbol{x}_t^T \boldsymbol{\beta}_t)}, \frac{\exp(-\boldsymbol{x}_i^T \boldsymbol{\beta}_i)}{1 + \exp(-\boldsymbol{x}_i^T \boldsymbol{\beta}_i)}\right) < \theta_t \right\} \\ &= \mathbb{I}\left(\ln \frac{1 - \theta_t}{\theta_t} < \boldsymbol{x}_t^T \boldsymbol{\beta}_t < \ln \frac{\theta_t}{1 - \theta_t} \right). \end{aligned} \tag{1.23}$$

The expectation of the acquisition probability p_t by the RAL-Agent given the labelled data pool $\mathcal{D}_t$ at time t is:

$$\mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{\beta}_t}[p_t] = F_{\boldsymbol{x}_t^T \boldsymbol{\beta}_t}(\ln \frac{\theta_t}{1 - \theta_t}) - F_{\boldsymbol{x}_t^T \boldsymbol{\beta}_t}(\ln \frac{1 - \theta_t}{\theta_t}), \tag{1.24}$$

where $F_{\boldsymbol{x}_t^T \boldsymbol{\beta}_t}(\cdot)$ is the cumulative density function of the random variable $\boldsymbol{x}_t^T \boldsymbol{\beta}_t$, θ_t is the certainty threshold at time t with predefined range $(0.5, 1]$. If $\theta_t \in (0, 0.5)$, the derivation and the proof will only be different by changing the lower and upper integral bounds as:

$$\mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{\beta}_t}[p_t] = F_{\boldsymbol{x}_t^T \boldsymbol{\beta}_t}(\ln \frac{1-\theta_t}{\theta_t}) - F_{\boldsymbol{x}_t^T \boldsymbol{\beta}_t}(\ln \frac{\theta_t}{1-\theta_t}), \tag{1.25}$$

Based on the asymptotic normality of the maximum likelihood estimation (MLE), we have:

$$\sqrt{n_t}(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}_q(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\beta})), \tag{1.26}$$

$$\boldsymbol{I}(\boldsymbol{\beta}) = -\mathbb{E}(s'(\boldsymbol{\beta}|\boldsymbol{D}_t))) = -H(\boldsymbol{\beta}|\boldsymbol{D}_t) = -\sum_{i=1}^{n_t} \boldsymbol{x}_i \boldsymbol{x}_i^T \cdot \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{(1+\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}))^2} \tag{1.27}$$

$$\hat{\boldsymbol{\beta}}_t \sim \mathcal{N}_q\left(\boldsymbol{\beta}, -\left[n_t \sum_{i=1}^{n_t} \boldsymbol{x}_i \boldsymbol{x}_i^T \cdot \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{(1+\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}))^2}\right]^{-1}\right), \tag{1.28}$$

where $\boldsymbol{\beta}$ is the unknown ground-truth parameter.

Denote $H = \boldsymbol{x}_t^T \boldsymbol{\beta}_t \in R$. Based on [1], the exact PDF of the product $H_i = \boldsymbol{x}_{t,i}^T \beta_{t,i}$ is given by:

$$\begin{aligned} f_{H_i}(h) = \exp\Bigg\{ & -\frac{1}{2(1-\rho^2)}\left(\frac{\mu_{k,i}^2}{\Sigma_{i,i}^{(k)}} + \frac{\beta_i^2}{\Sigma_{i,i}^{(\beta)}} - \frac{2\rho(h+\mu_{k,i}\beta_i)}{(\Sigma_{i,i}^{(k)}\Sigma_{i,i}^{(\beta)})^{1/2}}\right) \\ & \times \sum_{n=0}^{\infty}\sum_{m=0}^{2n} \frac{h^{2n-m}|h|^{m-n}\Sigma_{i,i}^{(k)(m-n-1)/2}}{\pi(2n)!(1-\rho^2)^{2n+1/2}\cdot\Sigma_{i,i}^{(\beta)(m-n+1)/2}} \binom{2n}{m}\left(\frac{\mu_{k,i}}{\Sigma_i^{(k)}} - \frac{\rho\beta_i}{(\Sigma_{i,i}^{(k)}\Sigma_{i,i}^{(\beta)})^{1/2}}\right)^m \\ & \times \left(\frac{\beta_i}{\Sigma_{i,i}^{(\beta)}} - \frac{\rho\mu_{k,i}}{(\Sigma_{i,i}^{(k)}\Sigma_{i,i}^{(\beta)})^{1/2}}\right)^{2n-m} K_{m-n}\left(\frac{|h|}{(1-\rho^2)(\Sigma_{i,i}^{(k)}\Sigma_{i,i}^{(\beta)})^{1/2}}\right)\Bigg\}, i = 1, ..., q, \end{aligned} \tag{1.29}$$

where ρ is the correlation coefficient

$$\rho = \frac{\mathbb{E}[(x_{t,i} - \mu_{k,i})(\beta_{t,i} - \beta_i)]}{\sqrt{\Sigma_{i,i\Sigma_{i,i}^{(\beta)})}^{(k)}}}, \forall i \in \{1, ..., L\}, \tag{1.30}$$

and $\Sigma_{i,j}^{(k)}$ is the $(i,j)-$th entry in the covariance matrix Σ_k. The summation of the p products leads to non-closed form derivation. If we consider $\boldsymbol{\beta}_t$ as an estimation based on $\mathcal{D}_t$, then H follows a normal distribution:

$$\mu_H = \mathbb{E}[H] = \sum_{i=1}^{q} \beta_q \mu_{k,i} \tag{1.31}$$

$$\sigma_H^2 = \mathbb{V}[H] = \sum_{i=1}^{q} \beta_i^2 \cdot \Sigma_{i,i}^{(k)} + 2\sum_{i=1}^{q}\sum_{j:j>i}^{q} \beta_i \beta_j \Sigma_{i,j}^{(k)}. \tag{1.32}$$

Therefore,

$$\mathbb{E}_{\boldsymbol{x}_t}[p_t] = \Phi\left(\frac{\ln\frac{\theta_t}{1-\theta_t} - \mu_H}{\sigma_H}\right) - \Phi\left(\frac{\ln\frac{1-\theta_t}{\theta_t} - \mu_H}{\sigma_H}\right) = \int_{-\ln\frac{\theta_t}{1-\theta_t}}^{\ln\frac{\theta_t}{1-\theta_t}} \frac{1}{\sigma_H\sqrt{2\pi}} \exp\left(-\frac{1}{2}\cdot\frac{h-\mu_H}{\sigma_H}\right)^2 dh \tag{1.33}$$

Theorem 2. *Given the labelled data pool $\mathcal{D}_t$ at time t, assume the center of the labelled samples in $\mathcal{D}_t$ is $\boldsymbol{\mu}_i \in \mathbb{R}^q$ and the incoming sample $\boldsymbol{x}_t \sim \mathcal{N}_q(\boldsymbol{\mu}_k, \sigma_k^2 \boldsymbol{I})$. With the increase of the distance between two centers $\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2$, there does not exist $M_3 \in \mathbb{R}$ such that $P\{|\mathbb{E}_{\boldsymbol{x}_t}[p_t] - M_3| \geq \epsilon\} = 0, \forall \epsilon \in \mathbb{R}$.*

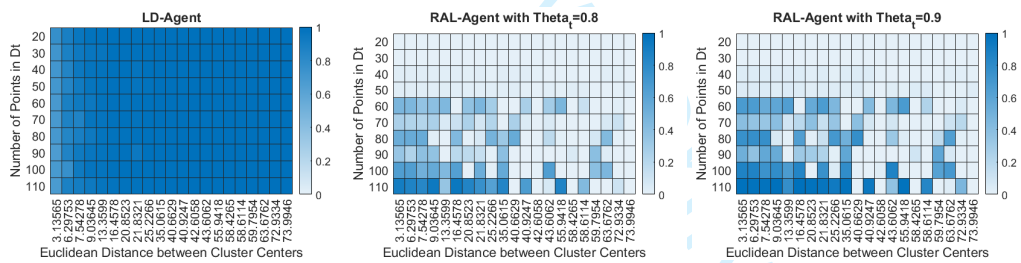
Proof. Given $\mathcal{D}_t$, σ_k, check the convergence property of $\mathbb{E}_{\boldsymbol{x}_t}[p_t]$ with the increase of the distance between the center of observed samples and the incoming sample (i.e., $\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\right\|^2$).

Since $\mathbb{E}_{\boldsymbol{x}_t}[p_t] \in [0, 1]$, assume there exists $M_3 \in [0, 1]$ such that $P\{|\mathbb{E}_{\boldsymbol{x}_t}[p_t] - M_3| \geq \epsilon\} = 0, \forall \epsilon \in \mathbb{R}$. To bound the integral (i.e., $P\{|\int_{-\ln\frac{\theta_t}{1-\theta_t}}^{\ln\frac{\theta_t}{1-\theta_t}} \frac{1}{\sigma_H\sqrt{2\pi}} \exp\left(-\frac{1}{2}\cdot\frac{h-\mu_H}{\sigma_H}\right)^2 dh - M_3| \geq \epsilon = 0\}$), there exists $M_4 \in \mathbb{R}$ such that $|\mu_H| < M_4$. As shown in Eq (1.26), $\boldsymbol{\beta}_t$ is bounded by the asymptotic normal distribution. However, since $\left\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_k\right\|^2$ increases, there does not exits $M_5 \in \mathbb{R}$ such that $\left\|\boldsymbol{\mu}_k\right\| \leq M_5$, which indicates $|\mu_H| = \sum_{i=1}^q \beta_q \mu_{k,i}$ can not be bounded, causing the contradiction.

Therefore, the expected acquisition probability $\mathbb{E}[p_t]$ of $\boldsymbol{x}_t$ by the reinforced agent does not converge with increasing distance between the center of the observed distribution and the incoming distribution. Hence, the acquisition decision made by RAL-Agent does not necessarily lead to an increasing variance when it decides not to acquire samples from a remote cluster because the decision is determined by the base learner's uncertainty instead of the distance. □

1.3 Numerical Comparison

For further comparison between exploration-oriented agents and exploitation oriented agents, we assume $\mathcal{W} = \mathcal{D}_t$ at time t, and the samples in $\mathcal{W}$ and $\mathcal{D}_t$ all belong to the same Gaussian distribution (i.e., $\boldsymbol{x}_1 \sim \mathcal{N}_q(\boldsymbol{\mu}_1, \sigma_1^2 \boldsymbol{I})$). Set $\sigma_1 = \sigma_2 = 1, q = 15, \delta_L = 0.5$, we have the following result:



(a) Set $\sigma_1 = \sigma_2 = 1$, change number of points in $\mathcal{W}$ and $\left\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right\|^2$.

(b) Set $\sigma_1 = 1, \left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|_2^2 = 4$, change number of points in $\mathcal{W}$ and σ_2/σ_1 ratio.

(c) Set $\sigma_1 = 1$, change number of points in $\mathcal{W}$ and $\left\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right\|^2$.

Fig. 2. Expectation of the probability of acquiring $\boldsymbol{x}_t$ of the exploration-oriented agent and exploitation-oriented agent with $\sigma_1 = \sigma_2 = 1, q = 15$, vertical axis as $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$, and horizontal axis as the window size L

With the derivation and numerical results, it can be illustrated that the proposed exploration-oriented agent is more likely to acquire a sample from a distinct distribution, thus increasing the variance of the samples in the labelled data pool. On the other hand, the exploitation-oriented agent focuses on samples close to the estimated decision boundary, which provides a relative smaller variance. The theoretical analysis and numerical results justify the exploration and exploitation capabilities of the proposed agents and motivates us to ensemble these two agents to dynamically adjust the trade-off between the exploration and exploitation in the online updating process.

2 NUMBER OF ACQUIRED SAMPLES

We show the result of the average number of the acquired samples by the proposed method and the candidate agents in the simulation study in Table 1.

Table 1. The average values and standard errors (in parenthesis) of number of the acquired samples. The smallest numbers are highlighted in **bold**.

Level		Method	Sparsity = 30%			Sparsity = 70%		
			Size of Data Stream			Size of Data Stream		
Disturbance	Percentage of Positive Samples		500	1000	1500	500	1000	1500
0%	10%	Opt Exploration	48.00 (0.00)	97.80 (0.19)	148.00 (0.00)	48.00 (0.00)	98.00 (0.00)	148.00 (0.00)
		Opt Exploitation	**37.20 (3.30)**	**81.00 (3.18)**	**108.00 (4.06)**	**44.00 (1.05)**	**79.60 (6.47)**	**114.70 (2.55)**
		CBEAL-2	45.60 (2.28)	91.90 (2.18)	119.20 (9.57)	47.60 (0.38)	98.00 (0.00)	123.20 (9.53)
		CBEAL-6	45.80 (1.42)	89.20 (3.25)	132.30 (2.42)	48.00 (0.00)	88.20 (4.31)	124.00 (5.78)
	5%	Opt Exploration	48.00 (0.00)	98.00 (0.00)	147.40 (0.57)	48.00 (0.00)	98.00 (0.00)	147.00 (0.95)
		Opt Exploitation	**30.90 (2.68)**	**58.00 (3.14)**	**71.80 (5.26)**	**37.50 (2.38)**	**59.20 (3.55)**	**80.10 (3.51)**
		CBEAL-2	40.30 (2.83)	68.70 (2.48)	78.20 (6.11)	43.60 (1.86)	70.60 (3.80)	83.70 (5.56)
		CBEAL-6	44.10 (1.78)	68.10 (2.73)	83.40 (2.78)	43.60 (1.02)	68.70 (4.35)	83.20 (4.19)
3%	10%	Opt Exploration	47.40 (0.57)	97.70 (0.29)	148.00 (0.00)	48.00 (0.00)	97.80 (0.19)	148.00 (0.00)
		Opt Exploitation	**41.50 (3.28)**	**71.90 (3.86)**	**109.10 (6.25)**	44.80 (0.99)	**81.40 (6.74)**	**102.50 (9.60)**
		CBEAL-2	48.00 (0.00)	90.70 (0.95)	120.60 (9.60)	48.00 (0.00)	97.00 (0.63)	127.60 (5.84)
		CBEAL-6	47.20 (0.76)	87.80 (6.37)	136.90 (3.50)	**43.70 (1.91)**	92.40 (3.30)	129.20 (6.65)
	5%	Opt Exploration	47.70 (0.20)	97.00 (0.95)	148.00 (0.00)	47.70 (0.20)	98.00 (0.00)	148.00 (0.00)
		Opt Exploitation	**30.10 (1.78)**	**51.80 (4.29)**	84.10 (2.72)	**30.10 (1.78)**	**56.90 (4.48)**	**86.20 (4.10)**
		CBEAL-2	40.90 (1.57)	64.60 (4.18)	89.80 (6.87)	40.90 (1.57)	73.10 (3.97)	101.60 (2.91)
		CBEAL-6	38.70 (2.87)	61.60 (4.10)	**83.00 (5.99)**	38.70 (2.81)	65.60 (4.52)	87.60 (4.56)

It is shown that the exploitation-oriented agent tends to acquire less samples under most of the scenarios whereas CbeAL (i.e., CbeAL-2 and CbeAL-6) lie in the middle of the two incorporated agents. This is reasonable since the ensemble method encode the exploration behabviour during the process, which will lead to a larger number of queries compared to pure exploitation.

3 ASSUMPTION ON INPUT DATA DISTRIBUTION

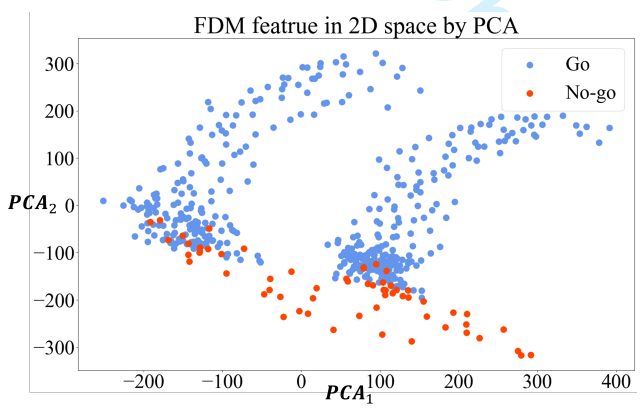


Fig. 3. Distribution of the first 500 samples with reduced dimensions by PCA in case study

To validate the third assumption made on the input data distribution, the input $\boldsymbol{X}$ of the first 500 samples in the FDM process is reduced to the two-dimensional space by principal component analysis (PCA) and visualized in figure 3. It can be clearly observed that there exist multiple clusters

in the class of good samples, which validates the assumption on the online data stream in the case study.

REFERENCES

[1] Guolong Cui, Xianxiang Yu, Salvatore Iommelli, and Lingjiang Kong. 2016. Exact distribution for the product of two correlated Gaussian random variables. *IEEE Signal Processing Letters* 23, 11 (2016), 1662–1666.
[2] Maurice Fréchet. 1927. Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Math. Polon.* 6 (1927), 93–116.
[3] Arakaparampil M Mathai and Serge B Provost. 1992. *Quadratic forms in random variables: theory and applications.* Dekker.