The Label Complexity of Mixed-Initiative Classifier Training

Jina Suh[*] JINSUH@MICROSOFT.COM
Xiaojin Zhu[*†] JERRYZHU@CS.WISC.EDU
Saleema Amershi[*] SAMERSHI@MICROSOFT.COM
[*]Microsoft Research. Redmond, WA, USA [†]University of Wisconsin–Madison. Madison, WI, USA

Abstract

Mixed-initiative classifier training, where the human teacher can choose which items to label or to label items chosen by the computer, has enjoyed empirical success but without a rigorous statistical learning theoretical justification. We analyze the label complexity of a simple mixed-initiative training mechanism using teaching dimension and active learning. We show that mixed-initiative training is advantageous compared to either computer-initiated (represented by active learning) or human-initiated classifier training. The advantage exists across all human teaching abilities, from optimal to completely unhelpful teachers. We further improve classifier training by educating the human teachers. This is done by showing, or explaining, optimal teaching sets to the human teachers. We conduct Mechanical Turk human experiments on two stylistic classifier training tasks to illustrate our approach.

1. Introduction

We study a topic at the intersection of human-computer interaction and statistical learning theory. We contrast three ways to train a classifier with interactive machine learning: **computer-initiated** where an active learning algorithm picks a query x and the human answers its label y, **human-initiated** where the human picks x and y, and **mixed-initiative** where both parties can pick x. Mixed-initiative machine learning has enjoyed success in practice (Wolfman et al., 2001; Fails & Olsen Jr, 2003; Fogarty et al., 2008), but a theoretical account for its effectiveness is lacking.

Our first contribution is a theoretical analysis on the label complexity of mixed-initiative classifier training, expressed by the following intervals: mixed-initiative $[\mathrm{TD}, \mathrm{TD} + \mathrm{AL}]$, computer-initiated AL, human-initiated $[\mathrm{TD}, \infty)$. Here TD is the teaching dimension of the hypothesis space which lower-bounds AL, the corresponding active learning label complexity. These quantities will be defined precisely below. Given that often $\mathrm{TD} \ll \mathrm{AL}$, mixed-initiative training is attractive among the three.

In more detail, we place a human teacher in one of three ability states and provide a guarantee for each state: (**Optimal teacher**) The teacher provides the optimal teaching set. The mixed-initiative label complexity is TD. (**Seed teacher**) When the teacher is not optimal but can provide at least one item per positive region, mixed-initiative training can significantly reduce the active learning label complexity for certain hypothesis spaces. (**Naive teacher**) For all other human teachers, even the completely unhelpful ones, the mixed-initiative label complexity has a fallback guarantee of $\mathrm{TD} + \mathrm{AL} \leq 2\mathrm{AL}$, namely no worse than twice the active learning label complexity.

Our second contribution is a framework for teacher education. The computer can try to move the human to a better "teacher ability state." It does so by showing or explaining example optimal teaching sets to the teacher. Although the computer does not know the target concept, it can automatically compute an example teaching set on *hypothetical target concepts* for the purpose of teacher education.

2. Problem Setting and Notations

We consider the standard learning-theoretic setting for classification. Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{-1, 1\}$ be the label space. It is straightforward to extend our results to multiple classes. We restrict the interactions between the human and the computer: they can interact through items $x \in \mathcal{X}$ and labels $y \in \mathcal{Y}$ but not other means (e.g. the human cannot label features; quantifying such richer interactions is left as future work).

There is a fixed test distribution $\mathcal{P}_{XY}$ on $\mathcal{X} \times \mathcal{Y}$. A classifier or a hypothesis is a measurable function $h : \mathcal{X} \mapsto \mathcal{Y}$. For any h, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, let $1_{h(x) \neq y}$ be the 0-1 loss function

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

which takes the value 1 if $h(x) \neq y$ and 0 otherwise. Define the risk of h as $R(h) = \mathbb{E}_{(x,y)\sim\mathcal{P}_{XY}} 1_{h(x)\neq y}$. Let $\mathcal{H}$ be the hypothesis space, namely the set of hypotheses considered by the computer. Let VC be the VC-dimension of $\mathcal{H}$. Let f^* be the risk minimizer in $\mathcal{H}$: $f^* = \text{argmin}_{h\in\mathcal{H}} R(h)$. We say the problem is realizable if $R(f^*) = 0$; otherwise we denote the minimum risk by $\nu = R(f^*)$. We may now phrase the classifier training task as follows.

Definition 1 (The Classifier Training Task)**.** *The human's task is to make the computer learn, with success probability of at least $1-\delta$, a hypothesis $\hat{h} \in \mathcal{H}$ such that the excess risk $R(\hat{h}) - R(f^*) \leq \epsilon$ for any given $\epsilon \geq 0, \delta \geq 0$.*

We will use label complexity as a proxy for teaching efficiency. Label complexity is mathematically convenient, but ignores potentially different human effort in coming up with training items. We discuss this issue in Section 7.

Definition 2 (Label complexity)**.** *Training achieves label complexity n if for every $\epsilon, \delta \geq 0$, every distribution $\mathcal{P}_{XY}$, and every integer $m \geq n$, the hypothesis $\hat{h}$ learned using m labeled items satisfies $R(\hat{h}) - R(f^*) \leq \epsilon$ with probability of at least $1-\delta$.*

We contrast three interactive machine learning methods:

Computer-initiated: The computer runs active learning until it achieves the (ϵ, δ) guarantee. In iteration i, it adaptively chooses item x_i to query the human based on previous items and labels $(x_1, y_1) \ldots (x_{i-1}, y_{i-1})$. The human acts as a labeling oracle by providing the stochastic label $y_i \sim \mathcal{P}_{Y|X=x_i}$.

Human-initiated: In iteration i the human can choose any item $x_i \in \mathcal{X}$ and the label y_i to show the computer. The human decides how many iterations to teach.

Mixed-initiative: There is an extra mechanism to determine whether each iteration is computer-initiated or human-initiated. In this paper we focus on Algorithm 1, narrated from the computer's perspective. The human can provide up to TD labeled items (x_i, y_i) in the beginning. Then the computer takes over and runs active learning starting from whatever data D the human already provided, and ends when it can achieve the (ϵ, δ) guarantee. During active learning, the human reduces to a labeling oracle.

We define the teacher ability states as follows. An **optimal teacher** manually chooses exactly TD training items and labels to form an optimal teaching set, and does not provide more than necessary items. A **seed teacher** is not optimal, but chooses at least one positive item for each positive region in the input space. The rest are **naive teachers**.

To help build intuition, we will use two tasks as running examples throughout the paper, including for human experiments. These tasks have been extensively studied in active learning (Hanneke, 2014). Our results generalize beyond the two tasks.

Algorithm 1 The Mixed-Initiative Mechanism

```
Data D = ∅
for i = 1 to TD do
  if human no longer wants to lead then
    break;
  else
    human chooses (x_i, y_i)
    append (x_i, y_i) to D
  end if
end for
run active learning starting from D
until completion
```

Example 1 (1D Threshold Task)**.** $\mathcal{X} = [0,1]$. $\mathcal{P}_X = \text{uniform}[0,1]$. *There is a target threshold $\theta^* \in [0,1]$ such that $\mathcal{P}_{Y=1|X=x} = 1$ if $x \geq \theta^*$ and 0 otherwise. $\mathcal{H}$ consists of all threshold classifiers with threshold θ: $\mathcal{H} = \{h_\theta \mid h_\theta(x) = 1 \text{ if } x \geq \theta \text{ and -1 otherwise}, \forall\theta \in [0,1]\}$. This task is realizable with $f^* = h_{\theta^*}$. Furthermore, $R(h_\theta) = |\theta - \theta^*|$ for all $h_\theta \in \mathcal{H}$. To succeed at the task is to make the learned threshold fall in $[\theta^* - \epsilon, \theta^* + \epsilon]$. $VC(\mathcal{H}) = 1$ since $\mathcal{H}$ always puts the positive class on the right and cannot shatter two points.*

Example 2 (1D Interval Task)**.** $\mathcal{X} = [0,1]$. $\mathcal{P}_X = \text{uniform}[0,1]$. *There is a target interval $[a^*, b^*] \subseteq [0,1]$ such that $\mathcal{P}_{Y=1|X=x} = 1$ if $x \in [a^*, b^*]$ and 0 otherwise. $\mathcal{H}$ consists of all interval classifiers: $\mathcal{H} = \{h_{[a,b]} \mid h_{[a,b]} = 1 \text{ if } x \in [a,b] \text{ and -1 otherwise}, \forall [a,b] \subseteq [0,1]\}$. This task is also realizable with $f^* = h_{[a^*,b^*]}$. For all hypotheses, $R(h_{[a,b]}) = |[a,b]\Delta[a^*,b^*]|$, which is the size of the symmetric difference between the two intervals. $VC(\mathcal{H}) = 2$.*

3. Label Complexity Analysis

We now analyze the label complexity to justify mixed-initiative training. The main results are listed in Table 1. These are worst-case analyses, including worst human teacher behaviors under the respective teacher states.

Table 1. Worst-Case Label Complexity

init \state	optimal	seed	naive
mixed	TD	TD + (AL − AL$_B$)	TD + AL
computer	AL	AL	AL
human	TD	∞	∞

3.1. Computer-Initiated Training

We first review existing results on active learning label complexity, which characterize computer-initiated training. In standard active learning analysis, the human teacher is assumed to be noiseless, always providing correct y for any query x. Because the teacher does not choose x, the three

teacher states (optimal, seed, naive teacher) do not apply to computer-initiated training.

There are two observations worth pointing out. The first observation is that AL has a $\log \frac{1}{\epsilon}$ (or worse in the unrealizable case) dependency on the required precision ϵ. Concretely, the following minimax lower bounds on active learning label complexity holds: In the realizable case, (Kulkarni et al., 1993) showed that $\text{AL} \geq \log(1-\delta) + \max\left(\text{VC}, \log \frac{1}{\epsilon}\right)$. In the unrealizable case with minimum risk ν, (Beygelzimer et al., 2009; Hanneke, 2014) showed that $\text{AL} \geq c\left(\frac{\nu^2}{\epsilon^2}\right)\left(\text{VC} + \log \frac{1}{\delta}\right)$. The logarithmic ϵ dependency has been a major theoretical achievement for active learning compared to passive learning's slower $1/\epsilon$ dependency. It is the theoretical justification for using active learning in computer-initiated classifier training.

Hanneke presented an analysis for the two example tasks (Hanneke, 2014). On the 1D threshold task, deterministic binary search is optimal for active learning, with a label complexity $\log \frac{1}{\epsilon}$. This can be understood by noting that every query reduces the version space length by half. The 1D interval task is more nuanced. One deterministic active algorithm proceeds in two phases. In the first "blind search" phase, it queries points on an increasingly dense grid 1/2, 1/4, 3/4, 1/8, 3/8, ... until it encounters a positive point or reaches grid spacing $\leq \epsilon$. This phase succeeds with at most $\text{AL}_B = \frac{2}{\max(b^*-a^*,\epsilon)}$ queries, where we use the subscript B to denote blind search label complexity. In the second phase it performs two binary searches from the positive point in opposite directions to identify the interval boundaries. This phase succeeds with at most $2\log \frac{2}{\epsilon}$ queries. The overall active learning label complexity is therefore $\text{AL} = \text{AL}_B + 2\log \frac{2}{\epsilon}$. Note the blind search phase depends on the width of the target interval: when $b^* - a^* \approx \epsilon$ the label complexity $\text{AL} \sim O\left(\frac{1}{\epsilon}\right)$, and active learning's advantage over passive learning disappears.

Hence our second observation: Such a blind search phase is common in active learning and unavoidable for some hypothesis spaces. Another example is active learning over axis-aligned hypercubes in $\mathbb{R}^d$ (Hanneke, 2014). The blind search phase has $\text{AL}_B = \frac{2}{\max(w^*,\epsilon)^d}$ where w^* is the narrowest side length among the d dimensions of the target hypercube. Yet another example of blind search is active learning on graphs (Dasarathy et al., 2015). Let the sampling probability of the smallest positive connected component in a graph be β. The blind search phase has $\text{AL}_B = \frac{\log(1/(\epsilon\beta))}{\log(1/(1-\beta))}$. When $\beta \approx \epsilon$ the denominator is around ϵ due to $\log(1-z) \approx -z$, again making $\text{AL}_B \sim O(1/\epsilon)$. The intuition is that, for these hypothesis spaces, the positive items form small isolated "islands" in the input space. Active learning really has no choice but to do blind (essentially random) search to find such islands. If an island is too small, the label complexity to find it can be as slow as the passive rate of $\frac{1}{\epsilon}$. However, once a single point in the island is found, active learning can very aggressively trace the island boundary at a rate of $\log \frac{1}{\epsilon}$. Note that the blind search phase is not always present for all hypothesis spaces: for example, it is not in the 1D threshold task.

3.2. Human-Initiated Training

When a human is in the driver's seat, teaching can become really good or bad depending on the ability of the teacher.

Optimal teacher. There is a fundamental difference between active learning and human-initiated training. In active learning, the computer does not know the target concept f^* (the risk minimizer) and must perform exploration. As we have seen, even the best active learning takes $O\left(\log \frac{1}{\epsilon}\right)$ queries. This is illustrated in our 1D threshold task: the computer does not know f^* to begin with, and binary search has label complexity $\log \frac{1}{\epsilon}$. In sharp contrast, a good human teacher *can* have the knowledge of f^*. She just needs to communicate f^* to the computer via the smallest training set. The size of the smallest training set to *exactly* teach f^* is known as the teaching dimension TD (Goldman & Kearns, 1995; Balbach & Zeugmann, 2009; Doliwa et al., 2014; Shinohara & Miyano, 1991). We need to relax the definition of classic TD to match the ϵ precision requirement in Definition 1. Let $D = (x_1, y_1), \ldots, (x_n, y_n)$ be a data set. Let $VS(D) = \{h \in \mathcal{H} : h(x_i) = y_i,\ i = 1 \ldots n\}$ be the version space, i.e. the set of hypotheses consistent[1] with D.

Definition 3 (ϵ-Teaching Dimension $\text{TD}(\epsilon)$)**.** *A data set D is called an ϵ-teaching set of f^* if $\forall h \in VS(D)$, $R(h) - R(f^*) \leq \epsilon$. The ϵ-teaching dimension of f^* is the size of the smallest ϵ-teaching set.*

Our ϵ-teaching dimension is a probabilistic generalization of (Kobayashi & Shinohara, 2009). The classic teaching dimension is $\text{TD}(0)$. For any $\epsilon_1 \leq \epsilon_2$, $\text{TD}(\epsilon_1) \geq \text{TD}(\epsilon_2)$. On our two example tasks, the classic TD from (Goldman & Kearns, 1995) would be $\text{TD}(0) = \infty$, since no finite data set can reduce the version space to precisely the target concept. Nonetheless, we can adapt their teaching set construction for $\text{TD}(\epsilon)$. For our 1D threshold task, an ϵ-teaching set D is $(x_1 = \max(\theta^* - \epsilon, 0), y_1 = -1)$, $(x_2 = \min(\theta^* + \epsilon, 1), y_2 = 1)$, and $\text{TD}(\epsilon) = 2$. The implication for human-initiated classifier training is profound: an optimal human teacher can choose the above data set D of only two items to train the classifier. Moreover, $\text{TD}(\epsilon) = 2$ for any small precision $\epsilon > 0$. This is in stark contrast to active learning, where we have a growing $\text{AL} = \log(1/\epsilon)$ as ϵ decreases. In other words, an optimal human teacher can train the 1D threshold classifier with far

[1] For unrealizable tasks VS is generalized to the set of low error hypotheses the learner retains.

fewer examples than active learning.

The same is true for our 1D interval task. An ϵ-teaching set D is $(x_1 = \max(a^* - \frac{\epsilon}{2}, 0), y_1 = -1)$, $(x_2 = \min(a^* + \frac{\epsilon}{2}, \frac{a^*+b^*}{2}), y_2 = 1)$, $(x_3 = \max(b^* - \frac{\epsilon}{2}, \frac{a^*+b^*}{2}), y_3 = 1)$, $(x_4 = \min(b^* + \frac{\epsilon}{2}, 1), y_4 = -1)$, and $\mathrm{TD}(\epsilon) = 4$. Note that the teaching dimension does not depend on the width of the target concept $b^* - a^*$ or ϵ. Again an optimal human teacher can use fewer examples than active learning. Henceforth we will omit the ϵ parameter from TD when it is clear from context. In fact, on any classification task an optimal human teacher can always use no more training examples than active learning. This has been observed in prior work (Cakmak & Thomaz, 2011; Angluin, 2004; Goldman & Kearns, 1995). We highlight it as follows.

Proposition 1 (The Fundamental Law of Interactive Classifier Training)**.** *On all classification tasks and for all* ϵ, $\mathrm{TD} \leq \mathrm{AL}$.

The proof is by definition. When active learning satisfies ϵ precision, its queries and oracle labels form an ϵ-teaching set. TD is the size of the smallest ϵ-teaching set. To get the benefit of TD, the optimal teacher must choose the teaching items and cannot rely on active learning.

Seed teacher and Naive teacher. Unfortunately, human-initiated training can have unbounded label complexity for non-optimal teachers. The best scenario is for a teacher to choose TD +1 items that form a (non-minimal) teaching set. For instance, for the 1D threshold task, the teacher may have unnecessarily added $(x_3 = 1, y_3 = 1)$ to the optimal teaching set. Far worse are teachers who never form a teaching set. To see why such teachers can lead to unbounded label complexity, consider a hypothetical teacher for the 1D threshold task who chooses to teach with $(x_i = 1 - \frac{1-\theta^*}{2^{i-1}}, y_i = 1)$ for $i = 1, 2, \ldots$. This training set can never reduce the version space sufficiently. Such inefficiency can happen on both seed teachers and naive teachers. For this reason, the entries are marked as ∞ in Table 1. They also cover the case where the teacher mistakenly thought that he has taught the concept and stopped providing training items too early.

3.3. Mixed-Initiative Training

The key to mixed-initiative training is to reap the benefits of optimal teaching in human-initiated training, while controlling for the unbounded label complexity. There are many possible mechanisms to do this. We focus on Algorithm 1 for its simplicity.

Optimal teacher. Algorithm 1 allows for TD rounds of human teaching in the beginning. This is designed to allow an optimal teacher to teach with an optimal teaching set. The algorithm will exhaust these TD human-chosen rounds and exit the loop with the data set D containing an optimal teaching set. Since D is already sufficient to teach to ϵ precision, active learning on line 10 detects this and issues no additional queries. The label complexity is TD.

Seed teacher. A non-optimal teacher may choose TD items that do not form a teaching set, or he may decide to stop choosing items altogether before there are TD items. In either case, Algorithm 1 exits the loop and forces the human to continue with active learning. In general, the data in D at this point will be beneficial to active learning. A seed teacher is able to choose at least one positive item for each positive region. For hypothesis spaces with an active learning blind search phase, we can quantify the benefit as removing AL_B from active learning label complexity (Table 1). For those hypothesis spaces where AL_B dominates, a seed teacher will substantially speed up active learning. The benefit of seed teachers are informally known in practice. For real-world tasks the positive class can often be rare. Various practical active learning systems offer ways for the human oracle to explicitly search and discover positive items, before active learning starts (Attenberg & Provost, 2010; Cakmak et al., 2010).

Naive teacher. For all other teachers, the worst case is that they choose TD items x that are non-informative. Note we still assume that the teacher gives labels according to the marginal $\mathcal{P}_{Y|X=x}$. In particular, for realizable tasks, the teacher always gives the correct label y for any x. Their non-informativeness stems solely from their choice of x. Because active learning takes over, the worst case label complexity is $\mathrm{TD} + \mathrm{AL}$. Obviously, this is upper bounded by $2\mathrm{AL}$. In other words, we have the fallback guarantee by preventing the teacher from teaching aimlessly; sometimes active learning is smarter than a human.

4. Teacher Education

So far we have taken a static view of the human teacher: they come and teach at their fixed teaching ability state. A natural question is: can we first teach the humans to be better teachers, before they train the classifier?

We view the teacher as a finite state machine, and that the computer can perform an action to cause a state transition. In this paper we consider the specific action of computer displaying a hint text to the human teacher. Designing richer actions is an interesting problem for future work. We want the hint text to be generated automatically by the computer for any hypothesis space. One good hint would have been showing an actual optimal teaching set D for the target concept f^*. Of course, the computer does not know f^* upfront. Nonetheless, the computer can arbitrarily pick a concept $f' \in \mathcal{H}$, and compute the optimal teaching set D' for f'. It can then use the pair f', D' to construct the following hint text for the human teacher: "To teach f' to

the computer, you could have used D'." Such teacher education is thus **education by analogues**. For our example tasks, we can go one step further by manually taking this optimal teaching set and explaining the effect of the teaching examples on the hypothesis space visualized in a number line. By providing simple quizzes and explanations, we encourage deeper understanding of the learning algorithm and higher rate of state transition. This is **education by explanation**. After a human teacher teaches, we will be able to determine her teacher state from the items she manually chooses. Our primary interest is to find out whether teacher education turns more teachers into optimal teachers.

5. Human Experiments

We conducted human experiments using Amazon Mechanical Turk (MTurk) on the two 1D classifier training tasks (threshold, interval). We designed a 2x3, between-subjects study where each experiment compared the three training paradigms (computer-initiated, human-initiated, mixed-initiative) and teacher education (no education, education by analogues). After running the initial pilot test, we saw evidence of shallow understanding from the participants' teaching strategies especially for the interval classifier where we saw less optimal teachers than the threshold classifier. To influence the teacher's understanding of the system, for the interval task, we further divided teacher education into three conditions (no education, education by analogues, education by explanation). Teacher education does not apply to the computer-initiated condition, which gave us a total of 5 conditions for threshold task and 7 conditions for interval task for the final run.

5.1. Participants, Tasks, and Procedure

We selected MTurk workers with Human Intelligence Task (HIT) approval rate $\geq 98\%$. We made sure each worker only participated in one condition. We recruited a total of 481 participants (282 male, 187 female). 49% of the participants were in the age range of 26 to 35.

The tasks are the integer variant of the 1D threshold and interval classifier training tasks, see Figure 1. That is, $\mathcal{X}$ is some finite integer range $c, \ldots, d$. $\mathcal{H}$ is also finite and contains hypotheses whose threshold or interval is on integers. The goal is to uniquely identify the integer target threshold or target interval. We use this integer variant because it is easier for the participants. The integer variant is similar to the continuous tasks with an $\epsilon < \frac{1}{2(d-c)}$. All the theoretical results still hold. We choose a cover story for the participant to teach a robot assistant a threshold or a range of acceptable prices when purchasing a car. In our consideration for the cover story, we wanted it to be relatable without any predefined notion of the positive class and reusable to many ranges within the hypothesis space where small differences in numbers mattered. Note the participants were informed that the robot's hypothesis space consisted of threshold or interval classifiers, respectively.

For the 1D threshold task, $\mathrm{TD} = 2$ with the optimal teaching set $\{(x_1 = 19000, y_1 = 1), (x_2 = 19001, y_2 = -1)\}$, while active learning using binary search would require 14 queries. For the 1D interval task, $\mathrm{TD} = 4$ with the optimal teaching set $\{(x_1 = 1259, y_1 = -1), (x_2 = 1260, y_2 = 1), (x_3 = 1360, y_3 = 1), (x_4 = 1361, y_4 = -1)\}$, while active learning requires 26 queries.

For the teacher education conditions with analogues, an additional piece of text is displayed to the participant as shown in Figure 2. The analogues are precomputed and consist of the optimal teaching set for two hypothetical target concepts. For the teacher education conditions with explanation, we provide a step-by-step tutorial illustrating the effects of teaching examples using the number line as well as three simple quizzes to test the understanding of the participants. The details are in supplementary materials.

The computer-initiated condition selects a dollar amount x, queries the teacher for the label y (acceptable or not), and terminates when the computer can identify a unique hypothesis. The human-initiated condition allows the teacher to enter x and y, and the teacher decides when to terminate. The mixed-initiative condition follows Algorithm 1.

Each participant was randomly assigned to one of the five or seven conditions. There were no time limits to the task. We analyzed the participant data first by filtering out all participants who incorrectly labeled any items. Among the 481 participants, 13 in the threshold task and 46 in the interval task were filtered out. This is in order to match the noiseless marginal $\mathcal{P}_{Y|X=x}$ human labeling assumption we made on the tasks.

5.2. Empirical Label Complexity

In this section we look at the no-teacher-education conditions. The human experiment results are summarized as histograms in Figure 3. Our empirical results verify the label complexity in Table 1 and clearly demonstrate the benefits of mixed-initiative training.

In the mixed-initiative condition, the most important observation is that all human teachers taught with from TD to TD+AL items as the theory predicts. For the threshold and interval tasks this interval is $[2, 16]$ and $[4, 30]$, respectively. The average label complexity is 6.6 ($n = 38, sd = 6.2$) for the threshold task and 13.1 ($n = 31, sd = 8.8$) for the interval task.

A closer look reveals that many mixed-initiative participants are optimal teachers (dark blue bars) who taught with TD items. 50.0% (19/38) of the threshold participants and

Imagine you are looking to buy a car. Car prices go from $10000 to $30000, but you will only accept a car priced at $19000 or below. You have a robot assistant who knows that your acceptable price falls at or below a threshold, but it does not know what your acceptable threshold is. Your task is to teach your robot what your acceptable threshold is:
· You can only give examples like "$_ is acceptable" or "$_ is unacceptable."
· You cannot afford any car over $19000 by even $1 because you only have $19000 in your bank account.
Provide the fewest number of examples possible while still making sure your robot has clearly understood your price threshold.

Imagine you are looking to buy a car. Car prices go from $500 to $1500, but you will only accept a car in the range of $1260 to $1360. You have a robot assistant who knows that your acceptable price falls into a range, but it does not know what your acceptable range is. Your task is to teach your robot what your acceptable price range is:
· You can only give examples like "$_ is acceptable" or "$_ is unacceptable."
· You believe any car under $1260 by even $1 will break down.
· You cannot afford any car over $1360 by even $1 because you only have $1360 in your bank account.
Provide the fewest number of examples possible while still making sure your robot has clearly understood your price range.

Figure 1. The integer variant of the 1D threshold (left) and interval (right) classifier training tasks for human experiments

If your price threshold was $20000 or below, you could show your robot these 2 examples: $20000 is acceptable, $20001 is unacceptable.
If your price threshold was $24000 or below, you could show your robot these 2 examples: $24000 is acceptable, $24001 is unacceptable.

If your price range were $900 to $1000, you could show your robot these 4 examples: $899 is unacceptable, $900 is acceptable, $1000 is acceptable, $1001 is unacceptable.
If your price range were $600 to $700, you could show your robot these 4 examples: $599 is unacceptable, $600 is acceptable, $700 is acceptable, $701 is unacceptable.

Figure 2. Education by analogues for the 1D threshold (left) and interval (right) tasks, respectively

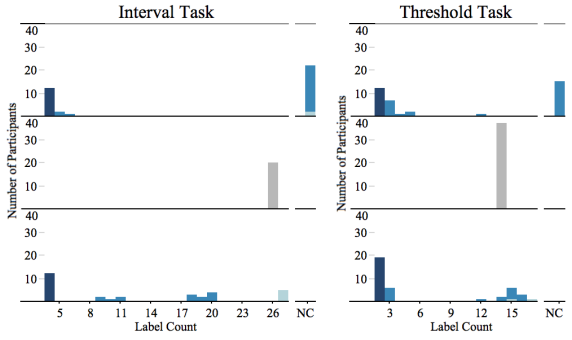

Figure 3. Number of human teachers (y-axis) who used a certain number of labeled items (x-axis) to teach. Rows from top down: human-initiated, computer-initiated, mixed-initiative conditions. All conditions without teacher education. NC: teaching not completed when the teacher stopped providing items. Blue color intensity represents observed teacher ability state: dark=optimal, medium=seed, light=naive. Human-initiated condition has a large number of NC participants, and mixed-initiative condition has the label count bounded by AL + TD.

38.7% (12/31) of the interval participants did so. This is the result of two separate benefits of mixed-initiative training: (1) mixed-initiative training *enables* optimal teaching, where as teachers in computer-initiated training are limited by active learning and cannot teach with TD items; (2) our mixed-initiative mechanism in Algorithm 1 can actually force some potentially suboptimal teachers into optimal ones. We observed from our human-initiated conditions that out of all participants that initially provided TD items forming an optimal teaching set, 20% (3/15) in the interval task and 40% (8/20) in the threshold task provided at least one additional item. If these participants were given our mixed-initiative mechanism, Algorithm 1 would cut them off at the initial TD items and stop with success. This can be seen by comparing to the shorter dark blue bars in the human-initiated condition in Figure 3 – in that condition there is nothing stopping such teachers from overteaching. Indeed only 31.6% (12/38) of the threshold participants and 32.4% (12/37) of the interval participants are optimal in the human-initiated condition.

In the computer-initiated condition, as expected the participants used exactly AL items: 14 ($n = 37, sd = 0$) for threshold and 26 ($n = 20, sd = 0$) for interval. For our tasks, AL $\gg$ TD. In fact one participant in our pilot run expressed the desire to do optimal teaching rather than active learning: "if I had my way I would only answer 4 questions: $(low - 1)$=bad, (low)=good, $(high)$=good, $(high + 1)$=good [*sic*].[2]"

In the human-initiated condition, the most important observation is the large number of not-completed (NC) participants: 39.5% (15/38) for the threshold task and 59.5% (22/37) for the interval task. These NC participants did not provide enough items to exactly specify the target concept before they decided to stop. Thus, this observation high-

[2]This participant correctly labeled $x = (high+1)$ as negative during active learning, so we believe this is a typo.

lights human inefficiency and corresponds to the ∞ cells in Table 1. Some of the remaining participants did manage to teach with TD items, though the fraction is smaller than in mixed-initiative training since some provided more labels than necessary.

We also observe the removal of "blind search complexity AL_B". All of the seed teachers in mixed-initiative, interval conditions used less than AL items to complete the task, while the computer-initiated conditions required 14 labels to even find the first positive item.

5.3. Effect of Teacher Education

Figure 4 shows the percentage of teachers in different states, without and with teacher education. Recall education is done by displaying the analogues in Figure 2 or providing step-by-step explanations. Our primary interest is the percentage of optimal teachers. In all conditions, there are more optimal teachers with teacher education intervention. We also observe higher percentage of optimal teachers when the education is given in the form of detailed explanations than in teaching set analogues.

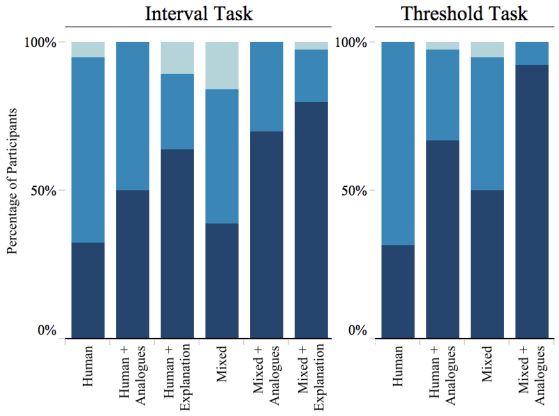

Figure 4. Effect of teacher education. Blue color intensity represents teacher state: dark=optimal, medium=seed, light=naive.

6. Related Work

Interactive machine learning has been hindered in practice by skewed class distributions and small hard to find "island" concepts (Attenberg & Provost, 2011). Recent work has compared human-initiated, computer-initiated and various mixed-initiative training strategies (e.g., turn taking) (Cakmak et al., 2010; Attenberg & Provost, 2010; Fogarty et al., 2008). The most directly related work is (Cakmak & Thomaz, 2011), which is the first to link optimal teaching and mixed-initiative training. Our paper generalizes and formalizes their work. Many alternatives to design mixed-initiative training remain (Horvitz, 1999). Human-computer interaction researchers have begun to improve the teaching ability of humans in interactive settings (Amershi et al., 2014; Cakmak & Thomaz, 2014; 2010; Fails & Olsen Jr, 2003). Finally, there is a growing literature on the teaching dimension and the corresponding optimal teaching sets for different hypothesis spaces and learners, including hypercubes, monomials, linear learners, and Bayesian models (Goldman et al., 1993; Goldman & Kearns, 1995; Liu et al., 2016; Zhu, 2013; 2015).

7. Discussions

Examination of participant labeling strategies in the human-initiated conditions show that, without the help of mixed-initiative mechanism the human teacher can be inefficient. 29.0% (threshold) and 8.1% (interval) of participants provided more than the necessary TD training items. For example, one participant in the human-initiated condition on the interval task (pilot) provided the robot with the TD training items, but additionally provided *every* positive item within the acceptable interval in order to teach the robot "all acceptable price ranges." These results have practical implications for machine learning based applications–namely that support for mixed-initiative training may prevent humans from wasted effort while also increasing machine learning efficiency.

We used analogues to educate the teachers. We observed that more people did provide the optimal teaching set when shown explanations or analogues than without these hints. Many of these participants, however, also provided additional and unnecessary training items either within their first TD opportunities or elsewhere during training. This suggests an opportunity to further educate people in *prioritizing* teaching set items over other possible training items. Furthermore, the teacher education action we consider in this paper focused on transitioning humans into the *optimal* teaching state. An alternative, or additional, action could include hints emphasizing that positive training items (for domains with rare positives) are more helpful to machine learning than negative ones. This may have the effect of transitioning humans *out of* the naive state into a more efficient one. Further research is necessary to examine the effectiveness of this as a viable teacher education strategy.

Participants' teaching strategies also reveal violations to some of the assumptions often made in theoretical analyses of machine learning, including our own presented in this paper. For example, we assumed a noiseless oracle for our two tasks. However, our empirical analyses show human teachers provided wrong labels (wrong y for some x) in 3.5% of cases, even though the task description was clear and unambiguous. The error rate of labeling was higher when the label was requested by the algorithm (4%) than when the human entered the examples (2.2%). We hypoth-

esize that the participants pay less attention to examples presented to them by the active learning algorithm than an example that they manual enter. If we treat labeling speed as a proxy to attention, our data shows that participants' labeling speed is faster in the computer-initiated condition compared to human-initiated condition by six-fold (discussed later), but further examination of these hypotheses is necessary. As another example, we assumed that participants understood the robot's hypothesis space. However, some participants in the threshold task, for example, indicated that they believed the hypothesis space consisted of intervals: "Entered lowest price acceptable, as well as, highest price acceptable."

Related to this, several participants indicated that they believed the robot *required* additional, prototypical training items to learn the target concept. For example, one participant in the human-initiated condition (1D interval, no education) provided the four teaching set items along with an additional prototypical training item from the middle of the positive interval. This participant explained their strategy as: "I set the lower bounds and upper bounds of the range. By doing that I set what the two out of bound items were acceptable. I then encapsulated a mid-range as being acceptable by selecting the midpoint of the acceptable amounts." This belief also manifested in the 1D interval with-education condition (e.g., "I showed the middle and extremes of the acceptable range") and the 1D threshold condition (e.g., "All the robot really needed to know was the maximum and one number over the maximum, to show what the maximum I would go is. 19000 being acceptable and 19001 being unacceptable, then a number below 19000 to show you can go down"). These statements reveal errors in our participants' mental models of how to communicate their concept to the learner. Mental models are internal representations that people formulate about systems they interact with (Johnson-Laird, 1983). As observed in these examples, incorrect mental models can result in inefficient or erroneous behaviors, emphasizing the need for further research in educating humans about how to most effectively interact with machine learning algorithms.

We observed that many participants are not able to behave in an optimal teaching state due to this mismatch between what the participants believe of the system behavior and how the system actually behaves. Even amongst participants who were in the optimal state, we see a varying level of certainty in their understanding of the robot. One participant in the human, interval condition said "Well, I really hope the robot understands ranges. I didn't want to cover every single number, so randomly pulling up numbers to say that they are acceptable or not didn't seem worth it. So I just marked the upper and lower limits and hoped it figured it out." In the human-initiated, interval condition with education, some participant comments suggest that analogues alone provide only a shallow understanding of the robot because they could simply mimic the behavior described in the instructions ("I followed the examples given based on other prices."). On the other hand, we see evidence of step-by-step explanations helping the participants gain deeper understanding of the robot and its hypothesis space. One participant in the human, interval condition with explanations said, "The robot knows there is a range. I provide the high and low end of the range as acceptable. Then I provide $1 under the low and $1 over the high as unacceptable. The robot knows the range is between the first two numbers." These examples as well as the empirical results illustrate that our approach to provide explanations can help bridge this gap.

Identifying actual human fallibilities also presents opportunities to guide users through better interface design. Some participants, for example, indicated that they were aware of the incorrect labels they provided by accident. This, along with the higher rate of error in computer-initiated labeling, suggests an interface for surfacing the history of items and labels provided along with mechanisms for correction.

One major assumption we made is to approximate human teacher effort with label complexity. This assumes that each item x and corresponding label y obtained for training requires equivalent effort on the part of the human. In practice, however, manually selecting x may take more cognitive effort than providing a label y on an x queried by the computer. This is well-known in cost-sensitive active learning (Settles, 2012). Our interval tasks provide some evidence for this, given that participants appeared faster at labeling in the computer-initiated condition (17.3 labels/min) than in the human-initiated condition (2.8 labels/min) while their overall task duration was similar (63 sec vs. 56.5 sec, respectively). The greater effort required in selecting x may be amplified in higher dimensional spaces where the user may not know x or know of all possible positive islands a priori. This presents additional opportunities to design rich interfaces equipped with efficient mechanisms to search for or generate training items (e.g., feature based interactions).

Finally, our analyses only included one possible mixed-initiative mechanism, where users have the opportunity to provide items only at the beginning, but other strategies may include interleaving human and computer-initiated phases. One could argue that, in a realistic application, it may be difficult for humans to come up with the exact teaching set at the beginning. What would happen if the humans provided subsets of the teaching set in batches interleaved with active learning queries? Such a method may provide different theoretical guarantees and empirical results and is one of many research opportunities to consider.

Acknowledgements

The authors would like to thank the Machine Teaching Group at Microsoft Research for their support. This work was done when XZ visited Microsoft Research.

References

Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, Winter 2014.

Angluin, D. Queries revisited. *Theoretical Computer Science*, 313(2):175–194, 2004. Algorithmic Learning Theory.

Attenberg, J. and Provost, F. Why label when you can search?: alternatives to active learning for applying human resources to build classification models under extreme class imbalance. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 423–432. ACM, 2010.

Attenberg, J. and Provost, F. Inactive learning?: difficulties employing active learning in practice. *ACM SIGKDD Explorations Newsletter*, 12(2):36–41, 2011.

Balbach, F. J. and Zeugmann, T. Recent developments in algorithmic teaching. In *Proceedings of the 3rd International Conference on Language and Automata Theory and Applications*, pp. 1–18, 2009.

Beygelzimer, A., Dasgupta, S., and Langford, J. Importance weighted active learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 49–56. ACM, 2009.

Cakmak, M. and Thomaz, A. Mixed-initiative active learning. *ICML Workshop on Combining Learning Strategies to Reduce Label Cost*, 2011.

Cakmak, M. and Thomaz, A.L. Optimality of human teachers for robot learners. In *Proceedings of the IEEE International Conference on Development and Learning (ICDL)*, pp. 64–69. IEEE, 2010.

Cakmak, M. and Thomaz, A.L. Eliciting good teaching from humans for machine learners. *Artificial Intelligence*, 217:198–215, 2014.

Cakmak, M., Chao, C., and Thomaz, A.L. Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development*, 2(2):108–118, 2010.

Dasarathy, Gautam, Nowak, Robert, and Zhu, Xiaojin. s^2: An efficient graph based active learning algorithm with application to nonparametric classification. *COLT*, 2015.

Doliwa, T., Fan, G., Simon, H.U., and Zilles, S. Recursive teaching dimension, VC-dimension and sample compression. *Journal of Machine Learning Research*, 15:3107–3131, 2014.

Fails, J.A. and Olsen Jr, D.R. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pp. 39–45. ACM, 2003.

Fogarty, J., Tan, D., Kapoor, A., and Winder, S. Cueflik: interactive concept learning in image search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 29–38. ACM, 2008.

Goldman, S. and Kearns, M. On the complexity of teaching. *Journal of Computer and Systems Sciences*, 50(1): 20–31, 1995.

Goldman, S.A., Rivest, R.L., and Schapire, R.E. Learning binary relations and total orders. *SIAM Journal on Computing*, 22(5):1006–1034, 1993.

Hanneke, S. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2-3): 131–309, 2014.

Horvitz, E. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 159–166. ACM, 1999.

Johnson-Laird, P.N. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press, 1983.

Kobayashi, H. and Shinohara, A. Complexity of teaching by a restricted number of examples. In *Proceedings of the Conference on Learning Theory*, pp. 293–302, 2009.

Kulkarni, S.R., Mitter, S.K., and Tsitsiklis, J.N. Active learning using arbitrary binary valued queries. *Machine Learning*, 11(1):23–35, 1993.

Liu, J., Zhu, X., and Ohannessian, H.G. The teaching dimension of linear learners. In *The 33rd International Conference on Machine Learning (ICML)*, 2016.

Settles, B. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2012.

Shinohara, A. and Miyano, S. Teachability in computational learning. *New Generation Computing*, 8(4):337–348, 1991.

Wolfman, S.A., Lau, T., Domingos, P., and Weld, D.S. Mixed initiative interfaces for learning tasks: Smartedit talks back. In *Proceedings of the 6th international conference on Intelligent user interfaces*, pp. 167–174. ACM, 2001.

Zhu, X. Machine teaching for Bayesian learners in the exponential family. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1905–1913, 2013.

Zhu, X. Machine teaching: an inverse problem to machine learning and an approach toward optimal education. In *AAAI*, pp. 4083–4087, 2015.