Data Acquisition for Improving Model Confidence

YIFAN LI, York University, Canada
XIAOHUI YU, York University, Canada
NICK KOUDAS, University of Toronto, Canada

In recent years, there has been a growing recognition that high-quality training data is crucial for the performance of machine learning models. This awareness has catalyzed both research endeavors and industrial initiatives dedicated to data acquisition to enhance diverse dimensions of model performance. Among these dimensions, model confidence holds paramount importance; however, it has often been overlooked in prior investigations into data acquisition methodologies. To address this gap, our work focuses on improving the data acquisition process with the goal of enhancing the confidence of Machine Learning models. Specifically, we operate within a practical context where limited samples can be obtained from a large data pool. We employ well-established model confidence metrics as our foundation, and we propose two methodologies, Bulk Acquisition (BA) and Sequential Acquisition (SA), each geared towards identifying the sets of samples that yield the most substantial gains in model confidence. Recognizing the complexity of BA and SA, we introduce two efficient approximate methods, namely *k*NN-BA and *k*NN-SA, restricting data acquisition to promising subsets within the data pool. To broaden the applicability of our solutions, we introduce a Distribution-based Acquisition approach that makes minimal assumption regarding the data pool and facilitates the data acquisition across various settings. Through extensive experimentation encompassing diverse datasets, models, and parameter configurations, we demonstrate the efficacy of our proposed methods across a range of tasks. Comparative experiments with alternative applicable baselines underscore the superior performance of our proposed approaches.

CCS Concepts: • **Computing methodologies** → **Machine learning**.

Additional Key Words and Phrases: query planning, cost estimation, conformal prediction

ACM Reference Format:
Yifan Li, Xiaohui Yu, and Nick Koudas. 2024. Data Acquisition for Improving Model Confidence. *Proc. ACM Manag. Data* 2, 3 (SIGMOD), Article 131 (June 2024), 25 pages. https://doi.org/10.1145/3654934

1 INTRODUCTION

The rapid development of Machine Learning (ML) techniques has brought about impressive breakthroughs, unleashing the potential of Artificial Intelligence (AI) to tackle complex problems and improve decision-making processes. However, a recurring theme that persists in the field is the insatiable hunger for data. As ML algorithms become more sophisticated and capable of handling intricate tasks, they demand an ever-increasing amount of diverse and high-quality data for training. This phenomenon has led to the popular adage that "data is the new oil" in the AI landscape.

Recognizing the significance of data acquisition for ML advancements, researchers and practitioners have actively pursued innovative approaches to collect and curate data efficiently. Recent years have witnessed substantial research interests focused on various aspects of data acquisition

Authors' addresses: Yifan Li, York University, Canada, yifanli@yorku.ca; Xiaohui Yu, York University, Canada, xhyu@yorku.ca; Nick Koudas, University of Toronto, Canada, koudas@cs.toronto.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published at https://doi.org/10.1145/3654934.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 2836-6573/2024/6-ART131
https://doi.org/10.1145/3654934

tasks, including active learning [33, 42], data market [4, 7, 28], data augmentation [11, 41], etc. Simultaneously, various industrial platforms have risen to facilitate the acquisition and sharing of datasets to fuel the growth of AI applications, such as Dawex [14], WorldQuant [45], and Xignite [46], to name a few.

In the area of data acquisition for ML, most existing efforts have centered around optimizing model performance. An equally vital aspect that has often been overlooked is model confidence [12, 22]. Model confidence refers to the ability of an ML model to express a measure of uncertainty or reliability in its outputs, and it has shown to be important in medical diagnosis, autonomous driving, and other tasks. Recognizing this crucial aspect, in this work we study the task of model confidence-oriented data acquisition and devise effective data acquisition strategies for the task. Previous works have established various ways to quantify the confidence for predictions of arbitrary models, which are generally based on some notion of distance between samples in an evaluation set and samples in the training set. We adopt established methods to quantify confidence [12] in our work, yet the approaches developed herein can be naturally applied to other measures of model confidence, as applicable.

In this paper, we consider a *model owner* whose objective is to improve the confidence of a model $\mathcal{M}$ by acquiring more samples to enrich the training set $\mathcal{T}$, and a *data pool* from which new samples can be acquired. The data pool can be a set of unlabeled samples and the annotation of each sample comes at a cost [34, 43] or a data market which allows users to purchase new data [9, 28]. Following previous work [8, 28], we assume that the model owner has a budget B denoting the maximum number of samples to acquire. The task is referred to as *Confidence-Oriented Data Acquisition*. Considering the well-studied variance-bias trade-off in the ML literature [24, 47], in this work we assume that the data pool and $\mathcal{T}$ follow the same distribution and focus on the high variance scenario. In this target setting, acquiring additional data is known to improve the model accuracy [5, 18, 21], and thus the data acquisition process oriented towards confidence improvement does not sacrifice model accuracy. We also empirically establish in Section 6.8 the positive correlation between confidence-oriented data acquisition and model accuracy in this case. Acquiring out-of-distribution samples [27, 32], or explicitly requiring the improvement of both confidence and accuracy of a model in a high bias scenario, are out of the scope of this paper; they form interesting multivariate optimization problems to explore as future work.

The major challenge to derive the optimal solution to the data acquisition task defined above comes from the mutual dependency between samples to be acquired. While it is straightforward to calculate the **C**onfidence **I**m**P**rovement of $\mathcal{M}$ resulting from each sample $s \in \mathcal{D}$ (denoted by CIP below), by definition, the CIP values of samples would change as new samples are acquired and added to $\mathcal{T}$. As will be demonstrated in Section 3.1, identifying the B samples in $\mathcal{D}$ that lead to the highest confidence improvement is NP-hard.

However, for certain scenarios, which we investigate, we are able to derive the optimal solution in polynomial time. More specifically, we consider two special cases, *Top-B Independence* and *Progressive Dominance*. Top-B independence describes the case when the B samples with the highest CIP values are independent from each other and acquiring any of them does not influence the CIP of the remaining samples. Progressive Dominance denotes the case when samples are acquired in rounds and there exists a sample that is guaranteed to be in the optimal solution in each round. We discuss how to observe and leverage the two cases in Section 3, together with the corresponding algorithms, *Bulk Acquisition* (BA) and *Sequential Acquisition* (SA) to derive the optimal solutions.

While BA and SA lead to the optimal solution in certain cases and aid the understanding of this problem, they have high complexity. To further improve the acquisition efficiency, we consider an important property in this problem: the effect to model confidence of a particular sample $e \in \mathcal{E}$ with a specific label depends on the distance between e and its nearest neighbors in $\mathcal{T}$ with the same

label and to the distance between e and its nearest neighbors in $\mathcal{T}$ with different label. Therefore, we propose two approximate variants of BA and SA, named kNN-BA and kNN-SA. For each $e \in \mathcal{E}$, only the kNN of e in $\mathcal{D}$ are retained, forming a candidate set. Algorithms kNN-BA and kNN-SA acquire new samples from the candidate set instead of $\mathcal{D}$. Although kNN-BA and kNN-SA do not lead to the optimal solution, as analyzed in Section 4, *their accuracy can be bounded*, providing theoretical guarantees regarding the model confidence improvement when they are utilized.

The data acquisition methods introduced above rely on the assumption that the feature vectors (coordinates) of $\mathcal{D}$ are available to the model owner. While the assumption is true for certain tasks [34, 43], there exist cases when the coordinates are hidden from the model owner [9, 28]. To facilitate effective data acquisition in generic cases, we devise *Distribution-based Acquisition* (DA), a two-phase approach that works in the absence of sample coordinates. In the first phase, samples from the same data space as $\mathcal{T}$ and $\mathcal{E}$ are generated (not necessarily belonging to $\mathcal{T}$ or $\mathcal{E}$), and their CIP values are computed. We then train a regression model F on the samples which maps each sample to its corresponding CIP. In the second phase, model F and budget B are passed to the data pool $\mathcal{D}$ (e.g., the owner of $\mathcal{D}$, a data broker, etc.), and F is used to estimate the CIP of each sample in $\mathcal{D}$ and return the B samples with the highest CIP values to the model owner. Therefore, DA avoids unnecessarily revealing the coordinates of $\mathcal{D}$ to the model owner.

We evaluate the performance of the proposed methods on both traditional ML tasks and Deep Learning tasks, including image classification and radar data classification, using classical ML models as well as state-of-the-art deep models. As will be shown in Section 6, the proposed methods demonstrate solid performance across a variety of settings, outperforming applicable baseline approaches. We also thoroughly compare the proposed methods analyzing the most suitable application scenario of each.

Our main contributions can be summarized below:

- We formally define and study the important problem of acquiring data to improve ML model confidence. The work does not only benefit the advancement of ML development, but also complement data acquisition tasks and related techniques.
- We study in depth various aspects of model confidence-oriented data acquisition, analyze its complexity, and propose two methods, BA and SA that could derive the optimal solution when certain assumptions are satisfied.
- We develop two lightweight acquisition strategies based on the properties of model confidence metrics, named kNN-BA and kNN-SA, which are efficient and approximate variants of BA and SA, with approximation guarantees.
- We devise a learned method which can estimate the CIP values of arbitrary samples, facilitating data acquisition without revealing the coordinates of the samples in the data pool, and thus the method can be applied in general settings.
- We thoroughly evaluate the performance of the proposed methods across a variety of settings and showcase their advantages compared with applicable baselines.

2 PRELIMINARIES

In this section we introduce the notations and terminologies that are used in subsequent sections, and formally define the problem studied in this work.

Data Domain. We consider a classification task on data domain $\Gamma = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{X}$ denotes the *feature* space and $\mathcal{Y}$ denotes the *label* space. The methods proposed in the paper can be applied to any models built on domain Γ.

Model Owner and Model Confidence. The model owner trains a Machine Learning model $\mathcal{M}$ on a training dataset $\mathcal{T} \subset \Gamma$, and evaluates the *model confidence* of $\mathcal{M}$ on an evaluation set $\mathcal{E} \subset \Gamma$. In this paper to facilitate presentation we compute confidence using [12]:

$$\text{conf}(\mathcal{M}, \mathcal{T}, \mathcal{E}) = \sum_{e \in \mathcal{E}} \frac{\bar{G}(e, \mathcal{T}) - G(e, \mathcal{T})}{2} \tag{1}$$

where $\bar{G}(e, \mathcal{T})$ and $G(e, \mathcal{T})$ are defined as follows:

$$\begin{aligned} \bar{G}(e, \mathcal{T}) &= \min\{d(e, x) | \forall x \in \mathcal{T} \wedge \mathcal{M}(e) \neq \mathcal{M}(x)\} \\ G(e, \mathcal{T}) &= \min\{d(e, x) | \forall x \in \mathcal{T} \wedge \mathcal{M}(e) = \mathcal{M}(x)\} \end{aligned} \tag{2}$$

where $d(e, x)$ denotes the Euclidean distance between e and x, and $\mathcal{M}(e)$ denotes the predicted label of e. We note that distance-based confidence metrics are a standard way to quantify model confidence [12, 22, 44], and the subsequent analysis in this paper and the techniques thus developed naturally apply to other distance-based confidence metrics as well.

From Equation (2) we know that $\bar{G}(e, \mathcal{T})$ is the minimal distance between e and samples in $\mathcal{T}$ with labels other than $\mathcal{M}(e)$, and $G(e, \mathcal{T})$ is the minimal distance between e and samples in $\mathcal{T}$ with label $\mathcal{M}(e)$. Note that only predicated labels (i.e., $\mathcal{M}(e)$) are used when calculating the model confidence, rather than true labels.

We also note that model confidence is additive, i.e., the confidence of $\mathcal{M}$ regarding $\mathcal{E}$ is the summation of the confidence of $\mathcal{M}$ regarding each sample in $\mathcal{E}$. In other words, we can also calculate $\text{conf}(\mathcal{M}, \mathcal{T}, \mathcal{E})$ as follows:

$$\text{conf}(\mathcal{M}, \mathcal{T}, \mathcal{E}) = \sum_{e \in \mathcal{E}} \text{conf}(\mathcal{M}, \mathcal{T}, \{e\}) \tag{3}$$

Data pool and Data Acquisition. Data pool is a dataset $\mathcal{D} \subset \Gamma$, from which the model owner can acquire samples to enrich $\mathcal{T}$, during *Data Acquisition* [25, 49]. Acquiring samples from a data pool is associated with a cost (referred to as data pricing); data pricing is orthogonal to the problem studied herein and we simplify cost calculation assuming that the model owner can acquire at most B samples from the data pool, which is denoted as the budget. Following previous studies [40, 43], we focus on a setting where the feature space $\mathcal{X}$ is exposed to the model owner while the label space $\mathcal{Y}$ is hidden, and thus acquiring a sample is essentially revealing the label of the selected sample. The data acquisition process may consist of one or more rounds, and in each round the model owner acquires one or more samples from $\mathcal{D}$. At the end of each round, the model owner retrains $\mathcal{M}$ with all of the possessed samples (i.e., $\mathcal{T}$ and samples acquired from $\mathcal{D}$), and future acquisition is guided by the current predictions of the model.

Confidence Improvement. The objective of the model owner is to improve the confidence of $\mathcal{M}$ (i.e., $\text{conf}(\mathcal{M}, \mathcal{T}, \mathcal{E})$). Each sample added to $\mathcal{T}$ may change the confidence of $\mathcal{M}$. Thus for each sample $s \in \mathcal{D}$, we compute the confidence of $\mathcal{M}$ before and after it is added to $\mathcal{T}$. The difference between the two values is the *Confidence Improvement* (CIP) s brings to $\mathcal{M}$. More specifically, we compute the CIP of sample s as follows:

$$\text{CIP}(s, \mathcal{M}, \mathcal{T}, \mathcal{E}) = \text{conf}(\mathcal{M}, \mathcal{T} \cup \{s\}, \mathcal{E}) - \text{conf}(\mathcal{M}, \mathcal{T}, \mathcal{E}) \tag{4}$$

Note that the CIP of sample s is equal to the actual model confidence improvement if only s is acquired, and can be lower than the actual model confidence improvement if multiple samples are acquired, as will be shown in Section 3.

Similar to $\text{conf}(*)$, CIP is also cumulative, i.e., the overall improvement s brings to model $\mathcal{M}$ is the summation of the individual confidence improvement s brings to each sample in $\mathcal{E}$; an equivalent

way to calculate CIP as in Equation (4) is:

$$\mathrm{CIP}(s, \mathcal{M}, \mathcal{T}, \mathcal{E}) = \sum_{e \in \mathcal{E}} \mathrm{CIP}(s, \mathcal{M}, \mathcal{T}, \{e\}) \tag{5}$$

Note that $\mathrm{CIP}(s, \mathcal{M}, \mathcal{T}, \{e\})$ might be (1) positive, i.e., inserting s into $\mathcal{T}$ increases the model confidence regarding e, which denotes that s becomes the new nearest neighbor of e with label $\mathcal{M}(e)$ and thus $G(e, \mathcal{T} \cup \{s\}) < G(e, \mathcal{T})$; (2) negative, i.e., adding s to $\mathcal{T}$ decreases the model confidence regarding e, which denotes that s becomes the new nearest neighbor of e with label other than $\mathcal{M}(e)$ and thus $\bar{G}(e, \mathcal{T} \cup \{s\}) < \bar{G}(e, \mathcal{T})$; and (3) zero, which means $G(e, \mathcal{T} \cup \{s\}) = G(e, \mathcal{T})$ and $\bar{G}(e, \mathcal{T} \cup \{s\}) = \bar{G}(e, \mathcal{T})$. In this work we assume that there exists at least B samples in $\mathcal{D}$ with positive CIP.

To facilitate the computation of CIP (more specifically, $\mathrm{conf}(*)$) which involves calculating the Euclidean distance between samples, we assume that the coordinates of the samples in $\mathcal{D}$ are exposed to the model owner, and the model owner pays a price to reveal the label of a sample. In Section 5 we discuss data acquisition when the assumption does not hold.

We introduce *ultimate model confidence improvement* (UMCI) below to quantify the utility of the acquired samples.

Definition 2.1. **Ultimate Model Confidence Improvement**. Let $\mathcal{D}_B$ of size B be a subset of $\mathcal{D}$. The ultimate model confidence improvement after acquiring $\mathcal{D}_B$ is $\mathrm{conf}(\mathcal{M}, \mathcal{T} \cup \mathcal{D}_B, \mathcal{M}) - \mathrm{conf}(\mathcal{M}, \mathcal{T}, \mathcal{M})$.

Note that since the confidence metric in Equation (1) is based on the distances between samples, which heavily depend on the dimensionality and scale of the data domain, UMCI is not a metric to be compared across datasets and tasks. Rather, it is designed for single dataset-related evaluation, e.g., measure the effectiveness of different data acquisition algorithms or compare different model options for the same task.

The problem studied in this work can be defined as follows:

Definition 2.2. **Confidence-oriented Data Acquisition**. Given (1) a Machine Learning model $\mathcal{M}$ and the dataset $\mathcal{T}$ on which $\mathcal{M}$ is trained, (2) an evaluation set $\mathcal{E}$ on which the confidence of $\mathcal{M}$ is measured, (3) a data pool $\mathcal{D}$ from which new samples can be acquired, and (4) a budget B, the objective of confidence-oriented data acquisition (or simply data acquisition when there is no ambiguity) is to choose B samples to acquire from $\mathcal{D}$ (denoted by $\mathcal{D}_B$) and retrain $\mathcal{M}$ on $\mathcal{T} \cup \mathcal{D}_B$, so as to maximize the ultimate model confidence improvement.

According to Definition 2.2, we use $(\mathcal{M}, \mathcal{T}, \mathcal{E}, \mathcal{D}, B)$ to denote a particular data acquisition instance. We refer to the subset of $\mathcal{D}$ with B samples that lead to the maximal confidence improvement as the *optimal solution* to the instance when there is no ambiguity.

3 THE STUDY OF OPTIMAL SOLUTIONS

In this section, we first analyze the hardness to obtain the optimal solution to the problem introduced in Definition 2.2 in generic cases, and then propose two special cases, under which the optimal solution can be derived in polynomial time, to more comprehensively study the data acquisition task from different perspectives.

3.1 Properties of the Data Acquisition Task

In order to better understand the data acquisition process and showcase its computational complexity, we first demonstrate an important property regarding confidence improvement.

Theorem 3.1. *Let $\mathcal{D}_B$ be the set of samples acquired from $\mathcal{D}$, then conf$(\mathcal{M} \cup \mathcal{D}_B, \mathcal{T}, \mathcal{E})$ can be larger than, smaller than, or equal to conf$(\mathcal{M}, \mathcal{T}, \mathcal{E}) + \sum_{s \in \mathcal{D}_B}$ CIP$(s, \mathcal{M}, \mathcal{T}, \mathcal{E})$*

Intuitively, Theorem 3.1 states that the overall confidence improvement a set of samples $\mathcal{D}_B$ brings to model $\mathcal{M}$ is not the cumulative CIP values of samples in $\mathcal{D}_B$. Thus, after sample s_i is acquired and added to $\mathcal{T}$, the CIP values of remaining samples in $\mathcal{D}$ need to be updated to reflect the influence of s_i on the confidence of $\mathcal{M}$. More specifically, after s_i is acquired, the CIP value of an arbitrary sample $s_j \in \mathcal{D}$ becomes CIP$(s_j, \mathcal{M}, \mathcal{T} \cup \{s_i\}, \mathcal{E})$. Theorem 3.1 can thus be equivalently rewritten into the following.

Theorem 3.2. *$\forall s_i, s_j \in \mathcal{D}$, CIP$(s_j, \mathcal{M}, \mathcal{T} \cup \{s_i\}, \mathcal{E})$ can be larger than, smaller than, or equal to CIP$(s_j, \mathcal{M}, \mathcal{T}, \mathcal{E})$*

We can prove Theorem 3.2 following the way conf$(*)$ is computed in Equation (2), the detailed proof is omitted for brevity but is available in the technical report of the paper.

Since it is non-trivial to predict how the CIP values of remaining samples would change after a certain sample is acquired, identifying the set of samples $\mathcal{D}_B \subset \mathcal{D}$ which leads to the highest confidence improvement requires exhaustively searching all subsets of size B of $\mathcal{D}$. We formally analyze the complexity of the task below.

Theorem 3.3. *Identifying the optimal solution for the Machine Learning Model Confidence-oriented Data Acquisition is NP-hard.*

Proof. The Max Cover Problem (MCP) [37] is a known NP-hard problem, which is defined as follows: given a number k and a collection of sets which may have some common elements, select the k sets so that the union of the selected sets has maximal size. Given a particular MCP instance, we can reduce it to the problem defined herein as follows:

(1) Each set in MCP is mapped to a sample in $\mathcal{D}$ in the data acquisition problem;
(2) Each element in MCP is mapped to a sample in $\mathcal{E}$ in the data acquisition problem;
(3) k in MCP is mapped to budget B in the data acquisition problem;
(4) Let s_i be an arbitrary sample in $\mathcal{D}$ and S_i be the corresponding set in MCP, for each sample $e_j \in \mathcal{E}$ and the corresponding element E_j in MCP problem, if $E_j \in S_i$, set CIP$(s_i, \mathcal{M}, \mathcal{T}, \{e_j\})$ to 1, otherwise set CIP$(s_i, \mathcal{M}, \mathcal{T}, \{e_j\})$ to 0;
(5) Solve the data acquisition problem instance defined above using a hypothetical algorithm A, and let $s_i, s_2, \cdots, s_B$ be the answer. The sets corresponding to $s_i, s_2, \cdots, s_B$ thus form the optimal answer to the given MCP instance.

In conclusion, if there exists an algorithm A that can solve the data acquisition problem in polynomial time, MCP can be solved in polynomial time as well, which is a contradiction. Therefore, algorithm A does not exist, and the data acquisition problem defined in Definition 2.2 is NP-hard. □

Although it is NP-hard to derive the optimal solution to the data acquisition problem in generic cases, in the subsequent sections we show that in certain scenarios, we can leverage properties of CIP to design algorithms identifying the optimal solution without exhaustively enumerating all subsets (of size B) of $\mathcal{D}$.

3.2 Top-B Independence and Bulk Acquisition

In this section, we analyze the case that the CIP values of samples in $\mathcal{D}$ satisfy a particular condition referred to as *Top-B Independence*. We develop an algorithm that can identify the samples which lead to the maximal confidence improvement in polynomial time.

As pointed out in Section 3.1, the major difficulty in solving the data acquisition problem comes from the mutual dependence between samples in $\mathcal{D}$: acquiring a sample may influence

the CIP values of the remaining samples, and since such influence is not predictable, exhaustive enumeration is required to identify the optimal solution. Looking at the mutual dependence from another perspective, if there exists a condition when the samples leading to the maximal confidence improvement have no influence on their corresponding CIP values, then the acquisition would become more tractable. We formally define the condition below.

Definition 3.4. **Top-B Independence**. Let $\mathcal{D}_B$ be the B samples in $\mathcal{D}$ with the highest CIP values. If $\forall \mathcal{D}_B^{sub} \subset \mathcal{D}_B$, and $\forall s \in \mathcal{D}_B \setminus \mathcal{D}_B^{sub}$, $\text{CIP}(s, \mathcal{M}, \mathcal{T} \cup \mathcal{D}_B^{sub}, \mathcal{E}) = \text{CIP}(s, \mathcal{M}, \mathcal{T}, \mathcal{E})$, then samples in $\mathcal{D}$ are said to satisfy the Top-B Independence condition.

While Top-B independence makes the data acquisition process tractable, checking if a particular problem instance satisfies this condition still requires the exhaustive search of $\forall \mathcal{D}_B^{sub} \subset \mathcal{D}_B$ and $\forall s \in \mathcal{D}_B \setminus \mathcal{D}_B^{sub}$. Below we introduce a case when we can verify Top-B independence simply by looking at the CIP values of samples in $\mathcal{D}_B$ calculated on $\mathcal{T}$, $\mathcal{M}$, and $\mathcal{E}$.

Theorem 3.5. *Let $s_1, s_2, \cdots, s_B$ be the B samples in $\mathcal{D}$ with the highest CIP values. If $\forall i, j \in [1, B],\ i \neq j$, and $\forall e \in \mathcal{E}$,*
$\text{CIP}(s_i, \mathcal{M}, \mathcal{T}, \{e\}) * \text{CIP}(s_j, \mathcal{M}, \mathcal{T}, \{e\}) = 0$, then Top-$B$ independence is satisfied.

Proof. The objective is to prove $\forall \mathcal{D}_B^{sub} \subset \mathcal{D}_B$, and $\forall s \in \mathcal{D}_B \setminus \mathcal{D}_B^{sub}$, $\text{CIP}(s, \mathcal{M}, \mathcal{T} \cup \mathcal{D}_B^{sub}, \mathcal{E}) = \text{CIP}(s, \mathcal{M}, \mathcal{T}, \mathcal{E})$, which is equivalent to proving $\forall s_i, s_j \in \mathcal{D}$, $\forall e \in \mathcal{E}$, $\text{CIP}(s_i, \mathcal{M}, \mathcal{T}, \{e\}) = \text{CIP}(s_i, \mathcal{M}, \mathcal{T} \cup \{s_j\}, \{e\})$. Since $\text{CIP}(s_i, \mathcal{M}, \mathcal{T}, \{e\}) * \text{CIP}(s_j, \mathcal{M}, \mathcal{T}, \{e\}) = 0$, we consider two cases:

(1) If $\text{CIP}(s_i, \mathcal{M}, \mathcal{T}, \{e\}) == 0$, according to Equation (4), we know that $\text{conf}(\mathcal{M}, \mathcal{T}, \mathcal{E}) = \text{conf}(\mathcal{M}, \mathcal{T} \cup s_i, \mathcal{E})$. And according to Equation (2), we know that $d(e, s_i) > \bar{F}(e, \mathcal{T})$ and $d(e, s_i) > F(e, \mathcal{T})$, i.e., there exists at least one sample in $\mathcal{T}$ with the same label as s_i that is closer to e than s_i, and evidently inserting s_j into $\mathcal{T}$ does not change the condition. Therefore, $\text{CIP}(s_i, \mathcal{M}, \mathcal{T} \cup \{s_j\}, \{e\})$ is still zero.

(2) If $\text{CIP}(s_i, \mathcal{M}, \mathcal{T}, \{e\}) \neq 0$, since $\text{CIP}(s_i, \mathcal{M}, \mathcal{T}, \{e\}) * \text{CIP}(s_j, \mathcal{M}, \mathcal{T}, \{e\}) = 0$, we know $\text{CIP}(s_j, \mathcal{M}, \mathcal{T}, \{e\}) = 0$, and according to Equation (2) the following is True:
   - $\text{conf}(\mathcal{T}) = \text{conf}(\mathcal{T} \cup \{s_j\})$
   - $\bar{G}(e, \mathcal{T}) = \bar{G}(e, \mathcal{T} \cup \{s_j\})$, or equivalently, $d(e, s_j) > \bar{G}(e, \mathcal{T})$
   - $G(e, \mathcal{T}) = G(e, \mathcal{T} \cup \{s_j\})$, or equivalently, $d(e, s_j) > G(e, \mathcal{T})$

   Since $\text{CIP}(s_i, \mathcal{M}, \mathcal{T} \cup \{s_j\}, \{e\}) = \text{conf}(\mathcal{M}, \mathcal{T} \cup \{s_i\} \cup \{s_j\}, \{e\}) - \text{conf}(\mathcal{M}, \mathcal{T} \cup \{s_j\}, \{e\})$, and from above statement we know $\text{conf}(\mathcal{M}, \mathcal{T} \cup \{s_j\}, \{e\}) = \text{conf}(\mathcal{M}, \mathcal{T}, \{e\})$, we only need to prove $\text{conf}(\mathcal{M}, \mathcal{T} \cup \{s_i\} \cup \{s_j\}, \{e\}) = \text{conf}(\mathcal{M}, \mathcal{T} \cup \{s_i\}, \{e\})$. And $\text{conf}(\mathcal{M}, \mathcal{T} \cup \{s_i\} \cup \{s_j\}, \{e\}) = \min\{\bar{G}(e, \mathcal{T} \cup \{s_i\}), d(e, s_j)\} - \min\{\bar{G}(e, \mathcal{T} \cup \{s_i\}), d(e, s_j)\} = \bar{G}(e, \mathcal{T} \cup \{s_i\}) - G(e, \mathcal{T} \cup \{s_i\}) = \text{conf}(\mathcal{M}, \mathcal{T} \cup \{s_i\}, \{e\})$. Therefore, $\text{CIP}(s_i, \mathcal{M}, \mathcal{T}, \{e\}) = \text{CIP}(s_i, \mathcal{M}, \mathcal{T} \cup \{s_j\}, \{e\})$.

□

Given a data acquisition instance, we can compute the CIP values of all samples in $\mathcal{D}$, and if the CIP values demonstrate the property introduced in Theorem 3.5, then Top-B independence is satisfied; as a result determining the optimal solution to the data acquisition problem becomes easy.

Theorem 3.6. *When Top-B Independence condition is satisfied, the B samples with the highest CIP values form the optimal solution.*

Proof. When the Top-B independence is satisfied, the B samples with the highest CIP values lead to an overall model confidence improvement that is equal to the cumulative CIP value of these

B samples. This is the highest possible accumulative CIP value of any B samples in $\mathcal{D}$. We provide the formal proof in the technical report of the paper. □

We summarize the process of calculating CIP values and acquiring samples in Algorithm 1. Since B samples are acquired all together, we refer to the method as *Bulk Acquisition* (BA).

Algorithm 1 Bulk Acquisition (BA)

Input: Model $\mathcal{M}$, training set $\mathcal{T}$, evaluation set $\mathcal{E}$, budget B, $\mathcal{D}$
1: **Initialization:** CIPs=2D matrix;
2: **for** $s \in \mathcal{D}$ **do**
3: **for** $e \in \mathcal{E}$ **do**
4: CIPs[s][e]=CIP($s, \mathcal{M}, \mathcal{T}, \{e\}$);
5: CIPs[s]['overall']=$\sum_{e \in \mathcal{E}}$ CIPs[s][e];
6: Sort CIPs[:]['overall'] in descending order;
7: Acquire samples corresponding to CIPs[0:B]

The major cost of BA comes from the calculations of CIP values regarding each pair of $s \in \mathcal{D}$ and $e \in \mathcal{E}$, while the acquisition cost after all CIP values are derived (and sorted) is linear. According to Equations (1) and (4), the complexity of BA is $O(|\mathcal{D}| * |\mathcal{T}| * |\mathcal{E}| * D)$, where D is the dimensionality of the feature space.

Another desired property of Top-B independence, as introduced in Definition 3.4, is that the confidence improvement of the model is known before the acquisition starts. In practice, when the condition in Theorem 3.5 is not satisfied, one still needs to know the potential confidence improvement prior to acquisition; we adopt a variant of Algorithm 1, and instead of acquiring the B samples with the highest CIP values, we traverse the CIP values in descending order, and acquire the first B samples which satisfy the condition given in Theorem 3.5. The cumulative CIP value of these B samples is guaranteed to be the confidence improvement of the model. Note that since with BA samples are acquired collectively in one round, model retraining commences after all samples are acquired.

3.3 Progressive Dominance and Sequential Acquisition

With Top-B independence we focus on a case when samples forming the optimal solution are mutually independent and can be acquired together, eliminating the need of updating the CIP values. In this section we view the problem from a different perspective and discuss a scenario in which we can acquire samples in the optimal solution gradually in iterations, with CIP values being repetitively updated along the process.

One challenge in selecting samples that are not mutually independent is that, acquiring a sample would change the CIP values of the remaining samples, and if samples are greedily chosen purely based on CIP values, we may not identify the optimal solution. In other words, for a specific round of acquisition and two samples s_i and s_j in $\mathcal{D}$, the CIP value of s_i being higher than that of s_j does not necessarily mean that acquiring s_i would lead to a higher ultimate model confidence improvement.

However, considering the property of model confidence and CIP (Equation (2) in particular), there are certain cases when we can conclude that a specific sample leads to higher ultimate confidence improvement than another, based on the CIP values. We introduce the notion of *dominance* below to formally define such cases.

Definition 3.7. **Dominance**. Let s_i and s_j be two samples in $\mathcal{D}$, if $\forall e \in \mathcal{E}$, $\text{CIP}(s_i, \mathcal{M}, \mathcal{T}, \{e\}) \geq \text{CIP}(s_j, \mathcal{M}, \mathcal{T}, \{e\})$, we say s_i dominates s_j.

Table 1. An Example of Dominance

CIP	e_1	e_2	e_3
s_1	0.4	0	0.2
s_2	0.5	0.1	0.3
s_3	0.1	0	0.1

Dominance describes the case when the CIP value of one sample in $\mathcal{D}$ is larger than the CIP value of another sample in $\mathcal{D}$ with respect to each evaluation sample. In Table 1 we illustrate dominance with an example, where s_1, s_2, s_3 are samples in $\mathcal{D}$, e_1, e_2, e_3 are samples in $\mathcal{E}$, and the value in the cell corresponding to s_i and e_j denotes $\text{CIP}(s_i, \mathcal{M}, \mathcal{T}, \{e_j\})$. As per Definition 3.7, s_2 dominates s_1 and s_3, and s_1 dominates s_3.

Theorem 3.8. *If s_i dominates s_j, then $\forall \mathcal{D}_{sub} \subset \mathcal{D}$, $conf(\mathcal{M}, \mathcal{T} \cup \{s_i\} \cup \mathcal{D}_{sub}, \mathcal{E}) \geq conf(\mathcal{M}, \mathcal{T} \cup \{s_j\} \cup \mathcal{D}_{sub}, \mathcal{E})$.*

Proof. We use $\mathcal{T}'$ to denote $\mathcal{T} \cup \mathcal{D}_{sub}$ and prove $\text{conf}(\mathcal{M}, \mathcal{T}' \cup \{s_i\}, \mathcal{E}) \geq \text{conf}(\mathcal{M}, \mathcal{T}' \cup \{s_j\}, \mathcal{E})$, which is equivalent to proving $\text{CIP}(s_i, \mathcal{M}, \mathcal{T}', \mathcal{E}) \geq \text{CIP}(s_j, \mathcal{M}, \mathcal{T}', \mathcal{E})$. According to Equation (5), we can prove $\forall e \in \mathcal{E}, \text{CIP}(s_i, \mathcal{M}, \mathcal{T}, \{e\}) \geq \text{CIP}(s_j, \mathcal{M}, \mathcal{T}, \{e\})$, which is true according to the definition of dominance. □

The property in Theorem 3.8 provides a way to compare two samples in terms of ultimate model confidence improvement, and leveraging the property we can identify samples in the optimal solution in certain cases.

Theorem 3.9. *For sample $s \in \mathcal{D}$, if s dominates $|\mathcal{D}| - B$ samples, then s is guaranteed to be in the optimal solution.*

Proof. Intuitively, if s dominates $|\mathcal{D}| - B$ samples, according to Theorem 3.8, there are at most $B - 1$ samples leading to higher ultimate model confidence improvement than s. Since the budget is B, s is in the optimal solution. □

With Theorem 3.9 we can identify samples that are guaranteed to be in the optimal solution. Since we make no assumptions regarding mutual independence, after samples dominating others are acquired, the CIP values of the remaining samples would change, and thus an update is needed.

Since iterative CIP updates are required, we consider an acquisition strategy named *Sequential Acquisition* (SA), in which we acquire one sample from samples yet to be acquired, in $\mathcal{D}$ and update the CIP values of the remaining samples in $\mathcal{D}$ in each round, and repeat the process for B rounds. Below we introduce a scenario in which we observe dominance repeatedly in each round of sequential acquisition, and showcase that SA leads to the optimal solution under this scenario.

Definition 3.10. **Progressive dominance**. Progressive dominance refers to the condition that in each round of SA, there is at least one sample $s \in \mathcal{D}$ dominating $|\mathcal{D}| - B$ samples in $\mathcal{D}$.

Note that with progressive dominance, while the number of samples to be dominated in each round is always $|\mathcal{D}| - B$, since samples are acquired along the process and thus removed from $\mathcal{D}$, the condition becomes more and more restrictive as more samples are acquired, and in the last round of SA, the selected sample dominates all remaining samples. We provide the detailed steps of SA when progressive dominance is satisfied as Algorithm 2.

Algorithm 2 Sequential Acquisition (SA)

Input: Model $\mathcal{M}$, training set $\mathcal{T}$, evaluation set $\mathcal{E}$, budget B, $\mathcal{D}$

```
while B>0 do
    for s ∈ D, e ∈ E do
        CIP[s][e] = CIP(s, M, T, {e})
    for s ∈ D do
        dominantCount=|D| − 1; //except for s
        for s′ ∈ D \ {s} do
            dominant=True
            for e ∈ E do
                if CIP[s][e] < CIP[s′][e] then
                    dominant=False;
                    break;
            if dominant == False then
                dominantCount-=1;
        if dominantCount≥ |D| − B then
            T = T ∪ {s}
            D = D \ {s}
            B = B − 1
```

Theorem 3.11. *SA leads to the maximal confidence improvement if progressive dominance is satisfied.*

Proof. According to Definition 3.10, the condition of Theorem 3.9 is satisfied in each round of the sequential acquisition process, and thus a sample in the optimal solution can be acquired in each round, and the optimal solution is gradually constructed. The formal proof is provided in the full technical report of the paper. □

Note that if we can identify more than one sample in each round satisfying the property in Theorem 3.9, we can acquire all of these samples in the same round, and it is evident that the acquired samples still form the optimal solution.

Even if the dominance condition is not satisfied in all rounds, we can still apply the sample selection method to only those rounds where the condition is met, and we leverage other acquisition methods for the rest of the rounds.

According to Algorithm 2, the complexity of Sequential Acquisition is $O(B * \max\{|\mathcal{D}| * |\mathcal{T}| * |\mathcal{E}| * D, |\mathcal{D}|^2 * |\mathcal{E}|\})$, where D is the dimensionality of the feature space.

Note that while top-B independence and progressive dominance do not necessarily hold in practice, the theoretical analysis in this section assist in the understanding of the nature of the data acquisition task studied in this work. In addition, while BA and SA are proposed in the context of top-B independence and progressive dominance respectively, their application is not limited to these two scenarios but can serve as the acquisition strategy in practice. As will be experimentally shown in Section 6, for acquisition tasks where top-B independence and progressive dominance are not necessarily satisfied, BA and SA still yield high quality acquisitions.

4 NEIGHBOR-BASED ACQUISITION

While the acquisition strategies introduced in Section 3 are able to derive the optimal solution under certain assumptions, their complexity, especially SA, is high, as repetitive computation of pair-wise distances is needed. In this section, we present more general solutions to this problem; namely we

present efficient algorithms for this problem under no assumptions, without significantly affecting optimality.

4.1 Neighbor-based Pruning

As discussed in Section 3.2 and Section 3.3, the complexity of the proposed acquisition strategies depends on $|\mathcal{T}|$, $|\mathcal{D}|$, $|\mathcal{E}|$ and D. While it is non-trivial to alter $|\mathcal{T}|$, $|\mathcal{E}|$ and D as they are intrinsic to the problem definition, it is possible to shrink the data pool $\mathcal{D}$ and thus reduce the number of operations required to update CIP values for samples in $\mathcal{D}$.

We can observe from Equation (1) and Equation (2) that the confidence of model $\mathcal{M}$ with respect to a particular sample $e \in \mathcal{E}$ depends on the distances between e and its nearest neighbor with the same/other labels in $\mathcal{T}$, i.e., $G(e, \mathcal{T})$ and $\bar{G}(e, \mathcal{T})$. Improving the model confidence with respect to sample e is equivalent to decreasing the value of $G(e, \mathcal{T})$. In other words, samples in $\mathcal{D}$ that are close to a sample in $\mathcal{E}$ (say e) with the same label are more likely to increase the overall confidence of $\mathcal{M}$. Inspired by this observation, we apply neighbor-based pruning on $\mathcal{D}$, only retaining a sample in $\mathcal{D}$ if it is close to a sample with the same label in $\mathcal{E}$, and conduct data acquisition on the retained subset of $\mathcal{D}$. More specifically, for each sample $e \in \mathcal{E}$, we identify its k nearest neighbors in $\mathcal{D}$, denoted by $k\text{NN}(e, \mathcal{D})$, and thus we generate a candidate set $C = \bigcup_{e \in \mathcal{E}} k\text{NN}(e, \mathcal{D})$. We then apply SA on candidate set C instead of the entire data pool $\mathcal{D}$, to reduce the cost of repetitive computation of pair-wise distances; this variant is denoted by kNN-SA. The same principle can be applied naturally to BA and we denote that variant by kNN-BA. In order to accelerate the construction of C, we can build indexes on $\mathcal{D}$, such as an R-Tree, to facilitate the kNN search, which can be reused.

Note that calculating model confidence regarding sample e only requires the distance to its nearest neighbors. Using $k = 1$ to construct C may lead to suboptimal results, especially under budget constraints, due to the inter-dependency in the selection of the sample to retain for different e's. We illustrate this issue in Figure 1, where samples s_1, s_2, s_3 are in $\mathcal{D}$, e_1, e_2 are in $\mathcal{E}$, and the nearest neighbors of e_1 and e_2 are s_1 and s_3 respectively. Suppose that $C = \{s_1, s_3\}$, and budget $B = 1$, then either s_1 or s_3 will be acquired, while s_2 may lead to the highest confidence improvement since it is close to both e_1 and e_2.

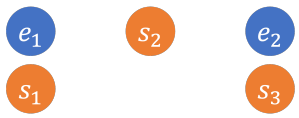


Fig. 1. Example of $k = 1$

As efficient approximations of BA and SA, kNN-BA and kNN-SA offer trade-off between confidence improvement and time cost by varying the value of k and thus the size of C, so that the users can choose the most suitable approach based on specific requirements and constraints.

4.2 Bounding kNN-BA and kNN-SA

In Section 3, we addressed how BA and SA can identify optimal solutions in specific cases. Here, we demonstrate that kNN-SA and kNN-BA, while designed as more efficient approximate variants of SA and BA have bounded approximate guarantees. The following analysis considers 1NN-SA and 1NN-BA without loss of generality and can be similarly generalized to other values of k.

Theorem 4.1. *Let V^C be the maximal possible model confidence improvement of applying SA or BA on C, and V^* be the maximal possible model confidence improvement of applying SA or BA on $\mathcal{D}$. Then $\frac{V^C}{V^*} \geq \frac{B}{|\mathcal{E}|}$.*

Proof. Suppose the optimal solution on C is $s_1, s_2, \cdots, s_B$, which are the nearest neighbors of $e_1, e_2, \cdots, e_B$ respectively. Let v_i^C $(i \in [1, B])$ be the confidence improvement s_i brings to e_i, then $V^C = \sum_{i=1}^{B} v_i^C + \Delta$, where Δ is the overall confidence $s_1, s_2, \cdots, s_B$ bring to samples $e_{B+1}, e_{B+2}, \cdots, e_{|\mathcal{E}|}$ (samples in $\mathcal{E}$ other than $e_1, e_2, \cdots, e_B$).

Since s_i is the nearest neighbor of e_i, v_i^C is the maximal possible confidence improvement regarding e_i. We can write V^* as $V^* = \sum_{i=1}^{B} v_i' + \Delta'$, where v_i' $(i \in [1, B])$ denotes the confidence improvement of sample e_i in the optimal solution, and clearly $v_i' \leq v_i^C$, and Δ' denotes the overall confidence improvement the optimal solution brings to samples $e_{B+1}, e_{B+2}, \cdots, e_{|\mathcal{E}|}$.

Since $s_1, s_2, \cdots, s_B$ form the answer, the confidence improvement of any sample in $\{e_{B+1}, e_{B+2}, \cdots, e_{|\mathcal{E}|}\}$ cannot be higher than that of $e_1, e_2, \cdots, e_B$. Therefore, $\Delta' \leq (|\mathcal{E}| - B) * \min\{v_i' | i \in [1, B]\}$.

Since the minimal value of Δ is zero, $\frac{V^C}{V^*} \geq \frac{\sum_{i=1}^{B} v_i^C}{\sum_{i=1}^{B} v_i^C + (|\mathcal{E}|-B) * \min\{v_i^C | i \in [1,B]\}}$, and clearly the right-hand side of equation takes the lowest value when $v_1^C = v_2^C = \cdots v_B^C$, and thus $\frac{V^C}{V^*} \geq \frac{B}{|\mathcal{E}|}$ □

Thus kNN-SA and kNN-BA are approximate solutions with accuracy guarantees and can benefit applications in which strict requirements regarding the confidence improvement are in place, while incurring lower data acquisition overhead compared to SA and BA.

4.3 BA and SA in Generic Cases

Although BA and SA make certain assumptions regarding the data distribution to build the optimal solution, they can still be utilized when this condition is not met, serving as a heuristic approximation method. We will demonstrate that it achieves high quality solutions in practical scenarios (Section 6).

Fig. 2. Example of Over Exploitation

We consider a generic case when the CIP values may have an arbitrary distribution, and demonstrate another advantage of kNN-BA. Considering the neighbors of each sample in $\mathcal{E}$ to construct C, kNN-BA reduces the probability of over-exploiting a particular region in the data space. Over-exploitation refers to the case when we acquire many samples in proximity that have non material influence on the model confidence, and as a result, acquiring these samples results in an overall confidence improvement that is much lower than the accumulative CIP values. As an example, samples s_1, s_2, s_3, s_4 in Figure 2 all have influence on the confidence of sample e_1; however, once the sample closest to e_1 (say s_1) is acquired, the remaining three samples have non material influence on the confidence of e_1 anymore, according to the definition in Equation (1). Since BA in the general case follows a greedy strategy (i.e., acquires samples with the highest CIP values) involving no CIP value updates, it likely leads to over-exploitation. With neighbor-based pruning in kNN-BA, we only retain k samples close to each sample in $\mathcal{E}$, and thus over-exploiting a particular region is less likely. As a result, the overall confidence improvement will potentially be higher. We illustrate this advantage using the same example in Figure 2 and set $k = 1$ (thus $C = \{s_1, s_5, s_6\}$). Assume the CIP values of samples in Figure 2 are as follows:

Table 2. Example of CIP values

Sample	S_1	S_2	S_3	S_4	S_5	S_6
CIP	0.71	0.7	0.7	0.7	0.65	0.64

Without neighbor-based pruning, samples close to e_1 (i.e., s_1, s_2, s_3, or s_4) will be acquired first, as they have higher CIP values, resulting in over-exploitation of the region around e_1. Constructing C forces the acquired samples to be more spread in the data space, overcoming the over-exploitation issue and leading to higher confidence improvement.

4.4 Algorithm and Analysis

Assuming an index on $\mathcal{D}$ has been built in a pre-processing step, we summarize the steps of kNN-BA and kNN-SA in Algorithm 3, which includes generating candidate set C (Lines 1-4) and conducting data acquisition on C(Line 5).

Algorithm 3 kNN-BA/kNN-SA

Input: Model $\mathcal{M}$, training set $\mathcal{T}$, evaluation set $\mathcal{E}$, budget B, $\mathcal{D}$, index R, k

1: $C = \emptyset$
2: **for** $e \in \mathcal{E}$ **do**
3: Find kNN$(e, \mathcal{D})$ using R;
4: $C = C \cup k$NN$(e, \mathcal{D})$;
5: Invoke BA or SA with $\mathcal{M}, \mathcal{T}, \mathcal{E}, B, C$;

According to Algorithm 3, kNN-BA/kNN-SA consists of two phases: (1) candidate generation, the cost of which is $O(|\mathcal{E}| * c)$, where c denotes the average cost of querying the index structure built on $\mathcal{D}$, and (2) sample acquisition, the cost of which is $O(|C| * |\mathcal{T}| * |\mathcal{E}| * D)$ for kNN-BA, and $O(B * \max\{|C| * |\mathcal{T}| * |\mathcal{E}| * D, |C|^2 * |\mathcal{E}|\})$ for kNN-SA. Based on the complexity analysis, we can conclude that kNN-BA and kNN-SA have clear efficiency advantage over BA and SA when $\mathcal{D}$ is large compared to C.

Acquiring data from a pruned dataset offers multiple advantages in our problem setting. First, it scales better with respect to data pool size and thus can be applied to scenarios when samples are acquired from a vast data source. Second, it is more suitable for cases when the coordinates of samples in $\mathcal{D}$ are not completely accessible to the model owner. For example, in a data market setting where the data provider sells data for profit [7, 28] and is thus motivated to minimize the exposed information (e.g., setting a constraint on the number of samples whose coordinates can be accessed by the buyer – the model owner), the model owner can send dataset $\mathcal{E}$ to the data provider and ask the data provider to build C and share the coordinates of samples in C only.

5 DISTRIBUTION-BASED ACQUISITION

The acquisition strategies introduced so far, including BA, SA, and their variants, have the following major limitations.

- They do not scale well in terms of $|\mathcal{T}|$ and $|\mathcal{E}|$, because the computation of model confidence involves calculating the pair-wise distance between samples in $\mathcal{T}$ and $\mathcal{E}$.
- They rely on the availability and accessibility of the coordinates of all samples in $\mathcal{D}$ for the computation of CIP values. As a result, the methods are not applicable to settings when these coordinates are not revealed, such as in a data market setting [28].

In this section, we seek to overcome these limitations and design a distribution-based acquisition strategy (DA) which is highly scalable, efficient, and does not require access to $\mathcal{D}$.

5.1 CIP Distribution

From Equation (4) we observe that the CIP of an arbitrary sample depends on its Euclidean distance to samples in $\mathcal{T}$ and $\mathcal{E}$. Imagine that we are given an infinite set $\mathcal{S}_{inf}$ which contains samples filling the entire data space. We can compute the CIP value of each sample in $\mathcal{S}_{inf}$ using Equation (4), producing the intrinsic CIP distribution of the data space with respect to the model's $\mathcal{T}$ and $\mathcal{E}$. Although in practice it is impossible to access such an infinite set $\mathcal{S}_{inf}$, we can generate a set of synthetic samples in the data space (consisting of points not neccesarily in $\mathcal{E}$) and use the CIP values of these samples to approximate the intrinsic CIP distribution. We illustrate the process of using synthetic samples to approximate the CIP distribution in Figure 3, which contains samples from a 1-dimensional data space with the same label, and defer the discussion of generating the synthetic samples to Section 5.2.

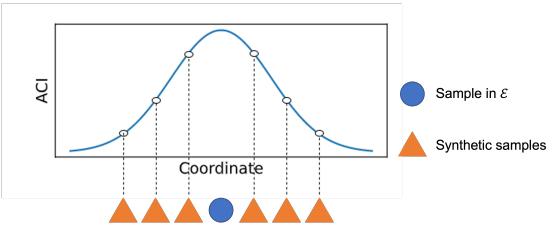


Fig. 3. Approximating CIP distribution

At the bottom of Figure 3 we show the positions of one sample in $\mathcal{E}$ and multiple synthetic samples, and at the top are the CIP values of the synthetic samples (hollow dots) and the CIP distribution thus approximated (blue solid curve). Note that while the figure shows a 1-dimensional example, similar CIP distributions can be built for higher-dimensional spaces as well.

The above observation motivates the construction of a function F to capture the intrinsic CIP distribution of the data space, so that for an incoming sample s_{new}, we can use F to directly predict the CIP of s_{new}, instead of going through the expensive process of computing CIP using Equation (4), to reduce the acquisition cost. In this work we use Machine Learning (ML) models to build F to better handle higher dimensionality and larger $\mathcal{T}$ and $\mathcal{E}$ (and consequently complex CIP distributions).

5.2 Fitting a CIP Distribution Model

To learn the CIP distribution using an ML model, the first step is to generate a training set $\mathcal{DT}$, which consists of a collection of samples $\mathcal{DS}$, drawn from data space Γ, and their corresponding CIP values.

There are two ways to generate $\mathcal{DS}$. The first approach is to randomly draw samples from data space Γ. This approach clearly suffers from severe sparsity as the dimensionality increases, and requires generating an exponentially increasing number of samples to ensure sufficient coverage of the data space, as it does not take into consideration how samples are truly distributed in the data space. Therefore, a more effective solution is to generate $\mathcal{DS}$ based on the actual underlying distribution of samples in Γ. Since the actual distribution is unknown, we can merge $\mathcal{T}$ and $\mathcal{E}$, which are available to the owner of the model and consist of samples drawn from the underlying distribution of Γ, to form $\mathcal{DS}$.

5.3 Analysis of Distribution-based Acquisition

With the model F obtained in Section 5.2, the complexity of obtaining the CIP of an arbitrary sample decreases from $O(\mathcal{T} * D)$ to $O(c_I)$, where c_I denotes the average cost of model inference, a

significant cost reduction especially when $\mathcal{T}$ is large. In addition, since the construction of F does not rely on samples in $\mathcal{D}$, we can further lift the assumption that $\mathcal{D}$ needs to be accessible to the model owner during the data acquisition process. Instead, the data acquisition process can now proceed as follows.

(1) The model owner trains F based on $\mathcal{T}$ and $\mathcal{E}$;
(2) The model owner provides F and a budget B to the data pool;
(3) The data pool utilizes F to predict the CIP values of samples in $\mathcal{D}$, and returns the B samples with the highest CIP values to the model owner.

We refer to this algorithm as *distribution-based acquisition* (DA).

One issue with DA is that, model F is never re-trained during the acquisition process and thus does not consider the mutual influence between the acquired samples. One solution could be to acquire a single sample in each round and re-train F iteratively. However, since model re-training also involves updating the CIP values of samples in $\mathcal{DS}$, the cost of frequent model retraining can easily outweigh the benefit gained from using a model for CIP prediction. Therefore, we propose to conduct DA in a batch mode, which is to acquire ΔB (i.e., the batch size) samples in each round of data acquisition, update the model, and repeat until the budget is exhausted. We provide the detailed steps of DA in Algorithm 4.

Algorithm 4 Distribution-based Acquisition (DA)

Input: Model $\mathcal{M}$, training set $\mathcal{T}$, evaluation set $\mathcal{E}$, budget B, batch size ΔB

1: **function** TRAIN_F($\mathcal{DS}$, $\mathcal{T}$, $\mathcal{E}$, $\mathcal{M}$):
2: $\quad\mathcal{DT} = \emptyset$;
3: $\quad$**for** $s \in \mathcal{DS}$ **do**
4: $\quad\quad$Compute CIP(s) using $\mathcal{T}, \mathcal{E}, \mathcal{M}$;
5: $\quad\quad\mathcal{DT} \cup \{(s, \text{CIP}(s))\}$;
6: $\quad$Train F on $\mathcal{DT}$;
7: $\quad$**Return** F
8: **function** DATA_POOL_PROCESS(F, $\mathcal{D}$, ΔB):
9: $\quad$scores = $\emptyset$;
10: $\quad$**for** $s \in \mathcal{D}$ **do**
11: $\quad\quad$scores = scores $\cup \{(s, F(s))\}$;
12: $\quad$Sort scores in descending order;
13: $\quad$res=scores[:ΔB];
14: $\quad\mathcal{D} = \mathcal{D}\backslash$\{samples in res\};
15: $\quad$**Return** \{samples in res\};
16: $\mathcal{DS} = \mathcal{T} \cup \mathcal{E}$;
17: **while** $B > 0$ **do**:
18: $\quad F$ =TRAIN_F($\mathcal{DS}, \mathcal{T}, \mathcal{E}, \mathcal{M}$);
19: $\quad n = \min\{\Delta B, B\}$
20: $\quad\mathcal{A}$=DATA_POOL_PROCESS(F, n);
21: $\quad\mathcal{T} = \mathcal{T} \cup \mathcal{A}$;
22: $\quad B = B - n$;

In Algorithm 4, to facilitate the distribution-based acquisition, we first define two functions, TRAIN_F (Lines 1-7) which trains the CIP distribution function F, and PROVIDER_PROCESS (Lines 8-15), which requests samples from the data pool. The detailed procedures of both functions have

been introduced at the beginning of Section 5.3. During the acquisition process, the model owner trains F in each round based on acquired samples (Line 18), and acquires ΔB new samples from the data pool (or the number that the remaining budget permits) (Lines 19-21). The above process is repeated until the model owner has acquired B samples (Lines 17 and 22).

The preprocessing step of the distribution-based acquisition method consists of the construction of $\mathcal{DT}$ and the training of F, and its complexity is thus $O((|\mathcal{T}| + |\mathcal{E}|) * |\mathcal{T}| * D + c_T)$, where c_T denotes the training cost of F. The acquisition cost of the approach is $O(B * c_I + \lfloor \frac{B}{\Delta B} \rfloor * ((|\mathcal{T}| + |\mathcal{E}|) * |\mathcal{T}| * D + c_T))$.

6 EXPERIMENTS

In this section, we embark on a series of experimental investigations involving the proposed techniques as well as applicable baselines. Our objective is to thoroughly assess these techniques, considering two key aspects: the performance of each method, measured by the improvement of model confidence following data acquisition, and the associated time overhead.

6.1 Settings

Datasets. We conduct experiments on four datasets, namely MNIST [15], CIFAR10 [26], CIFAR100 [26], Crop [16, 23], and HAR70+ [30]. MNIST, CIFAR10 and CIFAR100 are image classification datasets widely used for computer vision tasks, and Crop and HAR70+ are tabular datasets consisting of real-valued features. For each dataset we reserve a hold-out subset (denoted by $\mathcal{E}_{\text{ho}}$) to measure the ultimate model confidence improvement (UMCI), which is not accessible to the model owner and acquisition algorithm during the acquisition process. For MNIST, CIFAR10, and CIFAR100, we use the test set associated with the corresponding dataset as $\mathcal{E}_{\text{ho}}$; for Crop and HAR70+, since no designated test sets are provided, we use stratified random sampling to select 30% of the entire dataset as $\mathcal{E}_{\text{ho}}$. The characteristics of the four datasets are summarized in Table 3.

Table 3. Dataset Characteristics

dataset	*# records*	*# classes*	*# dimensions*
MNIST	60,000	10	784
CIFAR10	60,000	10	1,024
CIFAR100	60,000	100	1,024
Crop	325,834	7	175
HAR70+	141,714	7	6

Implementation and Environment. The algorithms proposed in the work are implemented using Python (version 3.8.10). Experiments are conducted on a Ubuntu 20.04 instance with Core i5 CPU, 24GB RAM, 256GB SSD, and 2TB HDD. For kNN-BA and kNN-SA, we adopt an off-the-shelf R-Tree implementation [39] to accelerate the construction of the candidate set.

Evaluation metrics. We evaluate each method from two aspects in the experiments, including the ultimate model confidence improvement (UMCI) after B samples are acquired (as introduced in Definition (2.1)), and the time overhead of the acquisition process. Note that the measured time includes the time of acquiring new samples, and excludes the pre-acquisition cost (such as generating synthetic samples and training the CIP distribution model for DA) and the post-acquisition cost (such as retraining the model after the acquisition process to compute UMCI). Each experiment is repeated five times and the average UMCI and time values are reported.

Construction of $\mathcal{T}$, $\mathcal{E}$, and $\mathcal{D}$. We randomly split each dataset used in the experiment into $\mathcal{T}$ (model training set), $\mathcal{E}$ (model confidence evaluation set), and $\mathcal{D}$ (data pool) following power-law

distribution (as per [28]). Unless otherwise stated, the ratio of $\mathcal{T} : \mathcal{E} : \mathcal{D}$ is $1 : 4 : 5$, and the observations under other ratios are similar.

Model Selection. To build model $\mathcal{M}$, we adopt VGG8B [38] for CIFAR10 and CIFAR100, and Multi-Layer Perceptron (MLP) for MNIST and CROP. Our experiments with other models show that the acquisition process and results are not sensitive to model choices.

Choice of F. For the CIP distribution model F used by Distribution-based Acquisition, we have utilized several standard built-in regression models in Scikit-Learn Package with default parameter settings, such as kNN regressor, Random Forest regressor, and MLP regressor, among which Random Forest regressor demonstrates the best trade-off between performance and efficiency. Therefore, for the experiments in this section, and without loss of generality, we use Random Forest regressor to build F.

Choices of B and k. We vary the value of B (budget) in a range to thoroughly study the proposed methods. The range of values is carefully selected to demonstrate the full spectrum of observable behaviours, from $0.01 * |\mathcal{D}|$ when only a small number of samples are acquired, to $0.2 * |\mathcal{D}|$ where we experience diminishing returns across all datasets used in the experiment. In the following we use relative values (i.e., percentage to $|\mathcal{D}|$) to denote B (e.g., 0.01, 0.05) for brevity. We study the effect of k on the UMCI and time required by kNN-BA and kNN-SA in Section 6.3, and set $k = 1$ for other experiments by default.

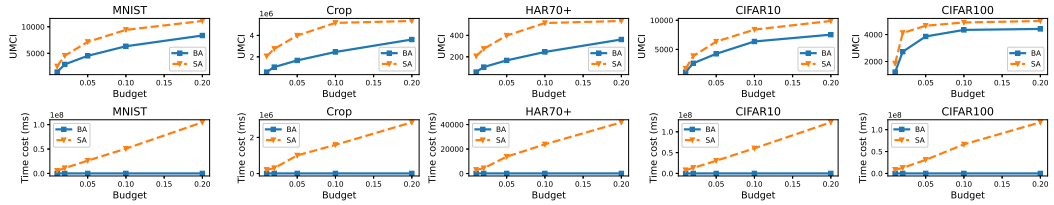


Fig. 4. BA vs. SA

Baselines. In addition to the methods proposed in the paper, we also consider two intuitive acquisition strategies, namely *Random Acquisition* (RA), and *Equality-oriented Acquisition* (EA), as baselines. With RA, we acquire B random samples from $\mathcal{D}$. With EA, samples are acquired in rounds, and in each round, we identify the class with the minimal number of samples in $\mathcal{T}$, say the i-th class, and acquire one sample from the i-th class (and add it into $\mathcal{T}$); the process is repeated until B samples are acquired.

6.2 Studying BA and SA

As discussed in Section 3, Bulk Acquisition (BA) and Sequential Acquisition (SA) can be applied in a general setting (without any specific assumptions on data properties). In this section, we apply both methods to the datasets used in the experiment, comparing their UMCI and time, and demonstrate the results in Figure 4.

The main observations from Figure 4 are listed and analyzed as follows. First, the overall UMCI improves after more samples are acquired, and for the datasets used in the experiment, the improvements exhibit diminishing returns when the budget B reaches $0.05 * |\mathcal{T}|$ to $0.1 * |\mathcal{T}|$, denoting that acquiring more samples has less effect on the model confidence. The reason is that, as more samples are acquired, the $G(e, \mathcal{T})$ value for each sample $e \in \mathcal{E}$ would decreases (as shown in Equation (1)), and acquiring more samples becomes less likely to further reduce $G(e, \mathcal{T})$; thus UMCI becomes increasingly more stable. In practice, it is sensible to monitor the UMCI for diminishing returns during the acquisition process, to avoid acquiring samples with no or negative influence

on model confidence. Second, SA results in higher UMCI compared to BA, and the benefits are more significant for larger B, as SA capitalizes on the frequent updates of CIP values and is able to identify samples leading to higher UMCI given the already acquired samples. Third, the time required by SA is usually hundreds of times higher than that of BA, depending on the budget, due to the repetitive updates of CIP values.

6.3 Studying the Influence of k

Methods kNN-BA and kNN-SA are proposed as efficient alternatives of BA and SA respectively, and their UMCI and time depend on the size of the candidate set we acquire new samples from, which is largely decided by the parameter k. In this section we study the influence of k to both UMCI and the time required by kNN-BA and kNN-SA. We showcase the results of kNN-BA in Figure 5, and similar observations are obtained for kNN-SA.

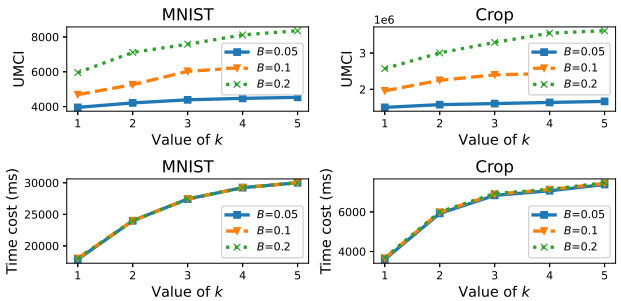


Fig. 5. The Influence of k

As depicted in Figure 5, as k increases, both UMCI and time would increase sub-linearly, for different datasets and budgets; UMCI exhibits diminishing returns approximately at $k = 4$. The reason is that, as revealed by Algorithm 3, as k increases, more nearest neighbors of samples in $\mathcal{E}$ are retained to construct the candidate set, thus more samples in $\mathcal{D}$ with high CIP values are included in the candidate set. However, since the kNNs of different samples in $\mathcal{E}$ may overlap, the candidate set size grows sub-linearly as k. In addition, according to Equation (1), samples in $\mathcal{D}$ that are distant from samples in $\mathcal{E}$ have less impact on the model confidence improvement, and consequently large k would result in diminishing returns. To determine a suitable value for parameter k for a new task, one way is to acquire new samples in rounds starting from $k = 1$. The value of k can then be increased in each round until diminishing returns in UMCI are observed; thus the k value achieving reasonable balance between UMCI and time can be selected accordingly.

6.4 BA vs. kNN-BA

Next we compare the UMCI and time requirements of BA and kNN-BA (with $k = 1$), and provide the results on Crop and MNIST in Figure 6; similar trends are observed on CIFAR10 and CIFAR100.

According to Figure 6, kNN-BA usually results in 20% lower UMCI and runs two to three times faster, due to the fact that kNN-BA acquires new samples from a candidate set smaller than the original data pool $\mathcal{D}$. One interesting observation is that, in certain cases, kNN-BA leads to higher UMCI than BA, especially when the value of B is small (e.g., when $B = 0.01$ in Figure 6). This is due to the fact that BA is not updating the CIP values during the acquisition process. As discussed in Section 4.3, acquiring samples greedily based on CIP values without updating the CIP values periodically causes over-exploitation, lowering the UMCI. In practice, one can adopt either method and set the value of k as suggested in Section 6.3 to achieve desired balance between confidence

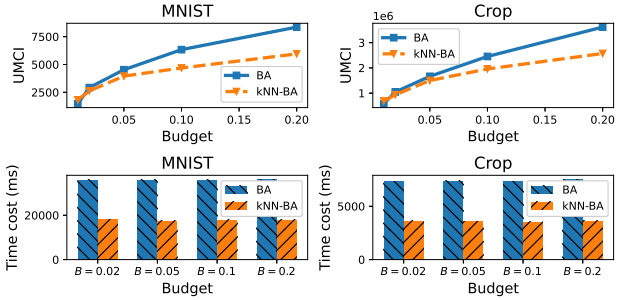


Fig. 6. BA vs. kNN-BA

improvement and time cost. On the other hand, since kNN-BA only retains the kNN of each sample in $\mathcal{E}$ to generate the candidate set, the candidate set is likely to contain a large number of samples in diverse areas of the data space, and the chances of over-exploitation are limited, leading to higher ultimate confidence improvement. As the budget increases, the advantage of acquiring samples from a larger data pool overweighs the benefit of limiting the probability of over-exploitation, and BA starts to outperform kNN-BA in terms of UMCI.

6.5 SA vs. kNN-SA

In this section we compare SA and its efficient variant, kNN-SA (with k=1), in terms of UMCI and acquisition time cost. The results are provided in Figure 7.

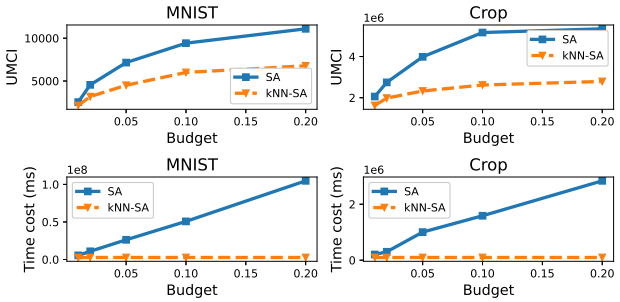


Fig. 7. SA vs. kNN-SA

As is clear from Figure 7, SA always leads to higher UMCI, and incurs a two to forty times higher time overheads than that of kNN-SA. Both SA and kNN-SA repetitively update the CIP values of samples yet to acquire, and thus do not suffer from over-exploitation. Since SA is associated with a larger data pool, it is more likely to acquire samples with higher CIPs (and thus leads to higher UMCI), compared to kNN-SA which is limited to a small candidate set. On the other hand, acquiring samples from a larger data pool also increases the time required by SA, compared to kNN-SA. For practical scenarios when SA or kNN-SA are being employed, it is recommended to set the value of k as suggested in Section 6.3 to seek trade-off between confidence improvement and time cost based on the requirements of specific tasks.

6.6 Studying the Influence of ΔB on DA

In section 5.3 we introduced parameter ΔB, i.e., the batch size, for Distribution-based Acquisition (DA), and the CIP distribution model F is retrained on every ΔB newly acquired samples. Next we study the influence of batch size ΔB on the UMCI and time requirements of DA. The results are provided in Figure 8.

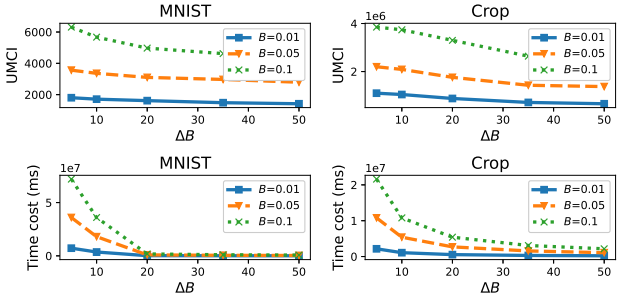
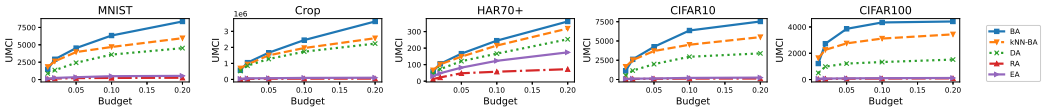


Fig. 8. The Influence of ΔB on DA

As demonstrated in Figure 8, smaller ΔB values lead to higher UMCI and incur higher acquisition time; in the experiments conducted in this work, we observe a good balance between UMCI and time cost when ΔB falls in the range $[20, 35]$. With a smaller ΔB, more frequent retraining of F is triggered and thus the acquisition process can quickly adapt to changes in $\mathcal{T}$, which also results in higher retraining cost. The value of ΔB in other scenarios can thus be adjusted accordingly to provide higher priority to UMCI or efficiency, and one way to choose ΔB is to start from a small value (e.g., 10) and gradually increase ΔB until an acceptable retraining cost has been reached.

6.7 BA, kNN-BA, DA vs. Baselines

In this section, we compare the methods proposed in this work with the two baselines introduced in Section 6. Since no CIP updates are involved in the acquisition process of Random Acquisition (RA) and Equality-oriented Acquisition (EA), we only compare them with BA, kNN-BA and DA with $\Delta B = B$, and exclude SA and kNN-SA. We present the comparison results in Figure 9. Note that the time in Figure 9(b) is in logarithmic scale.



(a) Comparison of UMCI

(b) Comparison of Time Cost

Fig. 9. Comparisons with Baselines

As is clear from the figure, BA achieves the highest UMCI, followed by kNN-BA and DA, and RA and PA lead to the lowest UMCI. In terms of acquisition time, RA and PA are significantly more efficient than the methods proposed in this work, as they require little to no calculation of the pair-wise distances between samples in $\mathcal{T}$, $\mathcal{E}$ and $\mathcal{D}$; DA is slightly faster than kNN-BA, and BA is the slowest. DA reduces the acquisition time cost by training F in advance, which can directly estimate the CIP of new samples, to avoid the expensive calculation of CIP in Equation 4. Compared to BA, kNN-BA limits the acquisition process to a smaller candidate set and thus incurs lower acquisition overhead, as detailed in Section 4.4. Note that the performance of DA is expected to improve when utilized jointly with well calibrated F; this could be an interesting direction for future work.

6.8 Studying the Influence on Model Accuracy

In this section we expand our study and investigate the influence of the samples thus acquired to the accuracy of model $\mathcal{M}$. More specifically, let *acc* be the accuracy of model $\mathcal{M}$ on the evaluation set $\mathcal{E}$ before the acquisition process, and *acc*′ be the accuracy of $\mathcal{M}$ after being retrained on $\mathcal{T}$ and the newly acquired samples. In Figure 10 we demonstrate the accuracy change (i.e., *acc*′ − *acc*) when varying the number of samples acquired.

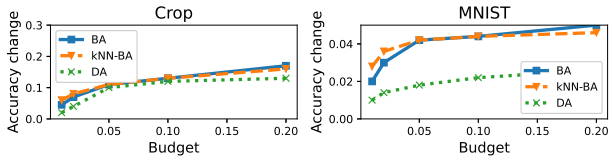


Fig. 10. Accuracy Changes

We observe a positive correlation between accuracy and confidence in our study, as shown in Figure 10. This correlation holds true across different datasets and acquisition methods. Specifically, the samples acquired with the main objective of improving confidence also contribute to improving model accuracy. It is important to note that this positive correlation is observed when the training data ($\mathcal{D}$) and the test data ($\mathcal{T}$) follow the same distribution, as mentioned in Section 1 and Section 6.1.

However, it is worth mentioning that when the training data ($\mathcal{D}$) and the test data ($\mathcal{T}$) follow different distributions, the relationship between confidence and accuracy may exhibit more complex behavior. Exploring this relationship in such cases presents an interesting direction for future research. In particular, ensuring that model accuracy does not decrease after confidence-oriented data acquisition when $\mathcal{D}$ follows an arbitrary distribution poses a challenging optimization problem.

6.9 Studying the Influence on Model Reliability

We have demonstrated that the proposed methods instruct data acquisition to improve model confidence, and it is known that confidence is closely related to the reliability of Machine Learning models [12, 22]. In this set of experiments, we verify whether the data acquired following our proposed algorithms can effectively improve the ultimate model reliability, adopting the ML system reliability quantification strategy designed in [13]. Note that the strategy in [13] consists of a comprehensive list of metrics describing ML system reliability from different perspectives, out of which we select the subset of metrics relevant to the target setting of the paper (i.e., high variance scenario with independent identically distributed samples), namely *in-distribution performance* and *calibration*. Following the design in [13], we calculate the values corresponding to each of the metrics and use the average as the reliability score. The changes in reliability score after the data acquisition process (with $B = 0.2 * |\mathcal{D}|$) are reported in Figure 11.

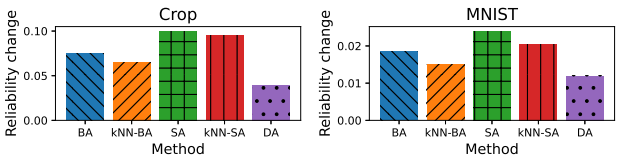


Fig. 11. Reliability Changes

As depicted in Figure 11, the model reliability increases after the data acquisition process, and the observation holds true across different datasets and acquisition methods, proving that the confidence gains resulted from the data acquisition process can effectively translate into model reliability improvements.

We note that the results in Figure 11 are mainly used to showcase the effectiveness of the proposed methods in boosting model reliability in the problem setting studied in this work. Comprehensive study and more insights of the relationship between model confidence and system reliability are out of the scope of the paper and can be located in related works [12, 13, 22].

6.10 Guidance to Choose Data Acquisition Algorithms

Based on the experimental results, we provide suggestions regarding which data acquisition algorithm to choose in different scenarios to maximize confidence improvement, giving particular consideration to time constraints. While it is difficult to directly translate a concrete time constraint (such as completing acquisition within 2 minutes) into actionable rules as the actual time cost heavily depends on the problem instance, the users can perform pre-acquisition probing to assist algorithm selection. More specifically, when dealing with a new task, the users can first conduct a data acquisition task with a small fraction of the budget (e.g., 0.1^*B) to obtain a basic understanding as to whether the given time constraint (e.g., 2 minutes, or 1 hour) for the particular problem instance is relaxed, moderately restrictive, or highly restrictive, and then adopt the following guidance to choose the most suitable algorithm.

(1) Based on the observation in Section 6.2, with relaxed time constraint, use Sequential Acquisition;
(2) Based on the observation in Section 6.5, with moderately restrictive time constraint and when an index on the data pool exists, use kNN-SA;
(3) Based on the observations in Section 6.2 and Section 6.4, with moderately restrictive time constraint and when an index on the data pool is not available, use Bulk Acquisition;
(4) Based on the observation in Section 6.7, with highly restrictive time constraint, use kNN-BA;
(5) Based on the analysis in Section 5.3 and the observation in Section 6.7, when the coordinates of samples in the data pool are not accessible and thus direct calculation of CIP is not possible, use Distribution-based Acquisition;
(6) For real time data acquisition, such as in a stream setting when frequent decisions about whether to acquire an incoming sample need to be made, use Distribution-based Acquisition.

We note that users can choose the parameters to use for each algorithm, based on the observations and detailed analyses in the corresponding experiment sections.

7 RELATED WORK

Data Acquisition. The study of this work falls into the area of data acquisition, which is to acquire data based on a given objective [8, 10, 25, 28, 36, 49]. Li et al. [28] study the problem of acquiring new samples from a data market using queries, with improving model accuracy as the objective, and propose two approaches to generate most promising queries so as to maximize the accuracy improvement. Chai et al. [8] consider the problem of acquiring data from sources in the wild, and devise an end-to-end solution collect and cluster data, as well as identify samples from each cluster that are more useful than others. Chen et al. [10] investigate the problem of acquiring data from diverse sources to enhance the precision of linear statistical estimators. Their study focuses on linear statistical methods, wherein specific guarantees pertain to the optimal approaches, in cases where data providers abstain from making their data available, resulting in fluctuating data costs.

Data Market. Data market is a platform where participants can share data [2, 6, 9, 17, 29, 35]. Fernandez et al. [17] present their vision for the design of platforms to support data markets, followed by various market models, protocols, and algorithms to incentivize the creation of datasharing markets [7]. Asudeh et al. [4] explore the challenge of providing cost-effective and distribution-aware query answering in data market, and envision a unified query answering framework which

integrates data from different sources in a data market and builds various data views for more efficient query answering. Zhao et al. [50] address the budget allocation and revenue allocation problems in data markets and develop a linear-time algorithm to tackle the two allocation problems simultaneously, with theoretical guarantees regarding the efficiency of the budget utilization. The proposal in this work could greatly benefit data sharing in data markets, especially for data buyers with the objective of improving model confidence.

Active Learning. The problem studied in this work share a similar setting with Active Learning, which is concerned with acquiring labels of a subset of unlabeled samples to improve the performance of machine learning models [3, 40, 43]. It has been applied to solve various problems in data management [34]. In a typical setting for active learning, we have access to the coordinates of new data and the main challenge is to identify samples that are more likely to boost model performance, and various approaches have been proposed to quantify the utility of unlabeled samples [31], such as Heterogeneity-based metrics, Representative-based metrics, Performance-based metrics. In contrast, in this work we deal with the scenario where the there exists a well-established utility measure (i.e., model confidence), and the focus is to efficiently identify samples with high utility.

Model Confidence. Model confidence (or uncertainty) is an important aspect of Machine Learning model performance and is very critical for certain types of application scenarios such as health care and autonomous driving, and existing works mainly strive to provide accurate and well-calibrated confidence measure [12, 19, 20, 22, 48]. Jiang et al. [22] propose a two-step framework to quantify the trustworthiness of classifier which first extracts high-density datasets for each class and then compute the trust score for each evaluation sample based on the distance from the sample to the nearest class different from the predicted class and the distance to the predicted class. Chouraqui et al. [12] put forward a geometric-based approach for uncertainty estimation by using the distance of a given evaluation sample from the existing training samples as a signal for estimating uncertainty and then calibrate that signal using standard post-hoc calibration techniques. [1] provides a comprehensive literature review of uncertainty/confidence quantification for deep learning. However, we are not aware of any existing work taking a well-established model confidence metric and solve the problem of confidence-oriented data acquisition, as this paper does.

8 CONCLUSION

In this work, we have investigated the important yet challenging task of model confidence-oriented data acquisition in this work. We have grounded our work in well-established model confidence measures, and devised two approaches to derive the optimal solution in certain cases, BA and SA. To optimize the efficiency of the acquisition process, we have designed two lightweight approximate variants of BA and SA, named kNN-BA and kNN-SA, with targeted search space. Distribution-based Acquisition has been proposed to provide a generic solution to the task which makes minimal assumption regarding the way the data acquisition is conducted. We have empirically studied the properties and performance of the proposed methods using various datasets and models, and under different experiment settings, as well as demonstrated their superiority over other applicable baselines.

ACKNOWLEDGEMENT

This work was supported in part by NSERC Discovery Grants and York University's Catalyzing Interdisciplinary Research Clusters (CIRC) program for the Research Cluster on Data Economy. We would like to thank the anonymous reviewers for their valuable comments that have helped improve this paper.

REFERENCES

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion* 76 (2021), 243–297.

[2] Anish Agarwal, Munther A. Dahleh, and Tuhin Sarkar. 2019. A Marketplace for Data: An Algorithmic Solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*. ACM, 701–726.

[3] Charu C Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and S Yu Philip. 2014. Active learning: A survey. In *Data classification*. Chapman and Hall/CRC, 599–634.

[4] Abolfazl Asudeh and Fatemeh Nargesian. 2022. Towards distribution-aware query answering in data markets. *Proceedings of the VLDB Endowment* 15, 11 (2022), 3137–3144.

[5] Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*. 26–33.

[6] Raul Castro Fernandez. 2022. Protecting Data Markets from Strategic Buyers. In *Proceedings of the 2022 International Conference on Management of Data*. 1755–1769.

[7] Raul Castro Fernandez. 2023. Data-Sharing Markets: Model, Protocol, and Algorithms to Incentivize the Formation of Data-Sharing Consortia. *Proceedings of the ACM on Management of Data* 1, 2 (2023), 1–25.

[8] Chengliang Chai, Jiabin Liu, Nan Tang, Guoliang Li, and Yuyu Luo. 2022. Selective data acquisition in the wild for model charging. *Proceedings of the VLDB Endowment* 15, 7 (2022), 1466–1478.

[9] Lingjiao Chen, Paraschos Koutris, and Arun Kumar. 2019. Towards Model-based Pricing for Machine Learning in a Data Marketplace. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*. ACM, 1535–1552.

[10] Yiling Chen, Nicole Immorlica, Brendan Lucier, Vasilis Syrgkanis, and Juba Ziani. 2018. Optimal Data Acquisition for Statistical Estimation. In *Proceedings of the 2018 ACM Conference on Economics and Computation, Ithaca, NY, USA, June 18-22, 2018*. ACM, 27–44.

[11] Nadiia Chepurko, Ryan Marcus, Emanuel Zgraggen, Raul Castro Fernandez, Tim Kraska, and David Karger. 2020. ARDA: Automatic Relational Data Augmentation for Machine Learning. *Proc. VLDB Endow.* 13, 9 (2020), 1373–1387.

[12] Gabriella Chouraqui, Liron Cohen, Gil Einziger, and Liel Leman. 2022. A geometric method for improved uncertainty estimation in real-time. In *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands (Proceedings of Machine Learning Research, Vol. 180)*, James Cussens and Kun Zhang (Eds.). PMLR, 422–432. https://proceedings.mlr.press/v180/chouraqui22a.html

[13] Anthony Corso, David Karamadian, Romeo Valentin, Mary Cooper, and Mykel J Kochenderfer. 2023. A Holistic Assessment of the Reliability of Machine Learning Systems. *arXiv preprint arXiv:2307.10586* (2023).

[14] Dawex. 2023. *Dawex*. https://www.dawex.com/en/

[15] Li Deng. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine* 29, 6 (2012), 141–142.

[16] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[17] Raul Castro Fernandez, Pranav Subramaniam, and Michael J Franklin. 2020. Data market platforms: Trading data assets to solve data problems. *Proceedings of the VLDB Endowment* 13, 12 (2020), 1933–1947.

[18] Stuart Geman, Elie Bienenstock, and René Doursat. 1992. Neural Networks and the Bias/Variance Dilemma. *Neural Comput.* 4, 1 (1992), 1–58.

[19] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.

[20] Chirag Gupta and Aaditya Ramdas. 2021. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International Conference on Machine Learning*. PMLR, 3942–3952.

[21] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE intelligent systems* 24, 2 (2009), 8–12.

[22] Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya R. Gupta. 2018. To Trust Or Not To Trust A Classifier. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 5546–5557. https://proceedings.neurips.cc/paper/2018/hash/7180cffd6a8e829dacfc2a31b3f72ece-Abstract.html

[23] Iman Khosravi and Seyed Kazem Alavipanah. 2019. A random forest-based framework for crop mapping using temporal, spectral, textural and polarimetric observations. *International Journal of Remote Sensing* 40, 18 (2019), 7221–7251.

[24] Ron Kohavi, David H Wolpert, et al. 1996. Bias plus variance decomposition for zero-one loss functions. In *ICML*, Vol. 96. Citeseer, 275–283.

[25] Yuqing Kong, Grant Schoenebeck, Biaoshuai Tao, and Fang-Yi Yu. 2020. Information Elicitation Mechanisms for Statistical Estimation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2095–2102.
[26] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
[27] Wendi Li, Xiao Yang, Weiqing Liu, Yingce Xia, and Jiang Bian. 2022. DDG-DA: Data Distribution Generation for Predictable Concept Drift Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 4092–4100.
[28] Yifan Li, Xiaohui Yu, and Nick Koudas. 2021. Data Acquisition for Improving Machine Learning Models. *Proc. VLDB Endow.* 14, 10 (2021), 1832–1844. https://doi.org/10.14778/3467861.3467872
[29] Jinfei Liu. 2020. Dealer: End-to-End Data Marketplace with Model-based Pricing. arXiv:2003.13103 [cs.DB]
[30] Aleksej Logacjov and Astrid Ustad. 2023. HAR70+. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5CW3D.
[31] R. Lomasky, C. E. Brodley, M. Aernecke, D. Walt, and M. Friedl. 2007. Active Class Selection. In *Machine Learning: ECML 2007*.
[32] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. 2018. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering* 31, 12 (2018), 2346–2363.
[33] Lin Ma, Bailu Ding, Sudipto Das, and Adith Swaminathan. 2020. Active learning for ML enhanced database systems. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 175–191.
[34] Lin Ma, Bailu Ding, Sudipto Das, and Adith Swaminathan. 2020. Active Learning for ML Enhanced Database Systems. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. Association for Computing Machinery, New York, NY, USA, 175–191.
[35] Sameer Mehta, Milind Dawande, Ganesh Janakiraman, and Vijay S. Mookerjee. 2019. How to Sell a Dataset?: Pricing Policies for Data Monetization. In *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*. ACM, 679.
[36] Fatemeh Nargesian, Abolfazl Asudeh, and HV Jagadish. 2022. Responsible Data Integration: Next-generation Challenges. In *Proceedings of the 2022 International Conference on Management of Data*. 2458–2464.
[37] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions—I. *Mathematical programming* 14 (1978), 265–294.
[38] Arild Nøkland and Lars Hiller Eidnes. 2019. Training Neural Networks with Local Error Signals. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 4839–4850.
[39] Brent Pedersen Matthias Adam Stewart Sean Gillies, Howard Butler. 2023. *R-Tree Implementation.* https://github.com/Toblerity/rtree
[40] Burr Settles. 2009. Active learning literature survey. (2009).
[41] Vraj Shah, Arun Kumar, and Xiaojin Zhu. 2017. Are Key-Foreign Key Joins Safe to Avoid when Learning High-Capacity Classifiers? *Proc. VLDB Endow.* 11, 3 (2017), 366–379.
[42] Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. 2020. Deep active learning: Unified and principled method for query and training. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1308–1318.
[43] Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. 2020. Deep Active Learning: Unified and Principled Method for Query and Training. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. 1308–1318.
[44] Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, and Dongsheng Li. 2021. Energy-based open-world uncertainty modeling for confidence calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9302–9311.
[45] WorldQuant. 2023. *WorldQuant.* https://data.worldquant.com
[46] Xignite. 2023. *xignite.* https://aws.amazon.com/solutionspace/financial-services/solutions/xignite-market-data-cloud-platform/
[47] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. 2020. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*. PMLR, 10767–10777.
[48] Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. 2020. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*. PMLR, 11117–11128.
[49] Meng Zhang, Ahmed Arafa, Ermin Wei, and Randall A. Berry. 2020. Optimal and Quantized Mechanism Design for Fresh Data Acquisition. arXiv:2006.15751
[50] Boxin Zhao, Boxiang Lyu, Raul Castro Fernandez, and Mladen Kolar. 2023. Addressing Budget Allocation and Revenue Allocation in Data Market Environments Using an Adaptive Sampling Algorithm. *arXiv preprint arXiv:2306.02543* (2023).

Received October 2023; revised January 2024; accepted March 2024