Active Data Acquisition - An Information Theoretic Approach

An Vuong
School of Electrical and Computer Engineering
Oregon State University
Corvallis, OR, 97331
Email: vuonga2@oregonstate.edu

Anthony Q Nguyen
Email: anth.nguyen2357@gmail.com

Thinh Nguyen
School of Electrical and Computer Engineering
Oregon State University
Corvallis, 97331
Email: thinhq@eecs.oregonstate.edu

***Abstract*—Insufficient data acquisition can lead to information loss, which may result in inaccurate classifications when using the database for supervised learning. On the other hand, expanding data collection comes with its own set of challenges, such as higher costs for data acquisition, storage, and extraction, as well as the risk of gathering and processing irrelevant data for specific purposes. To that end, we propose an information theoretic framework to address the problem of data acquisition and classification under resource constraints. An information-theoretic perspective offers a valuable approach to balancing specificity and generality in the active data acquisition problem. Instead of merely optimizing data acquisition to improve the accuracy of a particular learning algorithm, our goal is to collect data with the most relevant information for a broad range of potential tasks, under the resource constraints, and without involving specific learning algorithm in the process. Specifically, we use mutual information to quantify the relevance of the data for these tasks since it is more robust than other measures, and is not tied to any particular types of learning algorithms.**

Keywords: Entropy, mutual information, convexity.

I. Introduction

Modern data science discoveries depends on analyzing a large amounts of data, the more the better. However, as data acquisition methods can introduce biases, it's essential to use data acquisition techniques that mitigate biased conclusions in subsequent analyses and experiments. Yet, expanding data collection presents multiple challenges, including increased data acquisition, storage, and extraction costs, along with the potentially acquiring and processing irrelevant data for certain purposes. For instance, while capturing videos at 120 frames per second might be crucial for gaming experience, it proves excessive and costly for tasks like person identification, where 10 frames per second suffices. In particular, to train a deep neural network (DNN) using data consisting of 120 frames per second videos might incur more computational, storage costs, but produces a similar accuracy for a DNN using training dataset consisting of only 10 fps videos. Another example is object identification using electromagnetic wave. In this setting, a radar emits a signal and examines the reflected signal to identify an object. Some objects may only reflect waves at certain a frequency and power range, as such it is imperative to optimize the transmitted signals in certain frequency range and power.

These examples highlight the need to tailor the data acquisition to suit the specific goals or tasks under resource constraints. However, a goal-specific data acquisition approach can yield an overly specialized dataset, potentially omit important information necessary for similar tasks. Ideally, we want a sufficiently rich dataset that contain information relevant to the unspecified yet similar tasks. For example, a dataset containing images of animals can be more beneficial than one limited to just cats and dogs.

The goal-specific data acquisition approach can also be examined through the lens of the discriminative and generative learning framework. Discriminative learning features prominently in classifiers which learn a task-specific model, in particularly the conditional distribution $p(C|x)$, i.e., the probability of an object C present given the observation x. On the other hand, the generative learning aims to learn the joint distribution $p(C, x)$. Generally, the dataset for discriminative is smaller, and the models are easier to train compared to generative models. On the other hand, the generative model, trained on richer dataset, are more versatile at different tasks. Therefore, the data acquisition for the discriminative and generative models can be optimized to create a dataset that balances between specificity and generality.

Information theoretic perspective provides a useful way toward the balance between specificity and generality for the active data acquisition problem. Rather than focusing solely on optimizing data acquisition under resource constraints to enhance the accuracy of a specific task, our aim is to gather data containing the most relevant information for a wide range of potential tasks, without involving the learning algorithms in the process. Specifically, the mutual information is used to represent the relevant information for the set of tasks. It is much stronger than correlation and other measures, and is not specific to any type of learning algorithms.

II. Related Work

We first provide a brief review of the literature on on target detection and object identification since it is an important application of our proposed framework. The field of target identification is well studied in many scientific and engineering disciplines. For many years, target identification techniques

979-8-3315-2096-0/25/$31.00 ©2025 IEEE

were model-based driven due to its mathematical elegance and efficiency, but also due to the lack of data required for training. The model-based approach incorporates prior knowledge about a problem based on either physical laws or well-established intuitions to determine a few model parameters for accurate object identification. For this reason, model-based approach is highly efficient in many settings e.g., radar systems [1], [2] that can be accurately captured through a few parameters. On the other hand, with the abundance of available data, object identification in computer vision [3], [4] and radar-based target recognition [5] have transitioned from model-based approaches to supervised learning, where algorithms are trained directly from datasets. In contrast, our work does not provide a specific target detection algorithm. Rather, our work is focused on getting the optimal information to be used in subsequent classification algorithm. Our work is also related to active learning [6], [7], where the objective is to select specific samples for labeling to maximize the performance of a learning algorithm, particularly when data collection is costly. However, unlike most of active learning, our approach does not optimize for a particular learning algorithm but instead focuses on maximizing mutual information which is similar to that of [8]. Additionally, Shayovitz et al. recently introduced a universal active learning framework that minimizes conditional information [9]. While this approach is similar to our method of maximizing mutual information, our focus differs in specific goals.

III. Problem Description

A. Motivated Scenario

Object Identification: In this setup, an object detector, e.g., radar emits a signal $x \in \mathbf{R}^m$ towards an object $c \in \mathbf{C} = \{c_0, c_1, \ldots, c_{k-1}\}$, where $\mathbf{C}$ is a set denoting k possible objects. Depending on specific object c, a distinct signature signal $y \in \mathbf{R}^n$ is reflected back to the detector. The detector's objective is to identify the object c based on the received signature y.

In classical detection systems, the identification algorithm relies on physical modeling of various objects and their reflected properties to predict the target based on its signature. In contrast, with supervised learning techniques such as Deep Neural Networks (DNNs), a classifier: $f(y) : \mathbf{R}^n \longrightarrow C$ is learnt using the training samples consisting of a large number of feature-label pairs, i.e., (y, c). Using this approach, a large number of samples (y, c) must be collected first before the training can take place. In a typical systems, x is chosen using some prior knowledge about the targets and the environment. As a result, y, a function of x and c is also fixed. In many systems, there are associated significant costs with choosing x, e.g., larger $x^T x$ incurs larger transmit power during the data acquisition process. There are also costs associated with storing y, e.g., high dimensional y can lead to larger storage, potentially resulting in longer training time and energy to obtain a classifier. Furthermore, the performance of the classifier depends on the chosen x, and can be degraded if x is not properly chosen.

To that end, we propose to an information theoretic formulation which takes into account the costs associated with x and y. Specifically, we show how to select x so that the received signal y contains the most relevant information about an object c, and all the cost constraints are satisfied. Furthermore, the resulted dataset is not tuned for any particular classifier. Rather, the proposed framework produce a database that can be used for any subsequent classifier, i.e., the training database is decoupled from the classifiers.

B. Information Theoretic Formulation

The key to decouple the database from the classifier is to use the mutual information, rather than minimizing the loss function of a classifier. Let X be the random variable representing the sensing signal or action x, with distribution $p(x)$. Let C be the target random variable. C corresponds to the types of objects. Let Y be the random variable corresponding to the received signal with distribution $p(y)$. Y models the reflected signals. For clarity, assume that X, Y, C are discrete random variables.

To account for uncertainty in signal y, the conditional distribution $p(y|x, c)$ models the measured signal y when a signal x is applied to target c. Suppose n measurements are made by transmitting repeatedly the signals $x_1, x_2, \ldots, x_n$ and yield $y_1, y_2, \ldots, y_n$ as the received signals. Let $x^n = (x_1, x_2, \ldots, x_n)$, $y^n = (y_1, y_2, \ldots, y_n)$, and let $p(x^n)$ and $p(y^n)$ be their marginal distributions. The goal is to choose the distribution $p(x^n)$ that maximizes the mutual information $I(X^n, Y^n; C)$, where

$$I(X^n, Y^n; C) \triangleq I(X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n; C). \quad (1)$$

Let $p(x^n, y^n)$ be the joint distributions of X^n and Y^n, then

$$I(X^n, Y^n; C) \triangleq \sum_{x^n, y^n, c} p(x^n, y^n, c) \log \frac{p(x^n, y^n, c)}{p(x^n, y^n)p(c)} \quad (2)$$

Maximizing $I(X^n, Y^n; C)$ assume that both sensing signal X^n and the corresponding received signal Y^n are used to get information about the target C, i.e., both X^n and Y^n are available in the training data.

On the other hand, when X^n are not available, one wants to maximize the mutual information between the received signals Y^n and the target C. Formally, $p(x^n)$ is chosen to maximize

$$I(Y^n; C) \triangleq \sum_{y^n, c} p(y^n, c) \log \frac{p(y^n, c)}{p(y^n)p(c)} \quad (3)$$

This scenario is applicable to the data acquisition process where only Y^n are stored in the database, and only Y^n is used for training a classifier.

Also note that $p(y|x, c)$ can be viewed as a communication channel matrix. This channel is memoryless (i.e., $p(y_i|x_i, y_j, x_j, c)) = p(y_i|x_i, c), j \neq i$), then maximizing $I(X^n, Y^n; C)$ and $I(Y^n; C)$ over $p(x^n)$ is equivalent to maximizing $I(X, Y; C)$ and $I(Y; C)$ over $p(x)$. To that end, we propose two following optimization problems for the object

identification problem:

Problem P1:

$$\begin{aligned} \max_{p(x)} \quad & I(X,Y;C) \\ \text{s.t.} \quad & g_i(p(x)) \le 0, i = 1,\dots,N \\ & p(x) \ge 0,\ p(x) \le 1,\ \sum_x p(x) = 1. \end{aligned} \tag{4}$$

Problem P2:

$$\begin{aligned} \max_{p(x)} \quad & I(Y;C) \\ \text{s.t.} \quad & g_i(p(x)) \le 0, i = 1,\dots,N \\ & p(x) \ge 0,\ p(x) \le 1,\ \sum_x p(x) = 1. \end{aligned} \tag{5}$$

In both **P1** and **P2**, $g_i(p(x))$ are the given constraints modeling the costs associated the acquisition while the constraints on $p(x)$ enforce the validity of a distribution. Assume that $g_i(p(x))$ are convex functions with respect to $p(x)$, which is typical in real-world scenarios. For example, consider a scenario where the average transmit power must not exceed a certain threshold, expressed as $g(p(x)) = \mathbb{E}[X^2] < C$. In such cases, the problem exhibits a special structure that can be exploited to speed up the algorithm for finding the optimal solution, as will be discussed shortly.

IV. Solution Approach

A. Solution Characterization

Theorem 1. *Let X, Y, C be random variables denoting the sensing signal, the received signal, and the object. For a given $p(c)$, $p(y|x,c)$ as memoryless channel, $X \perp\!\!\!\perp C$, and $g_i(p(x))$ are linear, then* **P1** *is a linear programming problem. Furthermore, $p^*(x)$ must be one of the extreme points (vertices) from by intersections among the constraints. When there is no constraint $g_i(p(x))$, $p^*(x)$ must lie on one of the vertices of the probability simplex, i.e., the optimal solution is to send $X = x^*$ for some constant x^*.*

Proof. For clarity, assume X, Y, and C are discrete random variables.

$$I(X,Y;C) = \sum_{x,y,c} p(x,y,c) \log \frac{p(x,y,c)}{p(x,y)p(c)} \tag{6}$$

$$= \sum_{x,y,c} p(y|x,c)p(x)p(c) \log \frac{p(y|x,c)p(x)p(c)}{p(c)p(x)p(y|x)}$$

$$= \sum_{x,y,c} p(y|x,c)p(x)p(c) \log \frac{p(y|x,c)}{p(y|x)}. \tag{7}$$

Since $p(y|x) = \sum_c p(y|x,c)p(c)$ which is a constant. Thus, Eq. (7) is a linear with $p(x)$. Thus, maximizing $I(X,Y;C)$ with respect to $p(x)$ is a linear programming problem subject to the constraints on a valid distribution and linear $g_i(p(x))$.

From the well-known linear programming result [10], if $I(X,Y;C)$ is not a constant function then $p^*(x)$ must be at an extreme point formed by the intersections between the probability simplex and the linear constraints.

Similar proof can be carried out for continuous random variables X. Furthermore, without any constraint $g_i(p(x))$, $p^*(x) = \delta(x - x^*)$ for some x^*.

□

We now turn our attention to determine $p^*(x)$ that maximizes $I(Y;C)$. In this scenario, X is not available during training to obtain a classifier. We have the following theorem for this case.

Theorem 2. *Let X, Y, C be random variables denoting the sensing signal/action, the received signal, and the target. For a given $p(c)$, $p(y|x,c)$ as memoryless channel, $X \perp\!\!\!\perp C$, and $g_i(p(x))$ are convex functions, then* **P2** *is a convex maximization problem. Furthermore, $p^*(x)$ must be one of the extreme points (vertices) from by intersections of the given convex constraints. When there is no constraint $g_i(p(x))$, $p^*(x)$ must lie on one of the vertices of the probability simplex, i.e., the optimal solution is to send $X = x^*$ for some constant x^*.*

Proof. As before, assume that X, Y, C are discrete random variables. From a well-known result in information theory[11], for a fixed $p(c)$, $I(C;Y)$ is a convex function with respect to $p(y|c)$ where $p(y|c)$ is considered as a channel matrix, C as input and Y as output. Since $p(y|c) = \sum_x p(y|c,x)p(x) = T(p(x))$, where T is a linear map specified by the given $p(y|x,c)$. For any convex function $f(z)$ in z, let $z = T(w)$ where T is any affine map, then $f(T(w))$ is also convex in w [12]. A quick proof is as follows: If $f(z)$ is a convex function, then if and only if for any z_0, there exists an affine map $g(z)$ such that $g(z_0) = f(z_0)$ and $g(z) \le f(z)$ for all z. Applying an affine map T, $z = T(w)$ in the inequality, and note that $g(T(w))$ remains an affine map since the composite of two affine maps $g \circ T$ is a new affine map for which the inequality hold for all w. Therefore $f(w)$ is also convex in w. Consequently, since $g_i(p(x))$ are convex functions, then finding the optimal $p^*(x)$ that maximizes $I(Y;C)$ is a convex maximization problem. The proof for continuous random variables can be carried out in the same manner.

From the well-known optimization result [13], the solution to the convex maximization problem must be one of the extreme points formed by the intersections of the constraints.

□

B. Illustrative Examples

Example IV.1. (Additive Channel)
Let $X \sim Bern(p)$ represent the active signal X. Let $C \sim Bern(q)$ represent the two possible objects: $C = 0$ and $C = 1$. Let

$$Z_0 \sim Bern(q_0)$$
$$Z_1 \sim Bern(q_1).$$

If a signal X is transmitted, then the received signal is

$$Y = \begin{cases} X + Z_0 & \text{if object 0 is present} \\ X + Z_1 & \text{if object 1 is present} \end{cases}$$

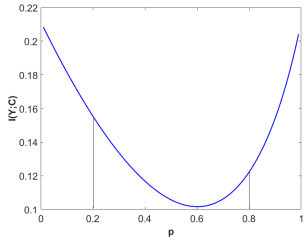

Figure 1. $I(Y;C)$ of additive channel; $a = 0.2$ and $b = 0.8$.

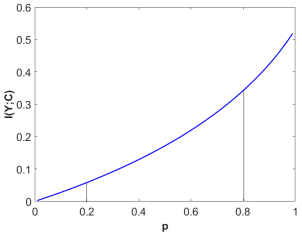

Figure 2. $I(Y:C)$ of multiplicative channel; $a = 0.2$ and $b = 0.8$.

The goal is to determine $p^*(x)$ that maximizes $I(X,Y;C)$ subject to $a \le \mathbb{E}[X^2] \le b$,, for some constants $a < b$.

Assuming C is independent with X, then

$$\begin{aligned} I(C;X,Y) &= H(Y,X) - H(X,Y|C) \\ &= H(Y|X) - H(Y|X,C=0)(1-q) \\ &\quad - H(Y|X,C=1)q \\ &= H(Y|X) - H(Z_0)(1-q) - H(Z_1)q \\ &= H(Y|X=0)(1-p) + H(Y|X=1)p \\ &\quad - H(Z_0)(1-q) - H(Z_1)q \end{aligned}$$

$$\begin{aligned} H(Y|X=1) &= H(Y|X=0) = H(q_0 - qq_0 + qq_1) \\ I(C;X,Y) &= H(Y|X=0)(1-p) + H(Y|X=1)p \\ &\quad - H(Z_0)q - H(Z_1)(1-q) \\ &= H(q_0 + qq_0 - qq_1) - H(q_0)(1-q) - H(q_1)q, \end{aligned}$$

where $H(r) \triangleq -r\log r - (1-r)\log(1-r)$ denotes the entropy of a binary random variable. This is a special case where the mutual information $I(C;X,Y)$ is constant regardless of $p(x)$. Hence, any distribution $p(x)$ would yield the same mutual information. On the other hand, to satisfy the constraint $a \le \mathbb{E}[X^2] \le b$, we can choose any $p^*(x) = (1-p, p)$ where $a \le p \le b$. Also,

$$\begin{aligned} &I(C;Y) = H(q) \\ &- (1-p)(1 - q_0 + q_0 q - q_1 q) \times H\Big(\frac{1 - q - q_0 + q_0 q}{1 - q_0 + q_0 q - q_1 q}\Big) \\ &- (p + q_0 - 2pq_0)(1-q) + (p + q_1 - 2pq_1)q) \\ &\times H\Big(\frac{(p + q_0 - 2pq_0)(1-q)}{(p + q_0 - 2pq_0)(1-q) + (p + q_1 - 2pq_1)q}\Big) \\ &- p(q_0 - q_0 q + q_1 q) \times H\Big(\frac{q_0 - q_0 q}{q_0 - q_0 q + q_1 q}\Big). \end{aligned}$$

Fig. 1 shows $I(Y;C)$ with varying p, where $q_0 = 0.4, q_1 = 0.9, q = 0.6, a = 0.2, b = 0.8$. The optimal solution occurs at $p^* = 0.2$ which corresponds to maximum $I(Y;C) = 0.1556$. Note that $p^*(x) = (0.8, 0.2)$ is an extreme point which confirms Theorem 2.

Example IV.2. (Multiplicative Channel)
Consider the same example IV.1, but using

$$Y = \begin{cases} XZ_0 & \text{if object 0 is present} \\ XZ_1 & \text{if object 1 is present.} \end{cases}$$

The goal is to determine $p^*(x)$ that maximizes $I(X,Y;C)$ subject to $\mathbb{E}[X^2] \le 0.5$.

$$\begin{aligned} I(C;X,Y) &= H(Y,X) - H(X,Y|C) \\ &= H(X) + H(Y|X) - H(X|C) - H(Y|X,C) \\ &= H(Y|X) - H(Y|X,C=0)(1-q) \\ &\quad - H(Y|X,C=1)q \\ &= H(Y|X) - H(XZ_0|X)(1-q) - H(XZ_1|X)q \\ &= H(Y|X) - H(Z_0)p(1-q) - H(Z_1)pq \\ &= H(Y|X=0)(1-p) + H(Y|X=1)p \\ &\quad - p\Big(H(Z_0)(1-q) + H(Z_1)q\Big) \\ &= \Big(H(q) + H(Z_0)(1-q) + H(Z_1)q\Big)p \\ &\quad - p\Big(H(Z_0)(1-q) + H(Z_1)q\Big) = pH(q). \end{aligned}$$

Since $H(q)$ is fixed, the optimal $p^*(x) = (1-p, p) = (0, 1)$, which leads to $\max I(C;X,Y) = H(q)$. This makes sense since sending $X = 0$ yields $Y = 0$ for both objects, resulting in no differentiation between the two objects. Thus, any effort of sending $X = 0$ will be wasted. Also, based on Theorem 1, without any constraint on $g_i(p(x))$, the structure of the optimal distribution must have the form $p^*(x) = \delta(x - x^*)$ for some optimal x^* which agrees with the optimal solution $p^*(x) = (0, 1)$. Also, it can be shown that

$$I(C;Y) = H(p(q_0 - qq_0 + qq_1)) - H(pq_0)(1-q) - H(pq_1)q$$

Fig. 2 shows $I(Y;C)$ with varying p, where $q_0 = 0.1, q_1 = 0.9, q = 0.5, a = 0.2, b = 0.8$. The optimal solution occurs at $p^*(x) = (0.2, 0.8)$ which is also an extreme point.

V. Simulation Results

In this section, we provide a detailed analysis of a scenario involving the identification of two objects c_0, c_1. When the signal $X \in \mathbf{R}^2$ interacts with object c_0, the reflected signal $Y_0 \in \mathbf{R}^2$ undergoes a linear fading with additive noise. In contrast, when X interacts with object c_1, the reflected signal $Y_1 \in \mathbf{R}^2$ undergoes a different linear fading with multiplicative noise. Specifically, we have the following model:

$$Y = \begin{cases} AX + X \odot Z_0 & \text{with probability } q \\ BX + Z_1 & \text{with probability } 1 - q. \end{cases}$$

where $Z_0, Z_1 \in \mathbf{R}^2$ are continuous noises with probability density functions $f_0(z_0)$ and $f_1(z_1)$, respectively. A and B are fading matrices, $\odot$ denotes the element-wise product, q is the probability that object c_0 is present. Let

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} z_0 = \begin{bmatrix} z_{01} \\ z_{02} \end{bmatrix}, z_1 = \begin{bmatrix} z_{11} z_{12} \end{bmatrix}$$
$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}.$$

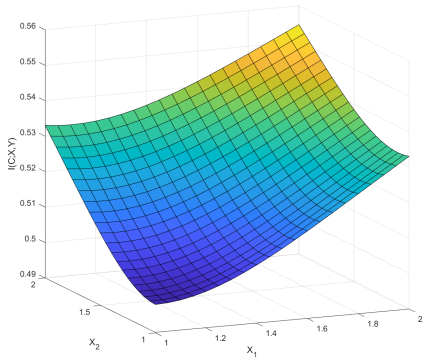

Figure 3. Mutual information $I(C;X,Y)$ varies with x_1 and x_2.

We now demonstrate the maximization of $I(X,Y;C)$ on some bounded region, e.g. $1 \preceq X \preceq 2$. Since

$$I(C;X,Y) = H(C) - H(C|X,Y)$$

and $H(C)$ is constant for a fixed q, maximizing $I(C;X,Y)$ is the same as minimizing $H(C|X,Y)$.

Let

$$z_{01} = \frac{y_1 - a_{11}x_1^* - a_{12}x_2^*}{x_1^*},\ z_{02} = \frac{y_2 - a_{21}x_1^* - a_{22}x_2^*}{x_2^*}$$
$$z_{11} = y_1 - b_{11}x_1^* - b_{12}x_2^*,\ z_{12} = y_1 - b_{21}x_1^* - b_{22}x_2^*,$$

then it can be shown that

$$\min_{f(x)} H(C|X,Y) =$$
$$-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \log\Big(\frac{qf_0(z_0)}{qf_0(z_0) + x_1^*x_2^*(1-q)f_1(z_1)}\Big) qf_0(z_0)\frac{1}{x_1^*x_2^*}dy_1dy_2$$

for some optimal $x^* = (x_1^*, x_2^*)$. The simulation results use the following setup: $q = 0.5$, Z_0 and Z_1 are multivariate Gaussians with zero means and covariance matrices $\Sigma_0 = \begin{bmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}$, $\Sigma_1 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$, $A = B = \begin{bmatrix} 0.4 & 0.2 \\ 0.3 & 0.1 \end{bmatrix}$.

Fig.3 shows the plot of $I(C;X,Y)$ as a function of x_1 and x_2, rather than $f(x)$. This is because the optimal distribution has the form $f(x) = \delta(x - x^*)$ in this case. Given that the feasible region is $1 \preceq X \preceq 2$, the optimal point is $x^* = (2,2)$, corresponding to $I(X,Y;C) = 0.556$. This is intuitive, as larger values of X lead to greater variance in Y due to the multiplicative noise from c_0, which aids in distinguishing between the two objects.

To further illustrate this intuition, Fig. 4 and Fig. 5 display sample plots of (y_1, y_2) for the optimal case $x^* = (2,2)$ and the suboptimal case $x = (1,1)$. Circles represent y samples corresponding to object c_0, while diamonds represent samples from object c_1. Visually, the differences in shapes between the distributions of diamonds and circles is more pronounced in Fig. 4 compared to Fig. 5. This greater distinction suggests that it would be easier to learn a classifier from the data in Fig. 4 than from Fig. 5.

VI. Conclusion

We present an information-theoretic framework to tackle the challenges of data acquisition and classification under resource constraints. This approach provides a balanced solution to the active data acquisition problem by addressing both specificity and generality. Rather than optimizing data collection solely to enhance the accuracy of a specific learning algorithm, our objective is to gather data with the most relevant information for a wide range of potential tasks, all within the constraints of available resources, and without relying on any particular algorithm. We leverage mutual information to measure data relevance, as it is more robust than other metrics and not tied to specific learning algorithms. We also provided simulation results to demonstrate the effectiveness of our approach.

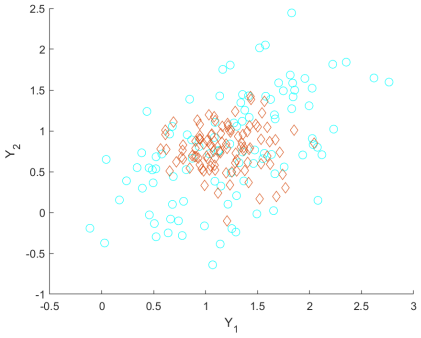

Figure 4. Samples (y_1, y_2) resulted from using the optimal x^*

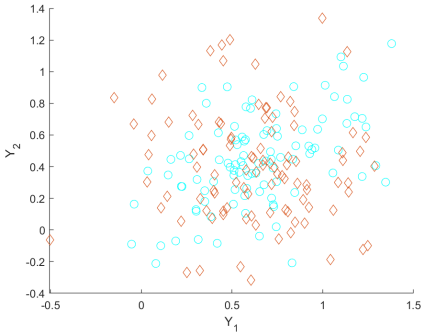

Figure 5. Samples (y_1, y_2) resulted from using the suboptimal optimal x

References

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*, Prentice Hall, 1998.

[2] Sudan Han, Linjie Yan, Yuxuan Zhang, Pia Addabbo, Chengpeng Hao, and Danilo Orlando, "Adaptive radar detection and classification algorithms for multiple coherent signals," *IEEE Transactions on Signal Processing*, vol. 69, pp. 560–572, 2021.

[3] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[4] Shivang Agarwal, Jean Ogier Du Terrail, and Frédéric Jurie, "Recent advances in object detection in the age of deep convolutional neural networks," *arXiv preprint arXiv:1809.03193*, 2018.

[5] Uttam Majumder, Erik Blasch, and David Garren, *Deep Learning for Radar and Communications Automatic Target Recognition*, Artech, 2020.

[6] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang, "A survey of deep active learning," *CoRR*, vol. abs/2009.00236, 2020.

[7] Ozan Sener and Silvio Savarese, "Active learning for convolutional neural networks: A core-set approach," 2018.

[8] Yuhong Guo and Russell Greiner, "Optimistic active-learning using mutual information.," in *IJCAI*, 2007, vol. 7, pp. 823–829.

[9] Shachar Shayovitz and Meir Feder, "Universal active learning via conditional mutual information minimization," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 720–734, 2021.

[10] Vašek Chvátal, *Linear programming*, Macmillan, 1983.

[11] Thomas Cover and Joy Thomas, *Elements of information theory*, Wiley-Interscience, 2 edition, 7 2006.

[12] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 2004.

[13] Philip B Zwart, "Global maximization of a convex function with linear inequality constraints," *Operations Research*, vol. 22, no. 3, pp. 602–609, 1974.