Data Mining and Knowledge Discovery

More for less: Adaptive labeling payments in online labor markets

--Manuscript Draft--

Manuscript Number:	DAMI-D-17-00623
Full Title:	More for less: Adaptive labeling payments in online labor markets
Article Type:	Manuscript
Keywords:	Machine learning; Supervised learning; Label acquisition; Crowdsourcing; Online labor markets; Adaptive labeling payments
Corresponding Author:	Maytal Saar-Tsechansky, PhD The University of Texas at Austin Austin, Texas UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	The University of Texas at Austin
Corresponding Author's Secondary Institution:	
First Author:	Maytal Saar-Tsechansky, PhD
First Author Secondary Information:	
Order of Authors:	Maytal Saar-Tsechansky, PhD
	Tomer Geva, Ph.D.
	Harel Loostinger, M.Sc.
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	In many predictive tasks where human intelligence is needed to label training instances, online crowdsourcing markets have emerged as promising platforms for large-scale, cost-effective labeling. However, these platforms also introduce significant challenges that must be addressed for these opportunities to materialize. In particular, it has been shown that different tradeoffs between payment offered to labelers and the quality of labeling arise at different times, possibly as a result of different market conditions and even the nature of the tasks themselves. Because the underlying mechanism giving rise to different tradeoffs is not well understood, for any given labeling task and at any given time, it is not known what labeling payments to offer in the market so as to cost-effectively produce accurate models. Importantly, because in these markets the acquired labels are not always correct, determining the expected effect of labels acquired at any given payment on the improvement in model performance is particularly challenging. Effective and robust methods for dealing with these challenges are essential to enable a growing reliance on these promising and increasingly popular labor markets for large-scale labeling. In this paper, we first present this new problem of Adaptive Labeling Payment (ALP): how to learn and sequentially adapt the payment offered to crowd labelers before they undertake a labeling task, so as to produce a given predictive performance cost-effectively. We then develop an ALP approach and discuss the key challenges it aims to address to yield consistently good performance. We extensively evaluate our approach over a wide variety of market conditions. Our results demonstrate that the ALP method we propose yields significant cost savings and robust performance across different settings. As such, our ALP approach can be used as a benchmark for future mechanisms to determine cost-effective selection of labeling payments.

Data Min Knowl Disc
DOI 10.1007/s10618-013-0306-1

More for less: Adaptive labeling payments in online labor markets

Received: 12 December 2017

Abstract In many predictive tasks where human intelligence is needed to label training instances, online crowdsourcing markets have emerged as promising platforms for large-scale, cost-effective labeling. However, these platforms also introduce significant challenges that must be addressed for these opportunities to materialize. In particular, it has been shown that different tradeoffs between payment offered to labelers and the quality of labeling arise at different times, possibly as a result of different market conditions and even the nature of the tasks themselves. Because the underlying mechanism giving rise to different tradeoffs is not well understood, for any given labeling task and at any given time, it is not known what labeling payments to offer in the market so as to cost-effectively produce accurate models. Importantly, because in these markets the acquired labels are not always correct, determining the expected effect of labels acquired at any given payment on the improvement in model performance is particularly challenging. Effective and robust methods for dealing with these challenges are essential to enable a growing reliance on these promising and increasingly popular labor markets for large-scale labeling. In this paper, we first present this new problem of Adaptive Labeling Payment (ALP): how to learn and sequentially adapt the payment offered to crowd labelers before they undertake a labeling task, so as to produce a given predictive performance cost-effectively. We then develop an ALP approach and discuss the key challenges it aims to address to yield consistently good performance. We extensively evaluate our approach over a wide variety of market conditions. Our results demonstrate that the ALP method we propose yields significant cost savings and robust performance across different settings. As such, our ALP approach can be used as a benchmark for future mechanisms to determine cost-effective selection of labeling payments.

Keywords Machine learning · Supervised learning · Label acquisition · Crowdsourcing · Online labor markets · Adaptive labeling payments

1 Introduction

Predictive modeling has radically impacted a wide variety of industries, becoming integral to the operations and competitive strategies of firms and giving rise to entirely new business platforms. Supervised learning of predictive models requires labeled training data—namely, instances for which the dependent variable value (label) is known. However, in many important applications, costly human intelligence is also needed to determine the labels for training data. Such labeling tasks include, for example, image classification tasks (e.g., whether an image contains a car, for application of surveillance, autonomous driving, and scene understanding), and text classification problems (e.g., whether a text contains hate remarks, or is humorous or sarcastic). Many similar labeling tasks are relatively intuitive and simple for humans to perform, but may require a large number of training examples for supervised learning to yield good performance. For such tasks, online crowdsourcing marketplaces, such as Amazon Mechanical Turk (AMT), have emerged as promising platforms for large-scale labeling, offering unprecedented scalability and immediacy. Markets such as AMT offer substantial savings and agility by allowing employers ("requesters") to offer simple micro-tasks to a large number of online workers simultaneously, thereby allowing to acquire labels cheaply and relatively quickly. In these markets, requesters typically post the task description and payment offered to all workers who meet certain criteria, and workers can review this information when choosing which tasks to undertake.

A fundamental challenge for the cost-effective use of such markets, however, is that for any given task and at any given time, it is not known what labeling quality can be acquired on the market for any given payment. Prior studies have found that different prevailing tradeoffs arise between payment and quality for different tasks and at different times (e.g., Kazai, 2011; Kazai et al., 2013; Mason and Watts, 2010). Yet the mechanisms giving rise to these different tradeoffs are complex and not well understood. They can be affected by a host of properties, such as the labeling task itself and the market conditions—including the impact of competing tasks on the market, or the composition and availability of workers on the market at any given time.

Because the cumulative costs of routine use of crowdsourcing labor markets for labeling can be substantial, a robust, data-driven methodology for identifying advantageous payments for labeling that yield a desired predictive performance in a cost-effective manner is essential.

In this paper, we refer to this problem as Adaptive Labeling Payment (ALP). Specifically, ALP refers to data-driven learning and adaptation of payment amounts to offer to crowd workers before they undertake a labeling task, so as to produce a given predictive performance in a cost-effective manner. The ALP problem considers online crowdsourcing labor markets where a task is offered to all workers for a given payment, and where no prior knowledge on an individual worker's performance for the particular task may be available (e.g., when a new worker is hired). In addition, we consider markets such as AMT, where the population of available workers and the competitive settings may also vary over time.

The ALP problem is related to active learning (Lewis and Gale, 1994; Abe and Mamitsuka, 1998; Kong and Saar-Tsechansky, 2014), and to prior work about online labor markets. However, as we discuss in detail below, both prior streams of work did not aim to adaptively identify advantageous payments to produce a desirable model performance cost-effectively (e.g., Raykar et al., 2010; Wang et al., 2017). The most closely related prior work considered theoretical settings in which labelers' quality for a given price is predetermined and known to the requester (Yang and Carbonell, 2012); however, such settings do not hold in real-world crowdsourcing labor markets we consdier here, where new crowd workers with no prior history are continuously encountered.

The ALP problem presents several challenges, which we discuss in the remainder of the paper. Perhaps the most fundamental challenge pertains to our objective of identifying the labeling payments that cost-effectively improve *predictive performance*. Indeed, unlike most prior work on crowdsourcing labor markets, in our problem settings, identifying advantageous labeling payments corresponds not to merely learning the effect of payment on the quality of the labels, but rather, to learning the effect of payment on the predictive performance of the model induced from the acquired, labeled training instances. Note that the acquired labels can be noisy (incorrect). Consequently, the choice of payments is ultimately affected by a host of factors in addition to those considered to improve labeling quality per se. These include the properties of the inductive modeling algorithm itself (e.g., resilience to noisy labels), the difficulty of the predictive task (i.e., the ability to distinguish the underlying patterns from noise), the benefits to induction of the labels purchased thus far, and even the current position on the learning curve.[1] Indeed, even labels of the same quality may not be equally cost-effective at different points along the predictive model's learning curve. This situation occurs because the marginal effect on model learning often differs substantially when the same set of labeled instances are added at different points along the learning curve, namely, when a different set of training instances are already available for induction. As we will see, these factors, along with the presence of noisy data, make accurate estimation about the effect of alternative payments on future model performance quite challenging.

To address this problem, we develop an ALP algorithm, and extensively evaluate its performance and robustness relative to alternatives under different settings. Our results demonstrate that the ALP method we propose often yields substantial cost savings compared to the existing benchmark, and that its performance is robust over a wide variety of settings. In addition, we find that when the underlying tradeoff between pay and labeling quality changes over time, our ALP method effectively adapts and

[1] The marginal improvement in predictive performance from acquiring the same 100 labeled training instances would likely be substantially different if we already have a large nunber of labeled training data instances, than if the training set is very small. Similarly, acquiring highly noisy labels may have a different impact on learning when the training set is small, than if added to a large training set.

continues to maintain robust performance. Overall, our results show that the proposed ALP method constitutes both an effective and robust approach for adapting payments to labelers in dynamic crowdsourcing markets.

The contributions of this work are as follows. Our study is the first to introduce the problem of Adaptive Labeling Payment for online crowdsourcing labor markets. We propose a data-driven ALP method that both learns and continuously adapts the payments offered to labelers, with the objective being to achieve a given predictive performance cost-effectively. Importantly, our ALP approach aims to directly estimate the expected benefits to predictive performance from labels acquired at different costs, rather than merely improve labeling quality. To do this, our algorithm introduces a novel approach for estimating the expected change in model performance from future acquisitions of labels at different payments; our proposed approach uses already-available noisy labeled data, and it does not rely on the acquisition of costly "gold standard" data. The ALP approach we develop here is also generic, and can be applied in order to cost-effectively acquire labeled training data for any given classification algorithm, population of workers, or set of market conditions. Finally, because our approach aims to identify advantageous payments for label acquisition, it can also be applied in conjunction with methods that address complementary problems, such as repeated labeling methods that use multiple labels per instance to improve the labeling quality. In the empirical evaluations, we demonstrate how such methods can be effectively applied simultaneously.

The rest of this paper is organized as follows. We review related research in the prior work section. We then discuss the desired properties of an ALP method, followed by the development of our proposed ALP approach. In the empirical evaluation section, we discuss in detail the setting and procedures we use to evaluate our ALP method's performance compared to alternatives in different settings. We then report our results, followed by conclusions and a discussion of the implications of our work and directions for future research.

2 Prior work

Prior work did not consider how to determine and continuously adapt the payments offered to labelers before they undertook a task, so as to cost-effectively produce a given predictive performance. Existing work about online labor markets in particular discussed a variety of mechanisms to improve the work quality, such as ways to screen workers, task design, and methods for acquiring multiple labels for the same instance so as to increase the probability of obtaining a correct label (e.g., Kazai, 2011; Downs et al., 2010; Lee et al., 2013; Paolacci et al., 2010).

One stream of research, which assumes that labeling payments are fixed and pre-determined, focuses on repeated acquisition of multiple labels for the same instance. Such repeated acquisitions can be used, for example, to infer the most likely label from multiple noisy ones. In contrast to the study we present here, these prior works do not address the problem of determining the payments to offer workers. Within this stream of research, some studies suggest methods to infer the likely label so as to learn better models using data instances that undergo repeated labeling for pre-determined payment (Dalvi et al., 2013; Kumar and Lease, 2011; Raykar et al., 2010; Rodrigues et al., 2013; Zhou et al., 2012). Some repeated labeling methods have also been applied together with active learning methods, in order to reduce the number of instances for which multiple labels are acquired for a pre-determined payment (Ipeirotis et al., 2013; Karger et al., 2011, 2014; Sheng et al., 2008; Wauthier and Jordan, 2011).

Our work differs from this stream of research in several important ways. As noted earlier, the key difference is that prior work did not aim to identify advantageous payments to be offered to crowd workers so as to cost-effectively yield a given predictive performance. Some prior work proposed means to assess work quality retrospectively, after workers have performed the task (e.g., Raykar et al., 2010; Wang et al., 2017). In this work, we aim to identify advantageous payments offered to workers before they agree to take on the task. We also do not assume the availability of individual worker-performance history. Therefore, the method we propose is suitable for recommending payments in popular crowdsourcing labor-market settings such as Amazon Mechanical Turk, where employers continuously encounter new workers, and where tasks and the corresponding payment are offered to all workers who meet certain criteria (e.g., workers are from a certain country). We also consider settings where the work quality that can be obtained for a given payment may not remain the same indefinitely, but rather that market conditions, and consequently the labeling quality that can be obtained for different payment levels, may vary over time. To remain cost-effective over time, our approach continuously adapts the payment so as to yield cost-effective improvements in model performance. Finally, in contrast to prior work that aimed to improve label quality efficiently per se (e.g., Wang et al., 2017), we aim to cost-effectively improve the predictive performance of the model induced from the acquired labels. Indeed, as we discuss in more detail below, the same quality of labels may have different implications for model performance across modeling tasks, and certainly across different points along the learning curve. Hence, our approach proposes to assess directly the cost-effectiveness of different label acquisitions toward the model's performance.

The ALP problem we consider here is also related to active learning (e.g., Ramirez-Loaiza et al., 2016; Sharma and Bolgic, 2016; Lewis and Gale, 1994; Abe and Mamitsuka, 1998; Kong and Saar-Tsechansky, 2014). Active learning methods typically assume that labels for unlabeled instances can be acquired at some fixed, pre-determined cost, and most studies also assume that the acquired labels are also correct; consequently, active learning methods aim to identify the training instances for which to acquire labels, so as to produce the best model performance for any given number of acquisitions. Because ALP aims to determine how much to offer for labels to produce cost-effective models, active learning addresses a complementary problem of identifying the instances for which to acquire labels, assuming the cost is predetermined. Importantly for the challenge we address here, because active learning methods also assume that the acquired labels are correct, such methods do not consider the challenge of estimating the benefit of alternative acquisitions, when the labels acquired are noisy.

Our study is perhaps most closely related to work by Yang and Carbonell (2012), but they make assumptions that do not hold in real online labor markets, which we consider here. Specifically, Yang and Carbonell (2012) assume settings in which labelers' performance quality for a given price is predetermined and known; thus, they do not consider settings where new crowd workers with no prior history are encountered continuously.

In what follows, we discuss the desired properties of an Adaptive Labeling Payment approach in our setting, followed by our proposed ALP approach from which we derive a specific ALP method.

3 The Adaptive Labeling Payment (ALP) approach

Perhaps the most important property of an ALP method is that it should effectively learn and continuously adapt to identify advantageous payments in a data-driven manner. In light of prior research findings on different relationships between payment and quality in different settings, one cannot assume any given tradeoff between payment and quality for either an arbitrary task or for market conditions. Therefore, a data-driven approach should identify cost-effective payments without making assumptions on the prevailing tradeoff. Furthermore, because market conditions, and consequently the relationship between pay and quality, may vary over time, an ALP approach should also continuously adapt to changing market conditions so as to remain cost-effective.

The goal of an ALP method is to improve the generalization performance of a supervised learning model induced from acquired labeled data. Thus, rather than merely assess the effect of payment on the quality of the labels, a key principle of our ALP approach is that it aims to directly estimate the tradeoff between labeling payment and the model's ultimate generalization performance.

As mentioned above, it is useful for an ALP method to estimate the model's generalization performance and how it may be affected by labels acquired at different payments. While a model's generalization can be estimated by evaluating its performance on an accurately labeled test set, the labels available in our setting are acquired via online crowdsourcing markets, and are thus inherently noisy. An alternative is to acquire "gold standard" data. These, however, are costly and may also become obsolete when the underlying concept being learned changes over time. Therefore, in this paper, we aim to develop an ALP method that assesses generalization performance directly from noisy labels acquired in online crowdsourcing markets, without the acquisition of additional data for model evaluation.

Finally, note that because ALP methods aim to determine advantageous labeling payments, they can be applied alongside methods addressing complementary problems, such as the acquisition of multiple labels for the same instance, as demonstrated in our empirical evaluations, or active learning for selecting advantageous instances to label.

3.1 Adaptive labeling payment algorithm

In this section, we develop an ALP algorithm to select adaptively the payment at which labels are acquired so as to cost-effectively improve the resultant predictive model's performance. Table 1 lists the key notations we use throughout this section. The approach we propose is iterative, where at each phase $\mathrm{k} = 1,2, \ldots, \mathrm{po}$, the labels of b instances are acquired on the market at pay level $\mathrm{c_k} \in \{\mathrm{c_1}, \ldots, \mathrm{c_{po}}\}$. At each phase, the ALP method determines the payment level at which the most cost-effective improvement in the model's generalization performance can be obtained. The acquisition o f labels proceeds sequentially until either the budget is exhausted or the model reaches a desired level of performance.

Table 1 Key Notations

Notation	Description
$\mathrm{C} = \{\mathrm{c_k}\}, \mathrm{k} = 1, .., \mathrm{po}$	Set of alternative payments per label, po is the number of payment options
n	Number of labeled instances acquired
Tc	Total cost incurred for labeling payments
$\mathrm{S_n}$	Set of labeled instances
b	Number of instances labeled at each acquisition phase
$\mathrm{B}^{\mathrm{j}}_{\mathrm{c_k}} \subset \mathrm{S_n}$	j-th random draw of b instances from $\mathrm{S_n}$, labeled previously at payment $\mathrm{C_k}$
m	Number of subsets $\mathrm{B}^{\mathrm{j}}_{\mathrm{c_k}}$
h	Number of recent acquisitions used to determine payment
$\mathrm{M}(\cdot)$	Model induced from training set $(\cdot)$
budget	Labeling acquisition budget
init_batch	Number of batches of labeled instances for initialization

To compile an initial dataset from which learning proceeds, our ALP algorithm is initialized by acquiring an equal number of labels at each pay rate (and unknown qualities), so as to yield a representative set of instances labeled at different levels of pay. Subsequent acquisitions are made iteratively, such that at each acquisition phase, the expected benefits to performance from each payment alternative are estimated, and then the next batch of b labels are acquired at the pay that is estimated to yield the best expected outcomes per acquisition cost.

Because our aim is to improve the model's predictive accuracy in a cost-effective manner, our ALP approach aims first to directly estimate the expected impact on model performance of labels acquired at different payments: $\{\mathrm{c_1}, \ldots, \mathrm{c_{po}}\}$. Specifically, ALP aims to estimate the expected model performance if the next batch of labels is acquired at each alternative payment level per label, and then acquire the next batch of labels at the payment expected to yield the greatest improvement per unit cost.

To estimate the expected model performance after acquisitions at different payments, one may consider an approach inspired by Roy and McCallum (2001) for active learning. In particular, recall that while active learning methods do not address the problem of determining which payments to offer, active learning methods do consider the problem of acquiring labels for instances at a fixed predetermined cost so as to improve a predictive model (e.g., Ramirez-Loaiza et al., 2016; Sharma and Bilgic, 2016; Roy and McCallum, 2001; Kong and Saar-Tsechansky, 2014). As such, active learning methods aim to acquire labels for unlabeled instances, which are likely to be particularly beneficial for model induction.

To estimate the expected benefit to induction from acquiring the label of a given instance, Roy and McCallum (2001) proposed to add to the training set the candidate instance (whose label is unknown) along with a possible label, induce a model M' from the augmented training set, and estimate generalization performance of the model, M', induced from the augmented set. Because the true labels are unknown, Roy and McCallum proposed that the contribution to performance from the prospective acquisition would be estimated in expectation, over all possible labels that an instance may have. Toward that expectation estimation, the probability of each possible label for a given instance is estimated from the available data. In principle, one may consider a similar approach for ALP. However, to estimate the performance change in expectation from adding instances labeled at different payments, for each payment level, and for each instance, it is necessary to estimate the likelihood that labelers will produce each possible label. This estimation is affected, not only by an instance's unknown, true label, but also, for different labeling payments, by the (unknown) likelihood of labeling error. Consequently, in our setting, there is significantly more uncertaintly about the label produced by a labeler for any given instance and at any given payment, thereby rendering particularly uncertain the estimation of the expected change in performance from acquiring labeled instances at different payments.

Rather than add prospective acquisitions and then estimate the likelihood of labels that will be acquired for these instances for different payments, we propose an approach that relies on omitting labels acquired previously for a given payment. Importantly, the benefit of considering the value to induction from instances whose labels were previously acquired for a given payment is that these labels are already known and need not be estimated.2

Specifically, to estimate the effect on model performance from acquiring the next batch of b labels at a pay c_k, at the current point along the learning curve, we

[2] Note that, a trivial approach to use known labels acquired previously for pay c_k, is to approximate the expected performance improvement from labels acquired at a given pay c_k $(k = 1,2, \ldots, po)$ by the improvement in performance already observed, when labels were previously acquired at the corresponding payment c_k. However, doing so would be misleading because the same set of labels acquired for payment c_k can yield rather different benefits to model induction if acquired at different times because different labeled (training) instances would be available in each case, that will invariably effect the marginal impact of additional acquisitions.

propose an approach that approximates the anticipated change in model performance by the change in performance resulting from removing from the current training data a set of b instances previously acquired at the corresponding pay c_k. The motivation underlying our omission-based approach is that if labeled instances acquired at pay c_k are more beneficial for induction, their omission from the training data would result in a greater drop in model performance at the current point along the learning curve.

Formally, let s_n denote the set of all labeled training instances acquired so far, let $s_n \backslash B_{c_k}$ denote the set of training instances after removing from s_n a subset B_{c_k} of b instances labeled at pay c_k, and $M(\cdot)$ be a model induced from training set $(\cdot)$ via the induction technique M. ALP approximates the expected change in model performance from acquiring a batch of b instances labeled at payment c_k by the Expected Performance Improvement (EPI_{c_k}): the difference between the estimated performance of the current model, $M(s_n)$, and the estimated performance of a model, $M(s_n \backslash B_{c_k})$, induced after omitting the set B_{c_k} of b instances previously labeled at pay c_k, given by:

$$(1) \quad EPI_{c_k} = \text{Performance}(M(s_n)) - \text{Performance}\left(M\left(s_n \backslash B_{c_k}\right)\right)$$

In Equation 1, $\text{Performance}(M(\cdot))$ corresponds to any relevant measure of model performance, such as Area Under the ROC Curve. The more advantageous a batch of labels acquired at pay c_k is for learning, the greater the drop in performance between model $M(s_n)$, induced from all labeled data acquired so far, and model $M(s_n \backslash B_{c_K})$, induced after removing the subset B_{c_k}.3

Figure 1 illustrates our proposed approximation of the expected impact on performance of new instances labeled at payment c_1 and c_2. The learning curve shows the model's performance, measured here by the Area Under the ROC Curve (AUC) obtained at different acquisition phases, and evaluated over the available training data. Point A corresponds to the AUC obtained by a model induced from all the data acquired so far, $Performance(M(s_n))$, and points 1 and 2 reflect $Performance\left(M\left(s_n \backslash B_{c_1}\right)\right)$ and $Performance\left(M\left(s_n \backslash B_{c_2}\right)\right)$—namely, the AUC of the models induced after excluding a batch of labels acquired at payment c_1 and c_2, respectively. As shown, omitting a set of labels acquired at payment c_1 results in a more significant drop in performance. ALP assesses the effect on modeling performance of acquiring new labeled instances at different payment levels $c_k \in C$, and selects the payment level that yields the most cost-effective improvement in performance. Specifically, the particular measure we propose to capture the cost-effectiveness of a labeling payment, called "Maximum Total Ratio" (MTR), simply corresponds to the ratio between the (estimated) model's performance after acquiring b labels at payment c_k, and the total labeling cost incurred after the next batch of b labels is acquired. Formally, at each acquisition phase and for each payment option, c_k, our ALP approach, henceforth refers to as ALP-MTR, computes the cost-effectiveness of acquiring a batch of b

[3] Note that this notion holds for all learning curves, including those that are not strictly increasing. For example, in unusual cases when model performance decreases with more instances, our method can identify that ommitting instances may improve performance.

instances labeled at payment $\mathrm{c_k}$ by:

$$(2)\quad \mathrm{MTR}_{\mathrm{c_k}} = \frac{\mathrm{Performance(M(s_n))+EPI}_{\mathrm{c_k}}}{\mathrm{Tc+(b\cdot c_k)}},$$

where $\mathrm{T_c}$ denotes the total labeling payments incurred thus far. At each phase, ALP-MTR selects the payment option $\mathrm{c_k}$, such that it yields the maximum $\mathrm{MTR}_{\mathrm{c_k}}$—namely, the selected payment $\mathrm{c_k^*}$ is given by $\mathrm{c_k^*} = \underset{\mathrm{c_k}}{\mathrm{argmax}}\,\{\mathrm{MTR}_{\mathrm{c_k}}\}$.

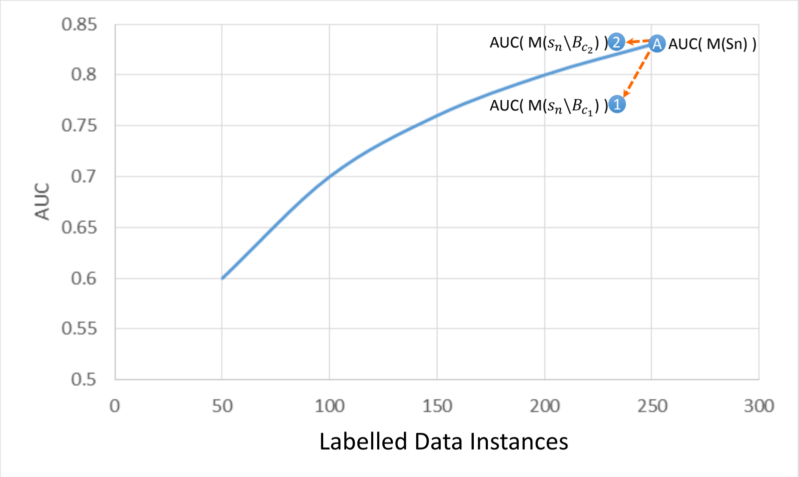


Fig. 1 Illustration of ALP's approximation of the expected impact on performance from acquiring labels at payment $\mathrm{c_1}$ and $\mathrm{c_2}$.

In what follows, we discuss the remaining elements of our ALP approach, address how it adapts to changing market conditions, and describe how our ALP-MTR approach aims to improve the estimation of model performance, $\mathrm{Performance(M(s_n))}$, in the presence of noisy labels.

3.2 Estimating EPI with noisy data

A key challenge in estimating a model's performance, and, consequently, a challenge in estimating $\mathrm{EPI}_{\mathrm{c_k}}$ (Equation 1) as well, is the presence of noisy labels. Recall from our earlier discussion, that we aim to use exclusively (noisy) labels acquired via crowdsourcing, without relying on the availability of "gold standard" labels. Relying on labels acquired from crowdworkers adversly affects the accuracy of the model performance's estimation. In particular, errors in the labels contribute to a higher estimation variance, compared to when the estimation relies on training instances that are correctly labeled. To improve the estimation of model performance, our approach incorporates several elements that aim to reduce the variance in the model's performance estimation.

Algorithm 1 ALP-MTR

```
Input:
Stopping criterion: Exhausting acquisition budget
I: Classifier inducer of choice, that returns class probability estimates
b: number of labeled instances acquired at each acquisition phase (batch size)
$\mathrm{h}, \mathrm{m}, \mathrm{C}$, init_batch

#Initialization:
Tc = 0   # total cost so far
Initial labeled set of instances: $\mathrm{S}_0 = \{\}$
Repeat init_batch times
    $\forall \mathrm{c}_\mathrm{k} \in C$:
    Purchase a batch of $b$ labeled instances at payment $\mathrm{c}_\mathrm{k}$ and add them to $\mathrm{S}_\mathrm{n}$
    $\mathrm{T}_\mathrm{c} \Leftarrow \mathrm{T}_\mathrm{c} + \mathrm{b} \cdot \mathrm{c}_\mathrm{k}$
End repeat

# Select payment
While Stopping criterion is not met (Total_pay≤ budget)
    $\mathrm{Performance}(\mathrm{M}(\mathrm{s}_\mathrm{n})) = \mathrm{AUC}(\mathrm{M}(\mathrm{s}_\mathrm{n}))$
    For $k = 1, \ldots, po$      # over all payment alternatives
        $\mathrm{D}_{\mathrm{c}_\mathrm{k}} \leftarrow$ from $\mathrm{S}_\mathrm{n}$ select h instances labeled at payment $\mathrm{c}_\mathrm{k}$ most recently
        For $j = 1, \ldots, m$
            $\mathrm{B}^{\mathrm{j}}_{\mathrm{c}_\mathrm{k}} \leftarrow$ randomly draw b instances from $\mathrm{D}_{\mathrm{c}_\mathrm{k}}$
            Compute AUC $\left(\mathrm{M}\left(\mathrm{S}_\mathrm{n} \setminus \mathrm{B}^{\mathrm{j}}_{\mathrm{c}_\mathrm{k}}\right)\right)$ via crross validation over labeled set $\mathrm{S}_\mathrm{n}$
        End For
        $\mathrm{Performance}\left(\mathrm{M}(\mathrm{s}_\mathrm{n} \backslash \mathrm{B}_{\mathrm{c}_\mathrm{k}})\right) = \frac{\sum_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{AUC}\left(\mathrm{M}\left(\mathrm{s}_\mathrm{n} \backslash \mathrm{B}^{\mathrm{j}}_{\mathrm{c}_\mathrm{k}}\right)\right)}{\mathrm{m}}$ # Calculate average performance
        $\mathrm{EPI}_{\mathrm{c}_\mathrm{k}} = \mathrm{Performance}(\mathrm{M}(\mathrm{s}_\mathrm{n})) - \mathrm{Performance}\left(\mathrm{M}(\mathrm{s}_\mathrm{n} \backslash \mathrm{B}^{\mathrm{j}}_{\mathrm{c}_\mathrm{k}})\right)$
    End For
    If Max$\{\mathrm{EPI}_{\mathrm{c}_\mathrm{k}}\} \geq 0$ :  $\mathrm{best\_pay} = \mathrm{argmax}_{\mathrm{c}_\mathrm{k}} \left(\frac{\mathrm{Performance}(\mathrm{M}(\mathrm{s}_\mathrm{n})) + \mathrm{EPI}_{\mathrm{c}_\mathrm{k}}}{\mathrm{Tc} + (\mathrm{b} \cdot \mathrm{c}_\mathrm{k})}\right)$
        else : best_pay = argmax$_{\mathrm{c}_\mathrm{k}}\{\mathrm{EPI}_{\mathrm{c}_\mathrm{k}}\}$
    Purchase b labeled instances at pay best_pay and add them to $\mathrm{S}_\mathrm{n}$
    $\mathrm{T}_\mathrm{c} \Leftarrow \mathrm{T}_\mathrm{c} + \mathrm{b} \cdot \mathrm{best\_pay}$
End while
```

Recall that our approach approximates the expected change in performance from acquiring b labels at a cost c_k by estimating $\mathrm{Performance}\left(\mathrm{s}_\mathrm{n} \backslash \mathrm{B}_{\mathrm{c}_\mathrm{k}}\right)$; namely, by estimating the performance of a model induced after omitting from our training data b instances labeled at payment c_k. To reduce the estimation variance in estimating $\mathrm{Performance}\left(\mathrm{s}_\mathrm{n} \backslash \mathrm{B}_{\mathrm{c}_\mathrm{k}}\right)$, we repeat the estimation multiple times by excluding random subsets of b instances acquired previously at payment c_k, estimating the model's performance after this omission. $\mathrm{Performance}\left(\mathrm{s}_\mathrm{n} \backslash \mathrm{B}_{\mathrm{c}_\mathrm{k}}\right)$ is then estimated by averaging the models' performances measured over the multiple repetitions. Specifically, we randomly draw *with replacement* m subsets, $\mathrm{B}^{\mathrm{j}}_{\mathrm{c}_\mathrm{k}}, \mathrm{j} \in 1 \ldots \mathrm{m}$, of b instances previously labeled at pay c_k. At each repetition, j, a different subset $\mathrm{B}^{\mathrm{j}}_{\mathrm{c}_\mathrm{k}}$ is removed from s_n and a measure of performance of the model $\mathrm{M}\left(\mathrm{s}_\mathrm{n} \backslash \mathrm{B}^{\mathrm{j}}_{\mathrm{c}_\mathrm{K}}\right)$, induced from the reduced training set $\mathrm{s_n \backslash B^{\,j}_{c_K}}$, is estimated via cross-validation. For now, we assume that the Area Under the ROC Curve, denoted $\mathrm{AUC(M(\cdot))}$, is the relevant performance measure for a model M induced from training set $(\cdot)$; however, any other performance measure of interest can be used. The expected model's performance after omitting a batch of b instances labeled at price $\mathrm{c_k}$ is estimated as the average over the m repeated experiments above:

$$\mathrm{Performance}\left(\mathrm{M}\left(\mathrm{s_n \backslash B_{\,c_k}}\right)\right) = \frac{\sum_{j=1}^{m} \mathrm{AUC}\left(\mathrm{M}\left(\mathrm{s_n \backslash B^{\,j}_{c_k}}\right)\right)}{\mathrm{m}}.$$

Finally, to reduce further the error in the estimation of both $\mathrm{AUC(M(s_n))}$ and $\mathrm{AUC}\left(\mathrm{M}\left(\mathrm{s_n \backslash B^{\,j}_{c_k}}\right)\right)$, we perform repeated applications (with different random seeds) of cross-validation, and estimate the model's performance as the average performance over multiple applications of cross-validation.

3.3 Continuous adaptation of payments

Recall that, changes in the market settings over time, such as those due to changes in the population of workers and competitive settings, may give rise to different tradeoffs between pay and labeling quality. To adapt labeling payments to the prevailing tradeoff, ALP estimates the EPI (Equation 1) at pay $\mathrm{c_k}$, based on a subset of recently labeled instances. More specifically, recall that our ALP approach estimates EPI from labels acquired at different payment levels, $\mathrm{c_K}$, by estimating the average change in performance between a model induced from the complete set of instances, $\mathrm{M(s_n)}$, and a model, $\mathrm{M}\left(\mathrm{s_n \backslash B^{\,j}_{c_k}}\right)$, induced after omitting b labeled instances acquired at payment $\mathrm{c_k}$. To facilitate adaptation, rather than draw subsets $\mathrm{B^{\,j}_{c_K}}$ from all prior acquisitions at pay $\mathrm{c_k}$, $\mathrm{B^{\,j}_{c_k}}$ is drawn from a set $\mathrm{D_{c_k}}$ of the most recent h instances acquired at pay $\mathrm{c_k}$ ($\mathrm{B^{\,j}_{c_k} \subset D_{c_k}}$), and we subsequently evaluate their average impact on performance. In the empirical evaluations that follow, we evaluate the benefits of this feature in settings where the tradeoff between labeling payment and labeling quality changes over time, as well as when the tradeoff is stable.

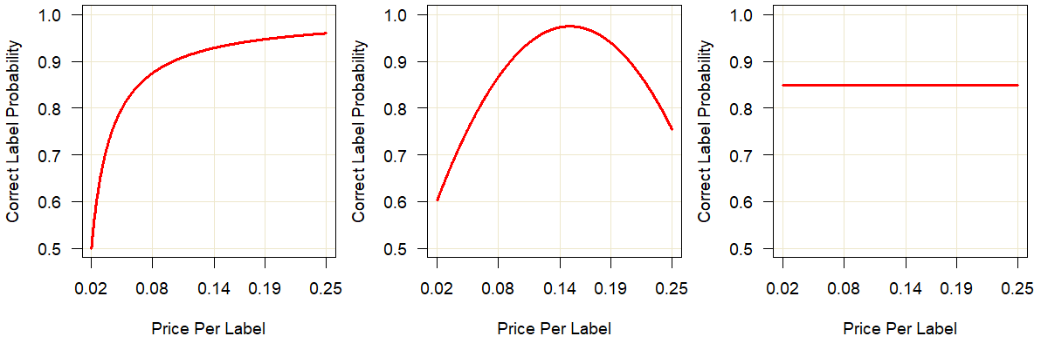
Algorithm 1 outlines the pseudo code for ALP-MTR. Note that our proposed approach estimates the cost-effectiveness of different payment alternatives by the Maximum Total Ratio measure (Equation 2), but other ALP methods may use other measures of cost-effectiveness.

4 Empirical evaluations

We evaluate our method's performance, compared to alternatives, under a variety of settings that correspond to different labeling tasks and different tradeoffs between payment and quality that have been reported in prior work as arising in outsourcing labor markets. In addition, because different tradeoffs may arise at different times, we also evaluate our approach's performance when the prevailing tradeoff between payment and quality changes over time. Across settings, we aim to identify whether a given method yields a robust performance. Specifically, a robust ALP method that can be relied on in practice ought to yield consistent performance across settings, and should not fail miserably in some settings.

As noted above, we evaluate the cost-effectiveness of our approach under different tradeoffs between payment and quality identified in prior work about online labor markets, illustrated in Figures 2(a)-2(c). In Figure 2 and throughout the paper, quality is characterized by the likelihood of error. Specifically, Kazai (2011), Kazai et al. (2013), and Feng et al. (2009) have found concave tradeoffs between payment and quality, illustrated in Figure 2(a), where quality improves initially with increasing payments, but then degrades beyond a certain payment level. In a different set of experiments, Kazai (2011) and later Kazai et al. (2013) have found an "asymptotic" tradeoff between payment and quality, illustrated in Figure 2(b). Mason and Watts (2010) have documented a fixed tradeoff, illustrated in Figure 2(c), where different payments yield the same quality. Because different tradeoffs arise in different market settings that can vary over time, it is also useful to examine the robustness of an ALP approach when the underlying tradeoff between payment and quality shifts from one tradeoff to another. In the evaluations reported below, we thus report ALP-MTR‘s performance when a given tradeoff switches to another after 50% of the acquisition budget has been used.

In the experiments we report below, once the payment to be offered to labelers for the next batch of labels is selected, the probability q that a label acquired at that payment is correct is determined by the prevailing tradeoff. Thus, the correct label for an instance is assigned at probability q, and the label is reversed at probability $(1 - q)$. We also followed prior work for the payment range and the number of payment alternatives from which payment methods can select. Specifically, prior studies considered 2-4 payment levels, ranging from a minimum payment of $0.01-$0.03 to a maximum payment of $0.1-$0.25 (e.g., Feng et al., 2009; Kazai, 2011; Kazai et al., 2013; Mason and Watts, 2010; Rogstadius et al., 2011); thus, in the evaluations that follow, we considered three payment alternatives within this range of $0.02, $0.14, and $0.25, reflecting low, mid, and high payment levels.



(a) "Asymptotic" tradeoff (b) "Concave" Tradeoff (c) "Fixed" Tradeoff

Fig. 2 Different tradeoffs between payment and quality observed in real experiments

For each of the tradeoffs outlined above, as well as for settings where the tradeoff changes over time, we report experimental results for labeling tasks corresponding to three publicly available datasets. The first two, Pen Digits and SPAM (Lichman, 2013), reflect typical labeling tasks that are easy for humans to produce. The third dataset, Mushroom

(Lichman, 2013), has been used in prior work on labeling via crowdsourcing markets (e.g., Ipeirotis et al., 2013).

4.1 Evaluation procedure and measures

Our evaluation procedure is illustrated in Figure 3 and includes three modules: (a) Payment Method, (b) Prevailing Tradeoff in the market at any given time between payment and quality, and (c) Reporting and Measurement.

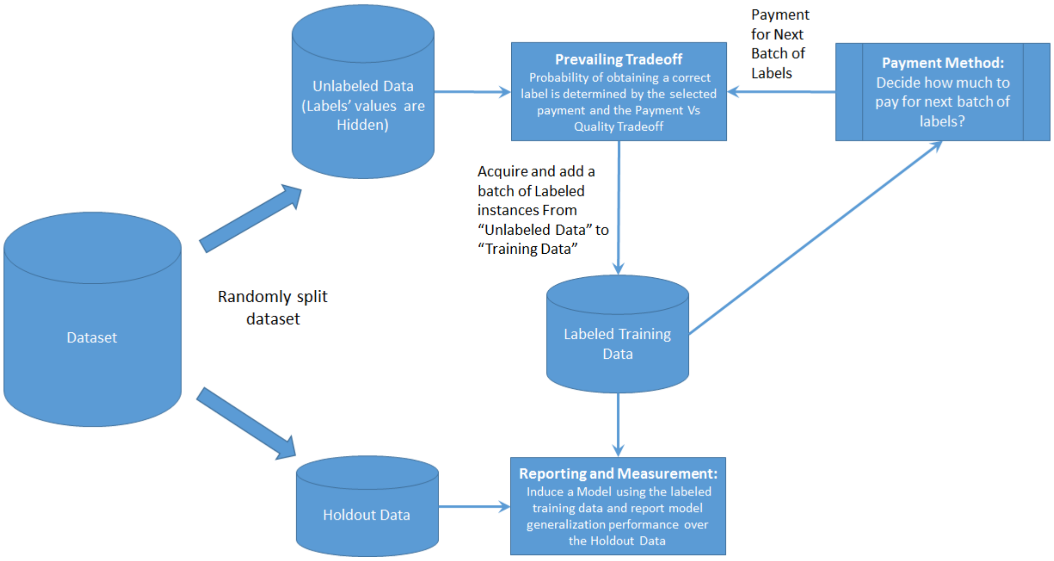

Payment
for Next
Batch of
Labels
Unlabeled Data
(Labels' values are
Hidden)
Prevailing Tradeoff
Probability of obtaining a correct label is determined by the selected payment and the Payment Vs Quality Tradeoff
Payment Method:
Decide how much to pay for next batch of labels?
Acquire and add a
batch of Labeled
instances From
"Unlabeled Data" to
"Training Data"
Dataset
Randomly split
dataset
Labeled Training
Data
Holdout Data
Reporting and Measurement:
Induce a Model using the labeled training data and report model generalization performance over the Holdout Data

Fig. 3 Overview of the evaluation process

The evaluation begins by randomly partitioning a labeled dataset into an Unlabeled dataset (70% of instances), for which the true labels are hidden and later can be acquired by different payment methods, and a Holdout dataset (the remaining 30%) containing instances with the correct labels. Note that the correctly labeled holdout set is used exclusively for the evaluation of the models produced from labeled instances acquired by the alternative approaches and is not used by any approach to inform its payment selection. For control, the same holdout set is used to evaluate ALP-MTR and baseline methods. At each iteration of batch label acquisition, the payment for purchasing the labels of a batch of 10 unlabeled data instances from the Unlabeled set is selected by a payment method (e.g., ALP-MTR or baseline), and the labeled instances are added to the corresponding method's training set. For control, at each iteration (acquisition phase), the labels of the same instances, drawn at random from the unlabeled set, are acquired by ALP-MTR and the baseline. Hence, any difference in cost-effectivness can be attributed exclusively to the payment selected by each approach and the corresponding labeling quality produced as a result.

For each approach, once a new set of labeled instances is acquired and added to the corresponding approach's training set, a new model is induced from the augmented training set, and we report its performance on the correctly labeled holdout data set, so as to assess the improvement in model performance achieved by each approach from all of its label acquisitions.

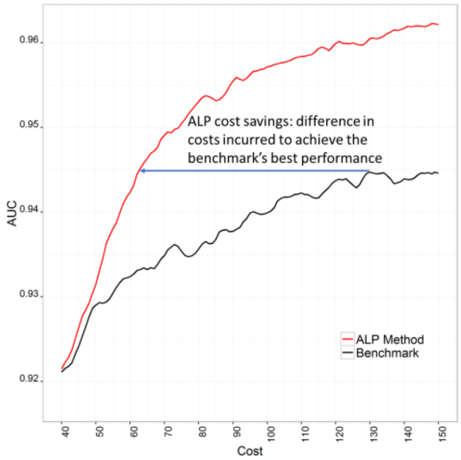


Fig. 4 Illustration of label acquisition costs incurred by two methods, A (red curve) and B (black curve), to achieve difference model performances. The horizontal line reflects the Cost Saving measure: the difference between the labeling costs incurred by B to yield its best performance, and the costs incurred by A to yield the same performance.

We report the average performance over 20 repetitions of the learned model's Area Under the ROC Curves (AUC) as a function of cumulative labeling costs incurred by each approach to achieve this performance. AUC is evaluated over the correctly labeled holdout set.

In addition, we report the labeling Cost Saving (CS) enabled by ALP-MTR to yield the highest AUC obtained by the baseline alternative. CS is illustrated using an stylized plot in Figure 4: The benchmark method (B) incurs costs of $130 to obtain its highest AUC of 0.945, while the ALP method (A) achieves the same level of performance after incurring $63, yielding cost savings of 52%.

We also report the statistical significance of the difference between the areas under the AUC curve as a function of the labeling costs generated by two competing methods. To this end, we use the BCA bootstrap method implemented in R, with 10,000 repetitions.

In the experiments reported here, we induced predictive models using the R implementation of the Random Forest algorithm (R Package: RandomForest). We used the default parameter setting with 100 base trees. We later replicate our main results for inducing SVM models. In the experiments reported below, ALP-MTR used eight-fold cross-validation repeated four times toward this estimation.[4] For the history parameter h , we used 100 recently labeled instances to evaluate alternative payments. We later explore the robustness of ALP-MTR's performance when varying the values of these parameters.

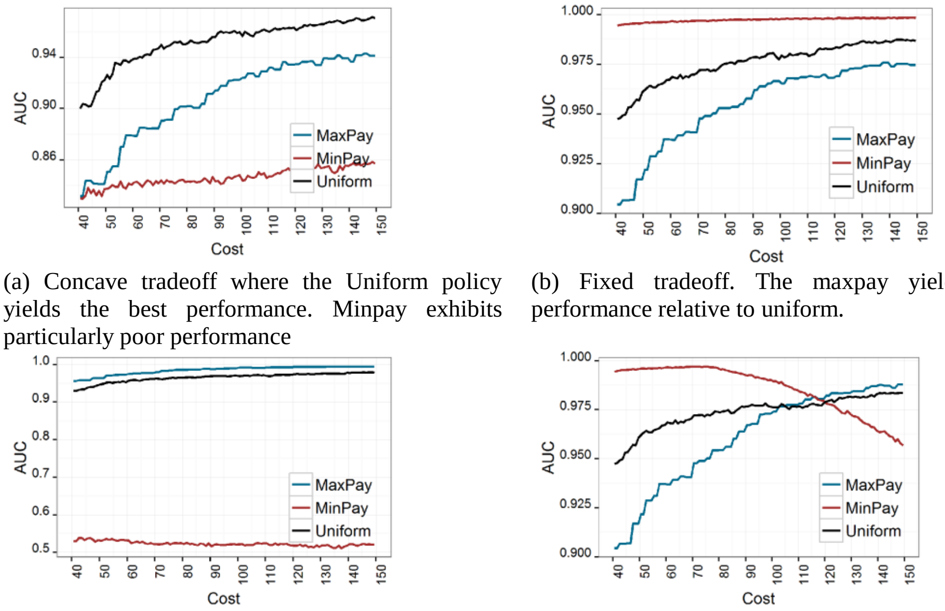
4.2 Alternative payment policies

To the best of our knowledge, there are no existing solutions to the ALP problem; hence, we consider alternative baseline payment policies, two of which are optimal under some settings.

[4] These cross-validation parameters were selected for efficiency using 32 core machines.

Specifically, the first policy, henceforth referred to as *Minpay*, always offers the lowest payment for labels. A second policy, *Maxpay*, always offers the highest payment. A third policy, henceforth referred to as *Uniform*, acquires labels at a representative set of payments—specifically, in each batch of payments, Uniform draws uniformly at random the payment at which labels are acquired. As we will demonstrate shortly, in some settings the above policies offer the optimal bias, and will yield the best model performance for a given payment. For example, Minpay is the most cost-effective policy in a setting where the prevailing tradeoff in the market between payment and quality is “fixed”—namely, when the least costly labels yield the same labeling quality as can be obtained at highest pay. Nevertheless, because our goal is to identify a robust benchmark policy that does not yield poor performance in some settings, we first examine the robustness of the three alternative policies: Minpay, Maxpay, and Uniform.

For the Pen Digits dataset, Figure 5 shows representative performances of Minpay, Maxpay, and the Uniform policy for different tradeoffs between payment and quality reported in prior work to arise in online labor markets. As shown, both Minpay and Maxpay exhibit good performance in some settings, but they also yielded particularly poor performance in other settings. For example, Figure 5(a) shows that neither Minpay nor Maxpay is optimal under a concave tradeoff between payment and labeling quality. In this setting, Minpay policy yields particularly poor performance. By contrast, Figure 5(b) shows that when labeling quality is fixed across payments, Minpay yields the best performance, and Maxpay yields the worst performance. When the tradeoff between payment and label quality is asymptotic, shown in Figure 5(c), Maxpay and Uniform produce good performances, whereas Minpay performs miserably. Overall, both Minpay and Maxpay lack robustness and they each exhibit very poor performance under some settings. By contrast, the Uniform policy yields the most robust performance, and can be relied on in practice not to fail miserably. The lack of robustness in the performances exhibited by Minpay and Maxpay policies can be even more significant when the underlying tradeoff between payment and quality changes. Figure 5(d) shows the policies’ performance when the underlying tradeoff changes from fixed quality across payments to an asymptotic tradeoff (change occurs once the labeling cost achieves $75). As shown, Maxpay yields poor performance initially in this setting, while Minpay exhibits good performance initially, but yields poor performance once the underlying tradeoff becomes asymptotic. Importantly, when the tradeoff can change over time, the approach of learning once, early on, which single payment level is best, and then acquiring labels at this payment indefinitely, does not yield good performance. In Appendix A, we present additional results for the Pen Digits and Spam datasets in settings where the tradeoff changes, as well as results for the Mushroom dataset. The complete set of experiments shows a similar pattern, where Minpay and Maxpay perform poorly under some settings, while Uniform yields robust performance across settings. Henceforth, we evaluate ALP-MTR’s performance relative to that of robust Uniform policy.



(a) Concave tradeoff where the Uniform policy yields the best performance. Minpay exhibits particularly poor performance

(b) Fixed tradeoff. The maxpay yields poor performance relative to uniform.

(c) Asymptotic tradeoff

(d) Tradeoff switches from fixed to asymptotic at the middle of cost range.

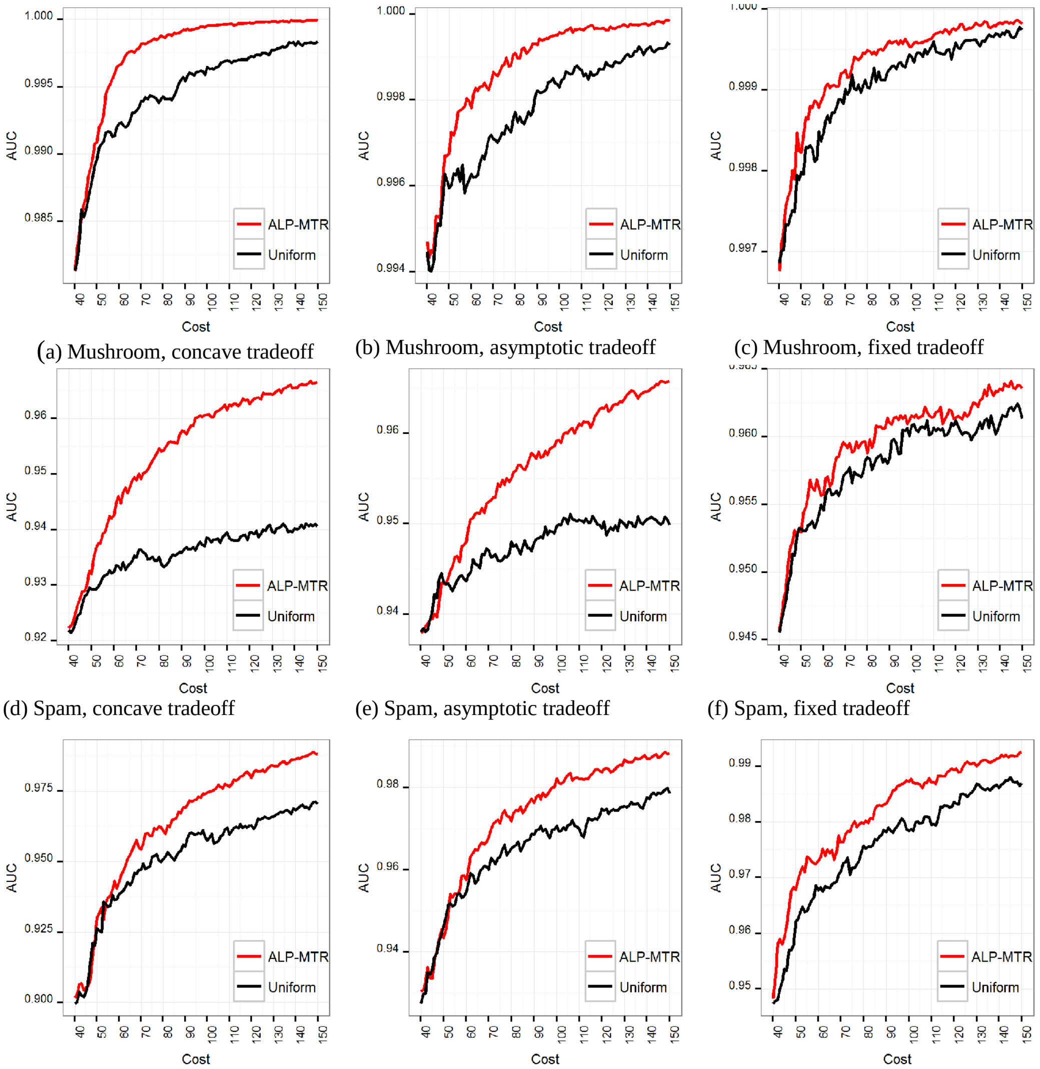
Fig. 5 Performance of alternative ALP policies for the Pen Digits dataset. Both Minpay and Maxpay lack robustness and yield poor performances under some settings. Uniform yields robust performance across settings and is never the worst policy.

5 Results

Table 2 summarizes the average cost savings achieved by ALP-MTR to yield the best performance exhibited by the UNIFORM policy for different data sets and tradeoffs. As shown, ALP-MTR yields a significant reduction in the costs necessary to yield the best model performance achieved by the UNIFORM payment policy. Specifically, Table 2 shows that, across different settings and labeling tasks, ALP-MTR yields an average of 35.6% savings in labeling costs.

Table 2 Cost Savings generated by ALP-MTR to achieve UNIFORM's best performance

Dataset	**Tradeoff**	**Cost Savings produced by ALP-MTR**
Mushroom	Asymptotic	39.9%
	Concave	50.6%
	Fixed	18.4%
Spam	Asymptotic	39.7%
	Concave	60.7%
	Fixed	15.6%
Pen Digits	Asymptotic	34.3%
	Concave	38.2%
	Fixed	23.0%
Average Savings		**35.6%**



(a) Mushroom, concave tradeoff (b) Mushroom, asymptotic tradeoff (c) Mushroom, fixed tradeoff

(d) Spam, concave tradeoff (e) Spam, asymptotic tradeoff (f) Spam, fixed tradeoff

(g) Pen Digits, concave tradeoff (h) Pen Digits, asymptotic tradeoff (i) Pen Digits, fixed tradeoff

Fig. 6 Comparison of ALP-MTR and the UNIFORM policy. ALP-MTR is often significantly better and is otherwise comparable to UNIFORM.

Figure 6 shows the performances obtained for different labeling acquisition costs incurred by ALP-MTR and the UNIFORM policy, under different labeling tasks and tradeoffs between payment and quality. As shown, the ALP-MTR method exhibits consistent performance and it is superior to the UNIFORM policy. For the learning curves presented in Figure 6, Table 3 shows significance test results for the difference between the areas under the AUC curves produced by ALP-MTR and UNIFORM and shown in Figure 6. For all datasets and underlying tradeoffs, the difference between ALP-MTR and the UNIFORM policy is statistically significant ($p < 0.01$).

Table 3 Significance Testing for Differences in Area under the AUC curves*

Dataset	**Tradeoff**	**Figure**	**Sigificance of difference between ALP-MTR and UNIFORM (*** p < 0.01)**
Mushroom	Concave	6a	***
	Asymptotic	6b	***
	Fixed	6c	***
Spam	Concave	6d	***
	Asymptotic	6e	***
	Fixed	6f	***
Pen Digits	Concave	6g	***
	Asymptotic	6h	***
	Fixed	6i	***

* Bootstrap *p*-values for significance test on difference between AUC produced by ALP-MTR and UNIFORM, shown in Figure 6. Bootstrap tests were performed using the bca method implemented in R. *** $p < 0.01$

5.1 Changing tradeoff between pay and labeling quality

A key motivation for an Adaptive Labeling Payment policy is that the tradeoff between pay and labeling quality is unknown for arbitrary labeling task and market settings, and the prevailing tradeoff in the market may also change over time (e.g., possibly due to changing market conditions). In what follows, we evaluate the proposed ALP-MTR's robustness in settings where the underlying tradeoff between pay and label quality varies. These evaluations aim to establish whether under these conditions the ALP method continues to be advantageous and to yield robust performance. In the experiments reported here, the prevailing tradeoff in the market changes once ALP incurs $75 in labeling costs.

Table 4 summarizes the average cost savings achieved by ALP-MTR relative to the UNIFORM policy when the underlying tradeoff changes from one tradeoff to another. As shown, ALP-MTR exhibits robust behavior and is superior to the UNIFORM policy, yielding a significant savings of 32.3% in labeling costs, on average. Figure 7 shows curves of the predictive performance obtained for different labeling acquisition costs. In the interest of space, we present results for a subset of the settings, and include in Appendix B results for all remaining settings. As shown, when the tradeoff varies, ALP-MTR often achieves superior performance and is otherwise comparable to the UNIFORM approach across settings. Figure 7(f) shows an interesting result in which ALP-MTR's performance briefly deteriorates soon after the underlying tradeoff changes; ALP-MTR then detects and adapts to the change, yielding cost-effective improvements in the model's performance.

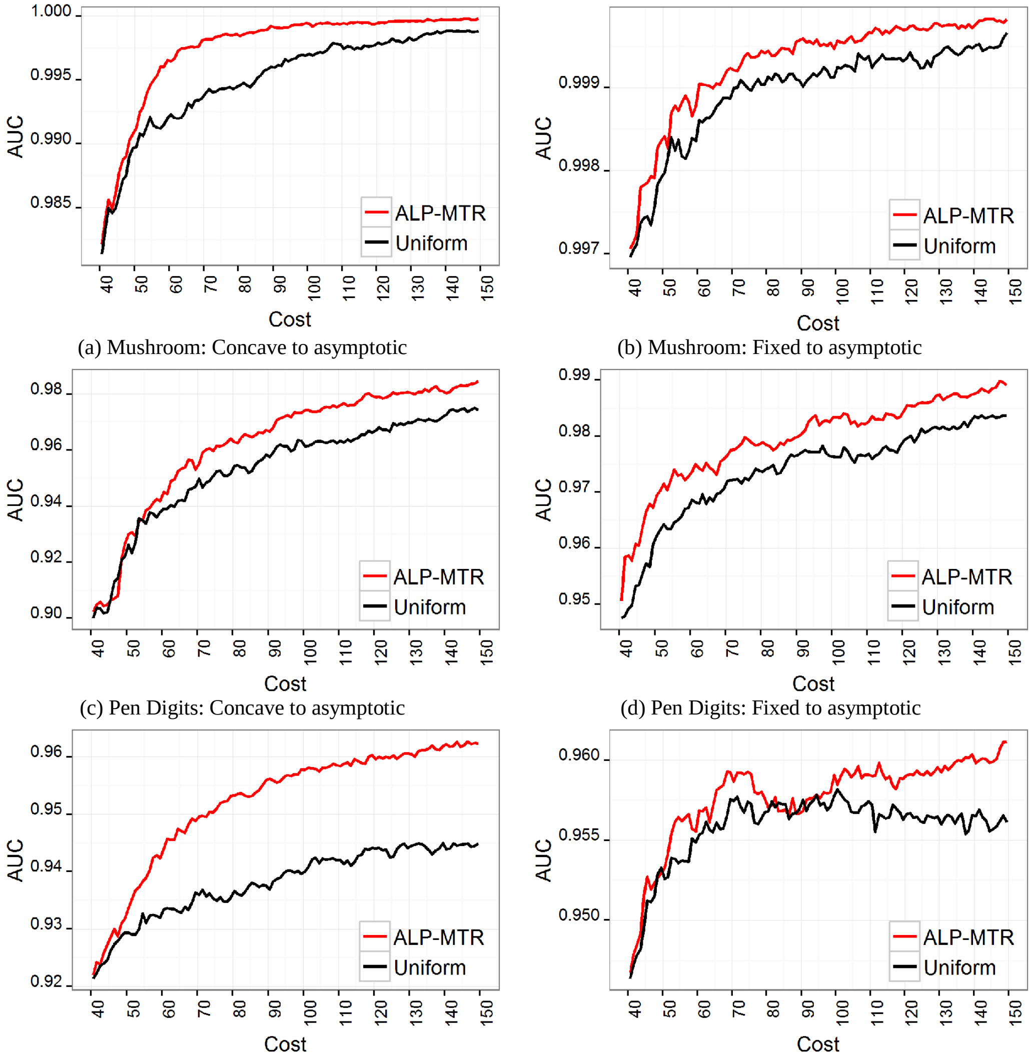
Table 4 Cost Savings achieved by ALP-MTR to yield UNIFORM'S best performance

Dataset	Change in Tradeoff (From → To)	Cost Savings produced by ALP-MTR
Mushroom	Concave → Asymptotic	38.5%
	Concave → Fixed	24.3%
	Asymptotic → Concave	46.3%
	Asymptotic → Fixed	29.1%
	Fixed → Asymptotic	26.0%
	Fixed → Concave	29.9%
Spam	Concave → Asymptotic	58.9%
	Concave → Fixed	48.1%
	Asymptotic → Concave	57.3%
	Asymptotic → Fixed	25.8%
	Fixed → Asymptotic	34.6%
	Fixed → Concave	10.4%
Pen Digits	Concave → Asymptotic	29.4%
	Concave → Fixed	9.2%
	Asymptotic → Concave	38.0%
	Asymptotic → Fixed	6.2%
	Fixed → Asymptotic	31.0%
	Fixed → Concave	38.8%
Average Savings:		**32.3%**

Table 5 Significance testing for difference in area under the ROC curves

Dataset	Change in Tradeoff Functions (From → To)	Figure	Difference between ALP-MTR and uniform
Mushroom	Concave → Asymptotic	7 (a)	***
	Concave → Fixed	B1 (a)	***
	Asymptotic → Concave	B1 (d)	***
	Asymptotic → Fixed	B1 (e)	***
	Fixed → Asymptotic	7 (b)	***
	Fixed → Concave	B1 (c)	***
Spam	Concave → Asymptotic	7 (c)	***
	Concave → Fixed	B1 (g)	***
	Asymptotic → Concave	B1 (j)	***
	Asymptotic → Fixed	B1 (k)	**
	Fixed → Asymptotic	7 (d)	
	Fixed → Concave	B1 (i)	***
Pen Digits	Concave → Asymptotic	7 (e)	***
	Concave → Fixed	B1 (m)	**
	Asymptotic → Concave	B1 (p)	***
	Asymptotic → Fixed	B1 (q)	**
	Fixed → Asymptotic	7 (f)	***
	Fixed → Concave	B1 (o)	***

Bootstrap *p*-values for significance testing of the difference between the area under the AUC curves produced by ALP-MTR and UNIFORM, shown in Figure 7 and Figure B1 (in Appendix B). Bootstrap tests were performed using bca method implemented in R. * $p < 0.1$, ** $p < 0.05$, *** p < 0.01



(a) Mushroom: Concave to asymptotic

(b) Mushroom: Fixed to asymptotic

(c) Pen Digits: Concave to asymptotic

(d) Pen Digits: Fixed to asymptotic

(e) Spam: Concave to asymptotic

(f) Spam: Fixed to asymptotic

Fig. 7 Performance of the ALP-MTR and the uNIFORM payment policy when the underlying tradeoff between labeling payment and quality changes. As shown, ALP-MTR's performance is often significantly better than that of UNIFORM, and ALP-MTR is never worse than UNIFORM.

For all the results reported in Figures 7 and B1 (Appendix B), Table 5 shows significance test results for the difference between the area under the AUC curves produced by ALP-MTR and UNIFORM. In almost all settings, the difference in the area under the curves produced by ALP-MTR and UNIFORM is statistically significant. Importantly, in the single setting where ALP-MTR did not yield statistically significant superior performance, ALP-MTR yields performance comparable to that produced by the UNIFORM policy.

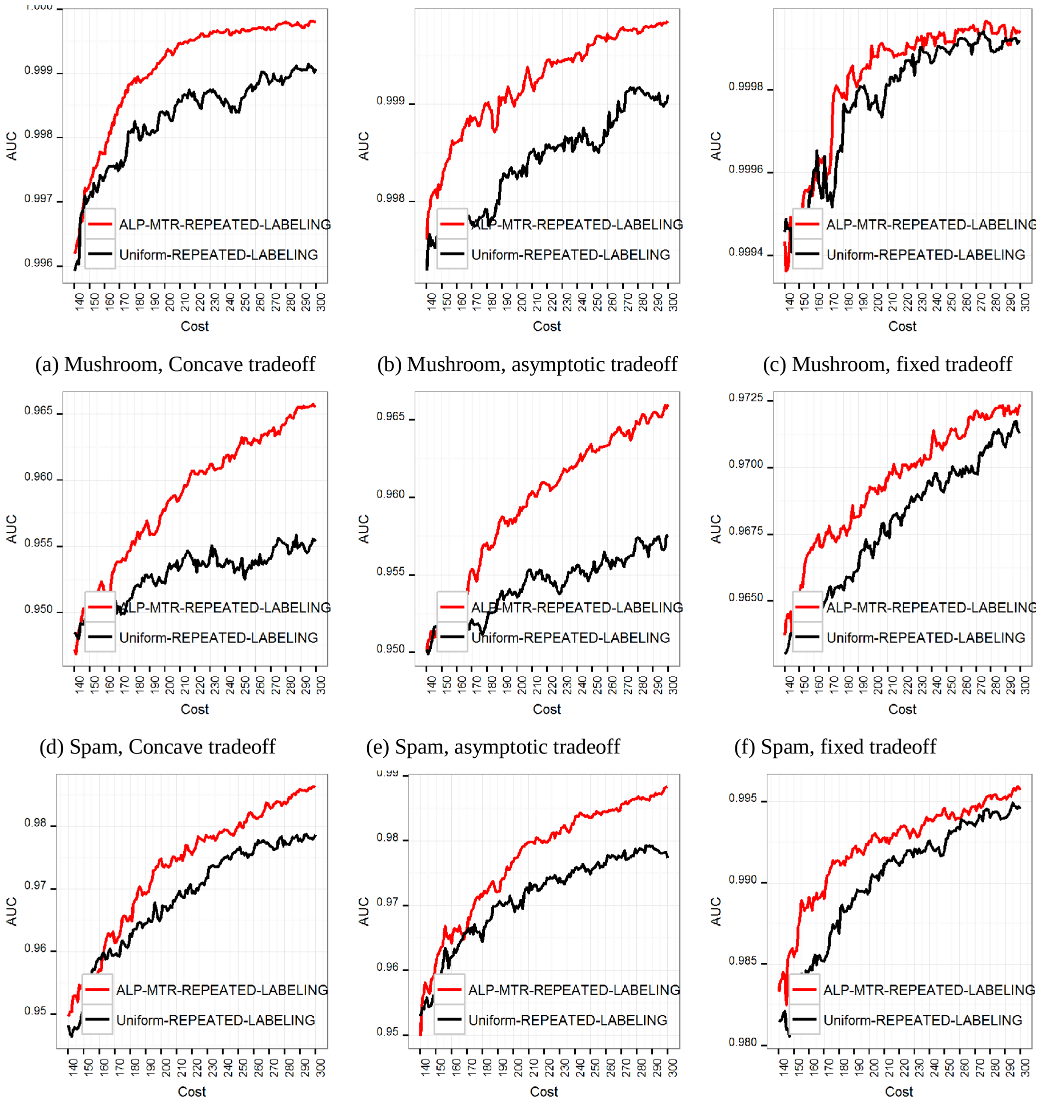
5.2 ALP-MTR with repeated labeling

As discussed earlier, repeated labeling refers to the acquisition of multiple noisy labels for the same instance so as to improve the accuracy of the labels. Methods for repeated labeling do not detemine which payments to offer for labels, but assume that the payment remains constant and is somehow pre-determined. Repeated labeling thus aims at reducing noise (error) in the data by acquiring multiple labels for the same instance at a fixed pre-determined cost. Subsequently, the most likely correct label is inferred and used for model induction. For example, a popular approach is to infer the majority label for each instance (e.g., Lee et al., 2013; Mason and Suri, 2012). Because ALP-MTR and repeated labeling address complementary goals, they can be applied in conjunction with benefit modeling. In particular, ALP-MTR can be applied first to identify the cost-effective payment at which to acquire the next batch of labels; multiple labels for the same instance can be then purchased at the payment determined by ALP-MTR and are used by repeated labeling.

Table 6 Cost savings enabled by ALP-MTR with repeated labeling relative to using UNIFORM with repeated labeling

Dataset	**Tradeoff**	**Cost savings enabled by ALP-MTR**
Mushroom	Asymptotic	31.0%
	Concave	33.9%
	Fixed	6.9%
Spam	Asymptotic	37.7%
	Concave	36.6%
	Fixed	10.8%
Pendigits	Asymptotic	27.8%
	Concave	19.0%
	Fixed	6.4%
Average Savings		**23.3%**

We evaluated the performance of ALP-MTR with repeated labeling compared to the performance of repeated labeling when the UNIFORM policy rather than ALP-MTR for selecting the payments for labels was used. In these experiments, repeated labeling refers to acquiring multiple labels per instance, and an instance is labeled by the majority of three label. Table 6 shows the average cost savings achieved by ALP-MTR, and Figure 8 shows the predictive performance achieved for different label acquisition costs by each approach. As shown, when multiple labels are acquired for repeated labeling, ALP-MTR can be used effectively to identify advantageous labeling payments, and it yields an average savings of 23% in labeling costs. Overall, using ALP-MTR to select the payments for repeated labeling often yields superior model performance for a given labeling cost and is otherwise comparable to when repeated labeling is applied naively, without ALP-MTR‘s selection of cost-effective payments.



(a) Mushroom, Concave tradeoff (b) Mushroom, asymptotic tradeoff (c) Mushroom, fixed tradeoff

(d) Spam, Concave tradeoff (e) Spam, asymptotic tradeoff (f) Spam, fixed tradeoff

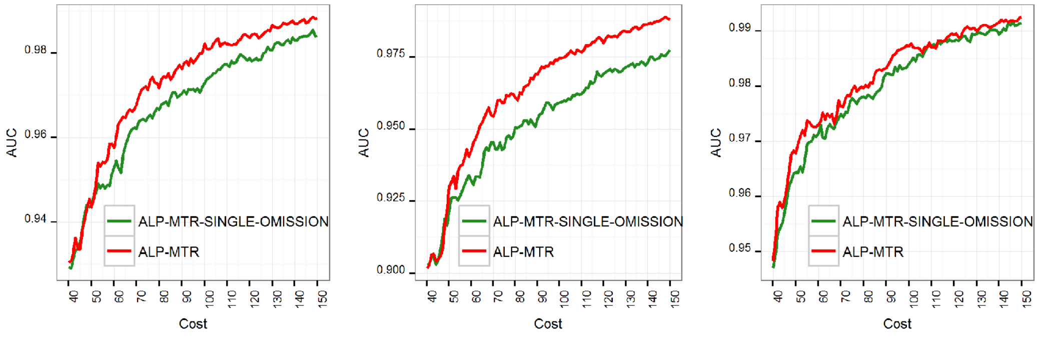
(g) Pendigits, Concave tradeoff (h) Pendigits, asymptotic tradeoff (i) Pendigits, fixed tradeoff

Fig. 8 Performance of ALP-MTR and UNIFORM with repeated labeling. Applying ALP-MTR to determine the payments used by repeated labeling yields better models for a given cost as compared to when payments are determined by UNIFORM.

5.3 Evaluation of the elements of ALP-MTR

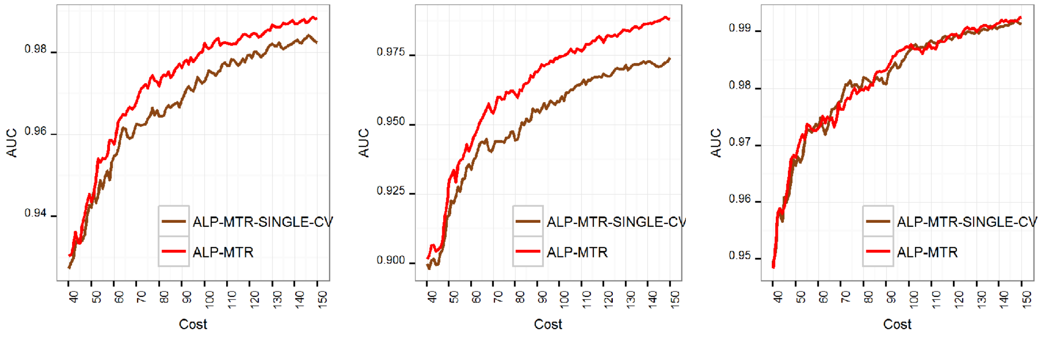
A key element of our ALP approach is the assessment of a model’s generalization performance in the presence of noisy labels. To that end, ALP-MTR performs repeated omissions of sets of instances that were previously labeled at a given payment level, drawn at random with replacement; the effect of these repeated omissions on the induced model's predictive performance is averaged.

We empirically evaluate the benefits of using repeated omission to ALP-MTR‘s selection of cost-effective payments. Figure 9 shows the performance of ALP-MTR when $m = 10$ repeated omissions are done to estimate the cost-effectiveness of each payment alternative, and the performance of an ALP-MTR variant, ALP-MTR-SINGLE-OMISSION, in which a single subset of instances is omitted toward this estimation. In the interest of space, Figure 9 shows results for the Pen Digits dataset, which are representative of the comparison across datasets. As shown, ALP-MTR‘s repeated sampling of subsets for omission benefits ALP-MTR's selection of cost-effective payments, thereby often producing better and otherwise comparable performance, compared to the effect of omitting a single set of instances.



(a) Pen Digits: Asymptotic Tradeoff (b) Pen Digits: Concave Tradeoff (c) Pen Digits: Fixed Tradeoff

Fig. 9 Comparison between ALP-MTR and ALP-MTR-SINGLE-OMISSION. ALP-MTR achieves superior or otherwise comparable performance to that of ALP-MTR- SINGLE-OMISSION.



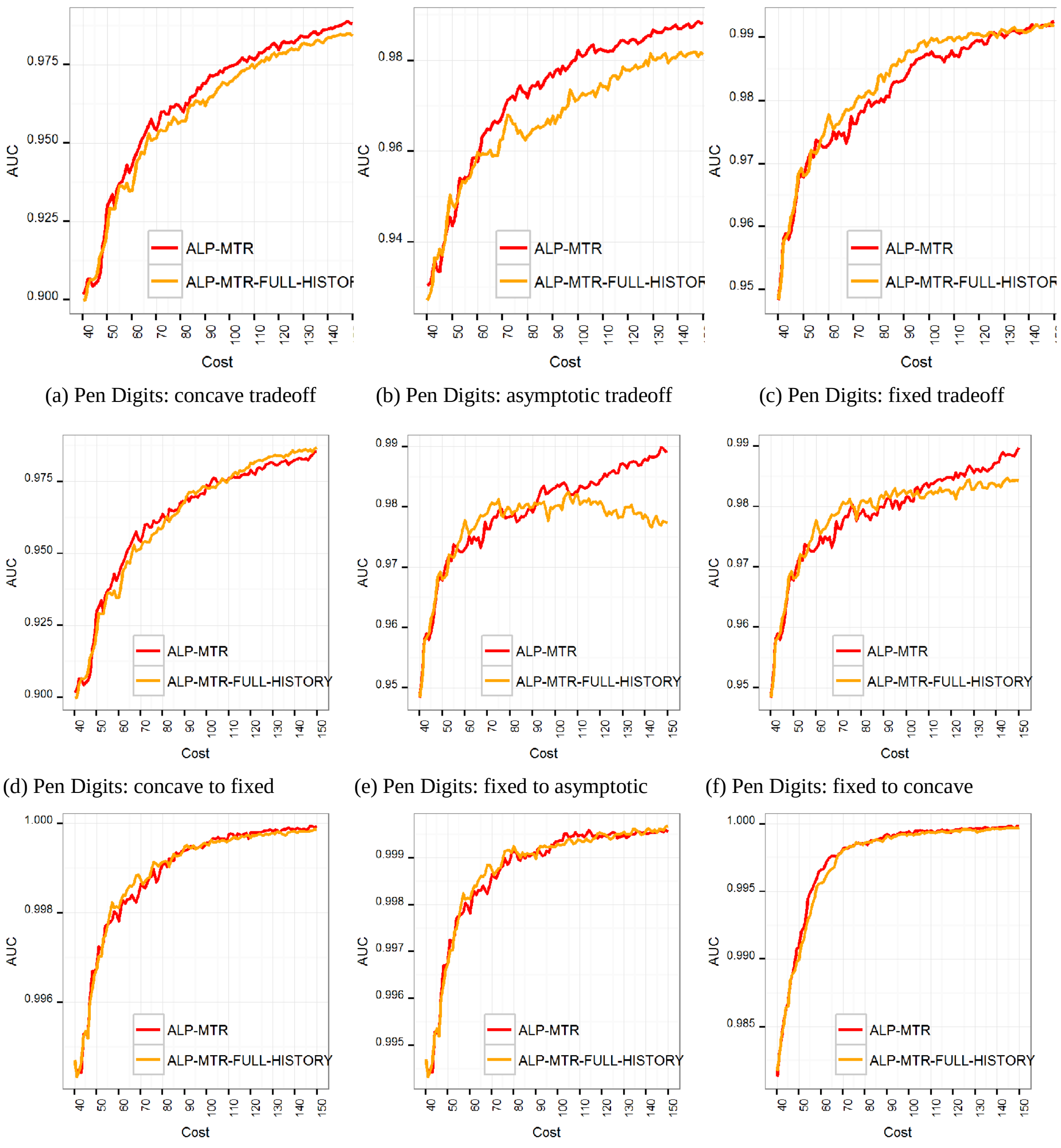
(a) Pen Digits: Asymptotic Tradeoff (b) Pen Digits: Concave Tradeoff (c) Pen Digits: Fixed Tradeoff

Fig. 10 Comparison between ALP-MTR and ALP-MTR-SINGLE-CV. ALP-MTR yields better or comparable performance to that of ALP-MTR-SINGLE-CV

Another element by which ALP-MTR aims to reduce variance in the estimation of the cost-effectiveness of different labeling payments is repeated applications of a cross-validation when estimating the expected change in performance when omitting instances labeled at alternative payment levels. Figure 10 shows a comparison of the standard ALP-MTR with four repetitions of cross validation and ALP-MTR-SINGLE-CV, where only a single cross-validation is used in the estimation. In the interest of space, we present representative results for the Pen Digits dataset and these findings are consistent across datasets. As shown, repeated cross-validations benefits the selection of cost-effective labeling payments, thereby often producing better and otherwise comparable performance as compared to when this procedure is not applied.

Recall that ALP-MTR evaluates alternative labeling payments by estimating the effect of omitting only labels acquired during the most recent h batches. The history parameter aims to enable ALP-MTR to adapt to any changes in the underlying tradeoff between labeling payment and quality. Figure 11 shows a comparison of ALP-MTR with an ALP-MTR variant, ALP-MTR-FULL-HISTORY, which evaluates the cost-effectiveness of alternative payments based on the entire purchase history. Figures 11(a-c) show results for settings in which the tradeoff between labeling payment and quality remains constant, and Figures 11(d-i) show learning curves for settings in which the tradeoff changes. In Figure 11, we show results for the Pen Digits dataset; results with the other datasets produced the same finding. As shown, across different tradeoff settings, ALP-MTR yields either superior or comparable cost-effective acquisitions as compared to ALP-MTR- FULL-HISTORY, suggesting that consideration of recent purchases when assessing alternative payments is beneficial.

Finally, because labels are noisy in our settings, having more acquisitions for a given payment can improve ALP-MTR's estimation of the contribution to induction of labels acquired for the corresponding payment. However, it is possible that due to noise, cost-effective payments should be deemed undesirable. Further, because such a payment is not selected, the estimation is not improved and the advantageous payment will continue to be ignored. This problem may be particularly significant when the tradeoff between labeling payment and quality varies over time. In Appendix D, we consider an ALP-MTR variant, ALP-MTR-PAYMENT-REG, which uses ALP-MTR's selection of payments as before; however, it also acquires labels at a given payment level if it has not been selected in the recent t consecutive batch acquisitions. Our results suggest that initiating acquisitions at different payment levels, as done by ALP-MTR-PAYMENT-REG, can improve performance in some cases, but often yields performance inferior to that achieved by the standard ALP-MTR.



(a) Pen Digits: concave tradeoff

(b) Pen Digits: asymptotic tradeoff

(c) Pen Digits: fixed tradeoff

(d) Pen Digits: concave to fixed

(e) Pen Digits: fixed to asymptotic

(f) Pen Digits: fixed to concave

(g) Pen Digits: asymptotic to concave

(h) Pen Digits: asymptotic to fixed

(i) Pen Digits: concave to asymptotic

Fig. 11 Comparison between ALP-MTR and ALP-MTR-FULL-HISTORY. ALP-MTR yields better or equivalent performance compared to ALP-MTR-FULL-HISTORY.

5.4 Additional results

We performed additional evaluations to assess ALP-MTR's robustness under different settings. Specifically, in Appendix C, we also replicate our main results for ALP-MTR when a different classification algorithm, Support Vector Machine, is used to induce predictive models from the labels acquired by ALP-MTR. Here as well, we find that ALP-MTR offers a superior labeling payment approach compared to UNIFORM. ALP-MTR often yields significantly better models for a given labeling acquisition cost and is otherwise comparable to the UNIFORM policy. In addition, we present in Appendix C an evaluation of ALP-MTR's robustness to different acquisition batch sizes (*b* in

Algorithm 1). We find that altering the batch size does not have a significant effect on performance.

6 Conclusions, limitations, and future work

As machine learning becomes integral to the routine operations of firms and the products and services they provide, the immediacy and accessibility of online labor markets present unprecedented opportunities for on-demand labeling to be brought to bear on machine-learning tasks. Yet for these opportunities to materialize, it is important to devise solutions for the novel challenges these markets present. Given the different tradeoffs that can arise between labeling payment and work quality under different labeling tasks and market conditions, in this paper we first formulate the problem of Adaptive Labeling Payments (ALP), then develop an approach to address it, and finally study extensively the performance of our proposed approach. Specifically, we develop an ALP approach, alp-mtr, designed to determine and continuously adapt the payment offered to crowd workers before they undertake a labeling task, so as to cost-effectively produce a desired predictive performance. Our ALP approach estimates the effect on induction from omitting training instances, previously acquired at different payments, and incorporates elements that benefit this assessment, particularly in the presence of noisy labels and changing prevailing tradeoffs between payment and quality.

We empirically evaluated the performance of alp-mtr relative to that of a robust alternative under a variety of market scenarios, reflecting different labeling tasks and tradeoffs found in prior work between labeling payments and quality, including settings where the tradeoff changes over time. Our results show that alp-mtr yields robust performance across settings and that it offers meaningful and substantial cost savings, with an average savings of 33.4% across settings. We also demonstrate that the design elements of alp-mtr, which are designed to improve its estimations of the expected benefits from alternative payments in the presence of noisy labels and its adaptation to changing market conditions, indeed contribute to its performance. Given our method's consistently robust benefits under different settings, it can be considered a benchmark for evaluating future mechanisms to determine labeling payments.

The practical implications of this research are important for enabling a growing reliance on instance labeling via crowdsourcing labor markets. Our ALP approach yields both reliably robust performance and meaningful cost savings. Because crowdsourcing for labeling tasks is becoming increasingly popular in academic research, academic efforts can similarly benefit from our method's efficiencies.

Our proposed approach offers a wealth of opportunities on which future research can be built, both to improve as well as to extend our work. One interesting direction would be an adaptation of our ALP approach to regression predictive tasks, in addition to the classification task that we consider here. As prescribed by our ALP approach, the effect of target (dependent variable) values acquired at different payments may, in principle, be evaluated by estimating the effect on the regression model's performance in omitting instances previously acquired at the corresponding payment. This approach, along with the averaging of repeated estimations proposed in our approach, can be advantageous for a regression setting as well. Similarly, while in this work we considered an important and therefore also popular measure of performance (namely, Area Under the ROC Curve) to evaluate a classification model's performance, our approach could accommodate other performance measures as well.

Our approach is designed to adapt to changing market conditions. However, inherent to all predictive modeling approaches is the assumption that the environment remains stable for some time so that the learned patterns can be exploited. Different solutions may be required in chaotic settings, where market conditions change significantly and very frequently. Finally, it would be interesting for future research to explore different ways of extending our approach. In particular, it would be beneficial for future work to explore the possibility of removing instances labeled at certain payments if this could improve the model's generalization performance. Similarly, it would be interesting to explore an extension of our approach that studies possible effective stopping criteria for acquisitions at a given time. Such stopping criteria could be based, for example, on the EPI measure outlined in Equation 1 to assess whether acquiring labels at a given payment might undermine induction.

7 References

Abe, N., and Mamitsuka, H. "Query learning strategies using boosting and bagging," In Proceedings of the International Conference on Machine Learning (ICML), pages 1–9. Morgan Kaufmann, 1998.

Alonso, O., Rose, D. E., and Stewart, B. 2008. "Crowdsourcing for Relevance Evaluation," ACM SigIR Forum (42:2), November, pp. 9-15.

Archak, N., Ghose, A., and Ipeirotis, P. 2011. "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," Management Science (57:8), pp. 1485-1509.

Bayus, B. L. 2013. "Crowdsourcing New Product Ideas over Time: An Analysis of the Dell Ideastorm Community," Management Science (59:1), pp. 226-244.

Benson, A., Sojourner, A. J., and Umyarov, A. 2015. "The Value of Employer Reputation in the Absence of Contract Enforcement: A Randomized Experiment," Available at SSRN 2557605.

Brabham, D. C. 2008. "Crowdsourcing as a Model for Problem Solving an Introduction and Cases," Convergence: The International Journal of Research into New Media Technologies (14:1), pp. 75-90.

Dalvi, N., Dasgupta, A., Kumar, R., and Rastogi, V. 2013. "Aggregating Crowdsourced Binary Ratings," in Proceedings of the 22nd International Conference on World Wide Web, New York: ACM, pp. 285-294.

Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. 2010. "Are Your Participants Gaming the System? Screening Mechanical Turk Workers," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 10, New York: ACM, pp. 2399-2402.

Feng, D., Besana, S. and Zajac, R. 2009. "Acquiring High Quality Non-Expert Knowledge from on-Demand Workforce," in Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, Stroudsburg, PA: Association for Computational Linguistics, pp. 51-56.

Ipeirotis P. G. 2010. "The New Demographics of Mechanical Turk," Blog. http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html

Ipeirotis, P. G., Provost, F., Sheng, V. S. and Wang, J. 2013. "Repeated Labeling Using Multiple Noisy Labelers," Data Mining and Knowledge Discovery (28:2), pp. 402-441.

Kazai, G. 2011. “In Search of Quality in Crowdsourcing for Search Engine Evaluation,” in Advances in Information Retrieval, P. Clough et al. (eds.), Berlin Heidelberg: Springer, pp. 165-176.

Kazai, G., Kamps J., and Milic-Frayling N. 2013. “An Analysis of Human Factors and Label Accuracy in Crowdsourcing Relevance Judgments,” Information Retrieval (16:2), April, pp. 138-178.

Kapelner, A., and Chandler, D. 2010. “Preventing Satisficing in Online Surveys: A ‘Kapcha’ to Ensure Higher Quality Data,” in The World’s First Conference on the Future of Distributed Work (CrowdConf2010).

Karger, D. R., Oh, S., and Shah, D. 2011. “Iterative Learning for Reliable Crowdsourcing Systems,” in Proceedings of Advances in Neural Information Processing Systems: 25th Annual Conference on Neural Information Processing, December 12-14.

Karger, D. R., Oh, S., and Shah, D. 2014. “Budget-Optimal Task Allocation for Reliable Crowdsourcing Systems,” Operations Research (62:1), February, pp. 1-24.

Kittur, A., Ed, H., and Chi, B. S. 2008. “Crowdsourcing User Studies with Mechanical Turk,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 5-10, New York: ACM, pp. 453-456.

Kong, D. and Saar-Tsechansky, M. 2014. “Collaborative Information Acquisition for Data-Driven Decisions.” Machine Learning, Volume 95, Issue 1, pages 71-86, 2014.

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. 2013. “The Future of Crowd Work,” In Proceedings of the 2013 Conference on Computer Supported Cooperative Work, NY: ACM, pp. 1301-1318, 4.

Kumar, A., and Lease, M. 2011. “Modeling Annotator Accuracies for Supervised Learning,” in Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM), at the Fourth ACM International Conference on Web Search and Data Mining (WSDM), pp. 19-22.

Lakhani, K. R., Jeppesen, L. B., Lohse, P. A., and Panetta, J. A. 2007. “The Value of Openness in Scientific Problem Solving,” Boston, MA: Harvard Business School (working paper).

Lee, D., Hosanagar, K., and Nair, H., "The Effect of Advertising Content on Consumer Engagement: Evidence from Facebook" (working paper). Available at SSRN 2290802 (2013).

Le, J., Edmonds, A., Hester, V., and Biewald, L. 2010. "Ensuring Quality in Crowdsourced Search Relevance Evaluation: The Effects of Training Question Distribution," in InSIGIR 2010 Workshop on Crowdsourcing for Search Evaluation, July 19-23, pp. 21-26.

Lewis, D., and W. Gale. "A sequential algorithm for training text classifiers," In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3–12. ACM/Springer, 1994.

Lichman, M. 2013. "UCI Machine Learning Repository," Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml

Mason, W., and Suri, S. 2012. "Conducting Behavioral Research on Amazon's Mechanical Turk," Behavior Research Methods (44:1), pp. 1-23.

Mason, W., and Watts, D. J. 2010. "Financial Incentives and the Performance of Crowds," ACM SigKDD Explorations Newsletter (11:2), May, pp. 100-108.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. 2010. "Running Experiments on Amazon Mechanical Turk," Judgment and Decision Making (5:5), pp. 411-419.

Ramirez-Loaiza, M. Sharma, M., Kumar,G., and Bilgic, M. 2016. "Active Learning: an Empirical Study of Common Baselines." Data Mining and Knowledge Discovery (DMKD), pp 1-27, 2016.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. 2010. "Learning From Crowds," The Journal of Machine Learning Research (1:11), March, pp. 1297-1322.

Rodrigues, F., Pereira, F., and Ribeiro, B. 2013. "Learning from Multiple Annotators: Distinguishing Good from Random Labelers," Pattern Recognition Letters (34:12), September, pp. 1428-1436.

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J. and Vukovic, M., 2011. "An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets," ICWSM (11), pp.17-21.

Roy, N., and A. McCallum. “Toward optimal active learning through sampling estimation of error reduction.” 2001. In Proceedings of the International Conference on Machine Learning (ICML), pages 441–448. Morgan Kaufmann, 2001.

Saar-Tsechansky, M. and Provost, F. “Active Sampling for Class Probability Estimation and Ranking.” Machine Learning, 54:2, 153-178, 2004.

Sharma M, Bilgic M. 2016. “Evidence-based uncertainty sampling for active learning. Data Mining and Knowledge Discovery, pp 1–39.

Shaw, Aaron D., John J. Horton, and Daniel L. Chen. 2011. “Designing incentives for inexpert human raters,” in Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work

Sheng, V. S., Provost, F., and Ipeirotis, P. G. 2008. “Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers,” in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 614-622.

Snow, R., O’Connor, B., Jurafsky, D., and Ng. A. Y. 2008. “Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing, October 25, pp. 254-263.

Wang, J., Ipeirotis, P. G., and Provost, F. (2017, forthcoming). “Cost-Effective Quality Assurance in Crowd Labeling,” Information Systems Research.

Wauthier, F. L., and Jordan, M. I. 2011. “Bayesian Bias Mitigation for Crowdsourcing,” in Advances in Neural Information Processing Systems (NIPS), P. Bartlett, F. Pereira, J. Shawe-Taylor, and R. Zemel (eds.), pp. 1800-1808.

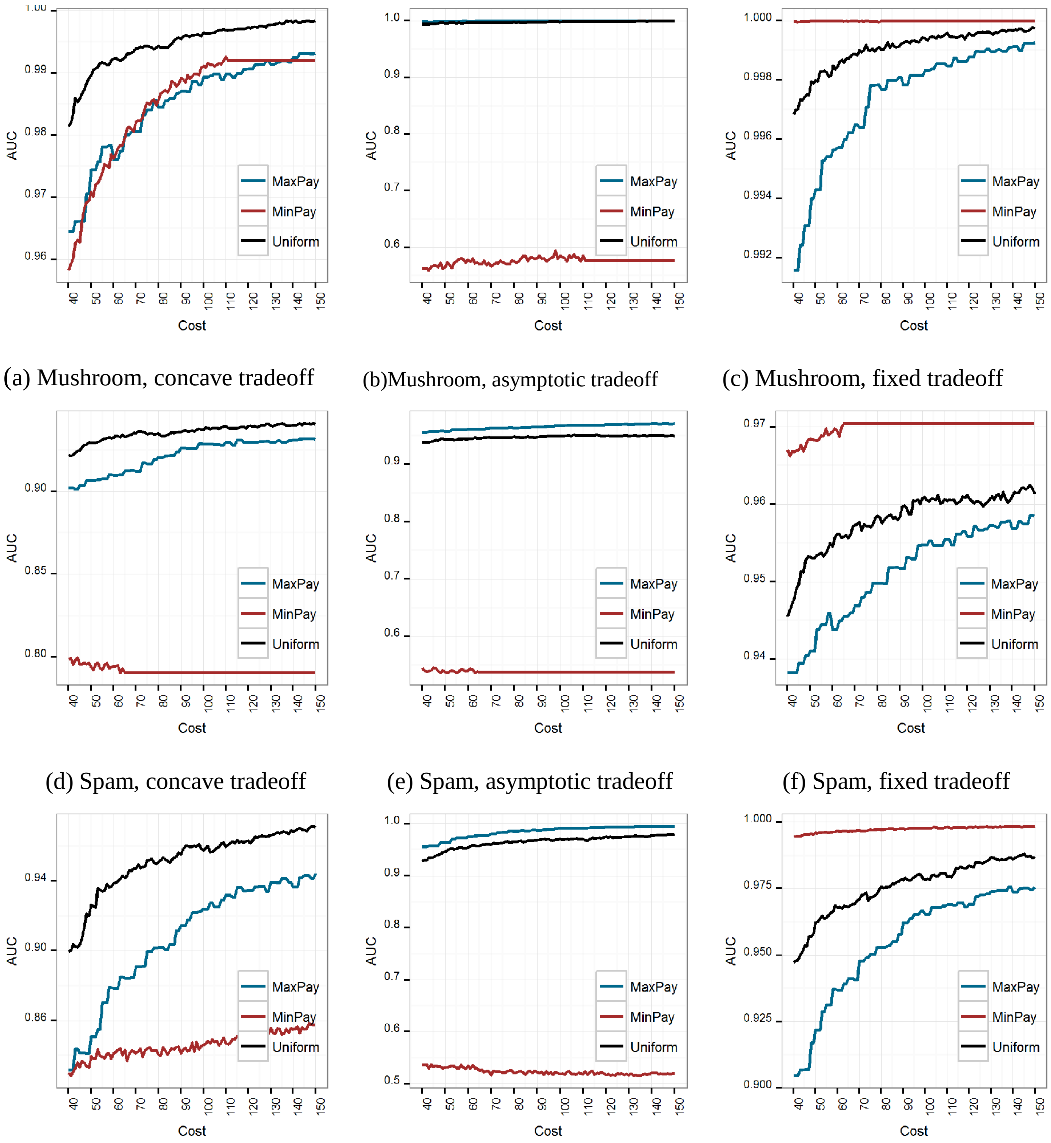
Yang, L., and Carbonell, J. 2012. “Adaptive Proactive Learning with Cost-Reliability Trade-off,” in Encyclopedia of the Sciences of Learning, N. M. Seel (ed.), New York: Springer, pp. 121-127.

Zhou, D., Basu, S., Mao, Y., and Platt, J. C. 2012. “Learning from the Wisdom of Crowds by Minimax Entropy,” in Advances in Neural Information Processing Systems, pp. 2204-2212.

Zhu, D., and Carterette, B. 2010. “An Analysis of Assessor Behavior in Crowdsourced Preference Judgments,” in SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation, July 23, pp. 21-16.

Appendix A: Evaluations of alternative policies

Figures A1 shows the performances of the MaxPay, MinPay, and the Uniform policies for different labeling tasks when the tradeoff between payment and quality remains constant. Figure A2 shows the performances of these policies for the same tasks, but when the tradeoff between payment and quality changes once half of the acquisition budget is exhausted. The Uniform policy yields the most robust performance and does not result in very poor performance. Minpay and maxpay both exhibit very poor performance under some settings.

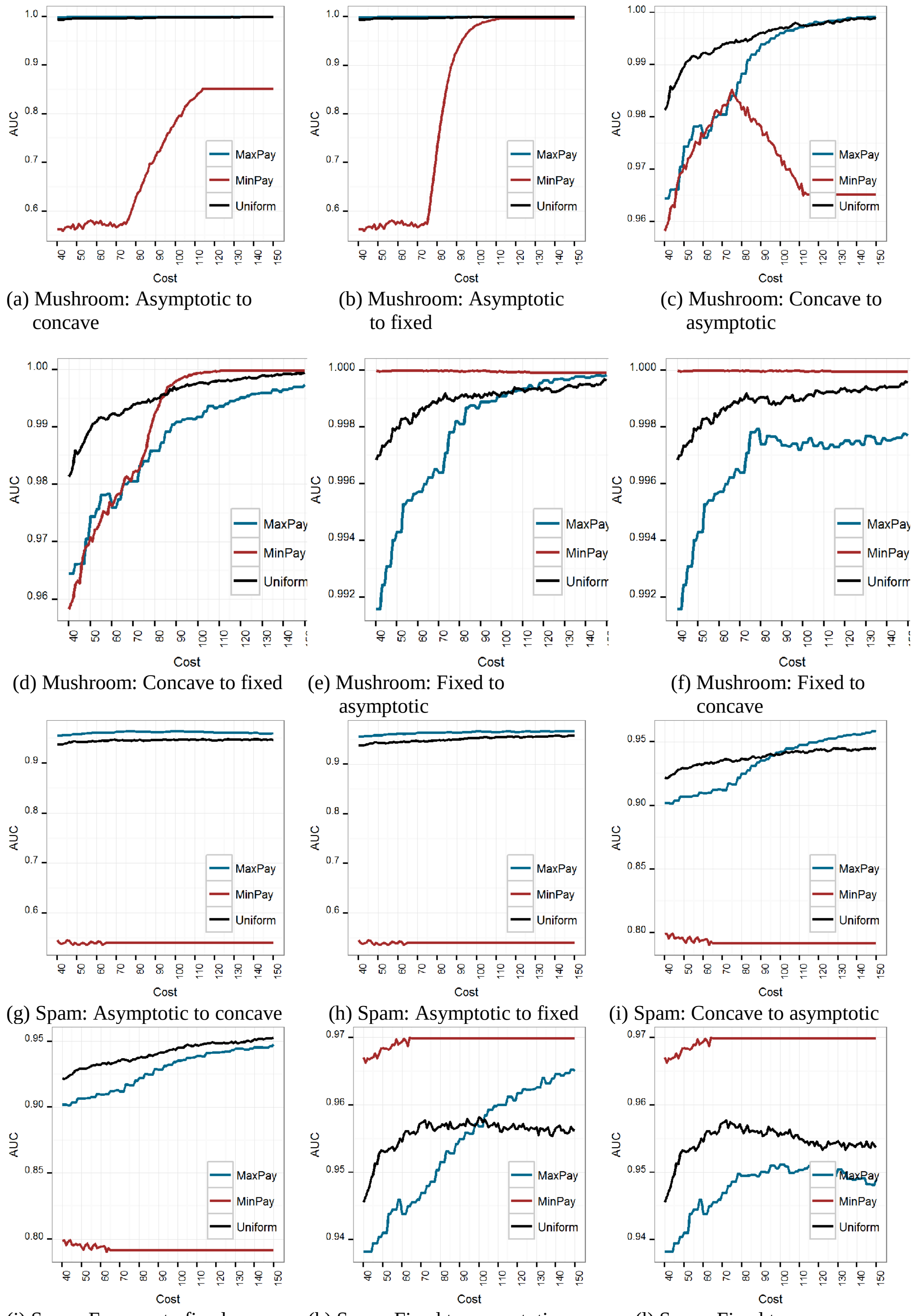


(a) Mushroom, concave tradeoff (b)Mushroom, asymptotic tradeoff (c) Mushroom, fixed tradeoff

(d) Spam, concave tradeoff (e) Spam, asymptotic tradeoff (f) Spam, fixed tradeoff

(g) Pendigits, concave tradeoff (h) Pendigits, asymptotic tradeoff (i) Pendigits, fixed tradeoff

Fig. A1 Performances of MinPay, MaxPay, and Uniform for different labeling tasks and Tradeoffs.



(a) Mushroom: Asymptotic to concave

(b) Mushroom: Asymptotic to fixed

(c) Mushroom: Concave to asymptotic

(d) Mushroom: Concave to fixed

(e) Mushroom: Fixed to asymptotic

(f) Mushroom: Fixed to concave

(g) Spam: Asymptotic to concave

(h) Spam: Asymptotic to fixed

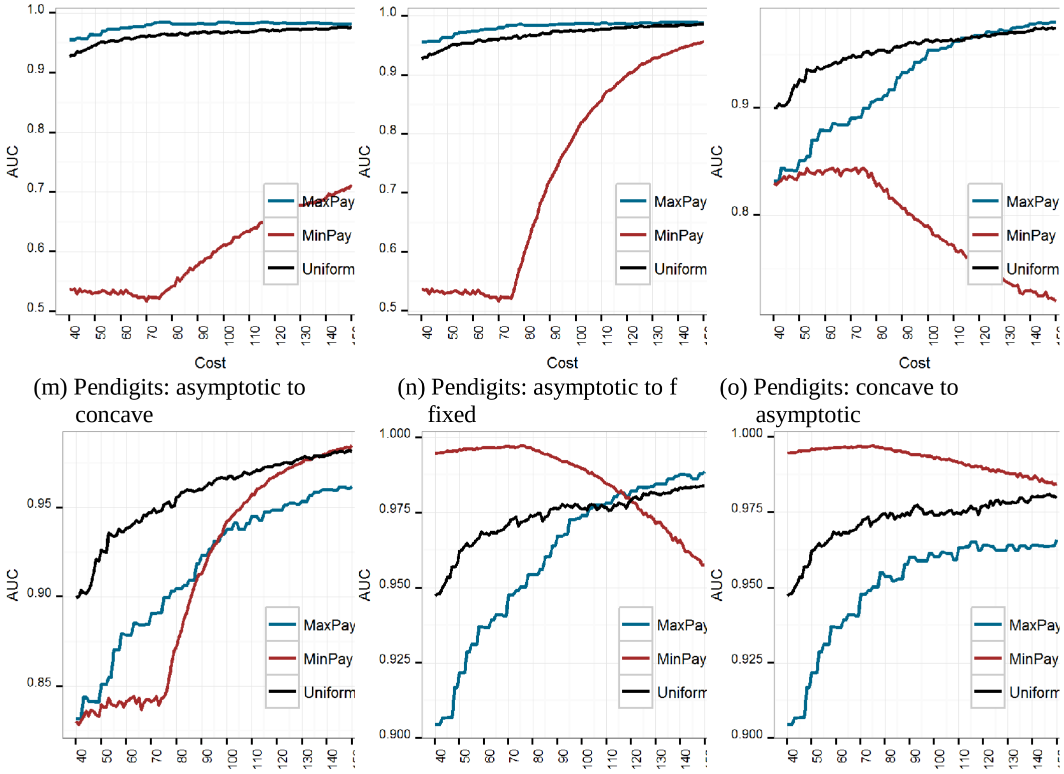
(i) Spam: Concave to asymptotic

(j) Spam: Foncave to fixed

(k) Spam: Fixed to asymptotic

(l) Spam: Fixed to concave

Fig. A2 Performances of MinPay, MaxPay, and Uniform for different labeling tasks and changing tradeoffs



(m) Pendigits: asymptotic to concave

(n) Pendigits: asymptotic to f fixed

(o) Pendigits: concave to asymptotic

(p) Pendigits: concave to fixed

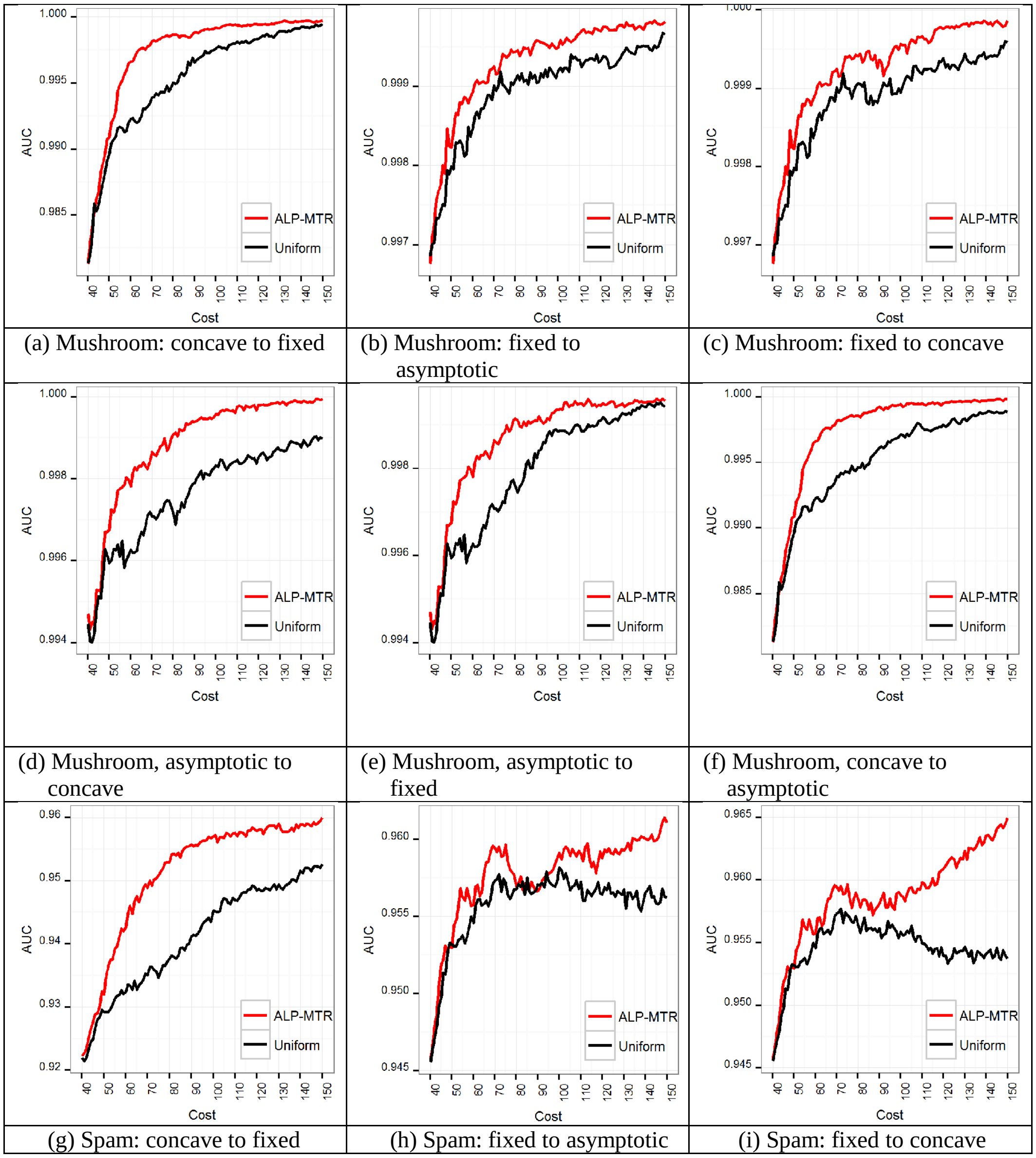
(q) Pendigits: fixed to asymptotic

(r) Pendigits: fixed to concave

Fig. A2 (continued from previous page): Performances of MinPay, MaxPay, and Uniform for different labeling tasks and changing tradeoffs.

Appendix B: Evaluations of when the prevailing tradeoff between payment and quality changes

Figure B1 shows the performances of ALP-MTR and the UNIFORM policy under market settings in which the tradeoff between payment and quality changes from one tradeoff between payment and quality to another. As shown, ALP-MTR‘s performance is often significantly better and is otherwise comparable to that of the UNIFORM policy. Overall, ALP-MTR is preferable to UNIFORM in these settings.



(a) Mushroom: concave to fixed

(b) Mushroom: fixed to asymptotic

(c) Mushroom: fixed to concave

(d) Mushroom, asymptotic to concave

(e) Mushroom, asymptotic to fixed

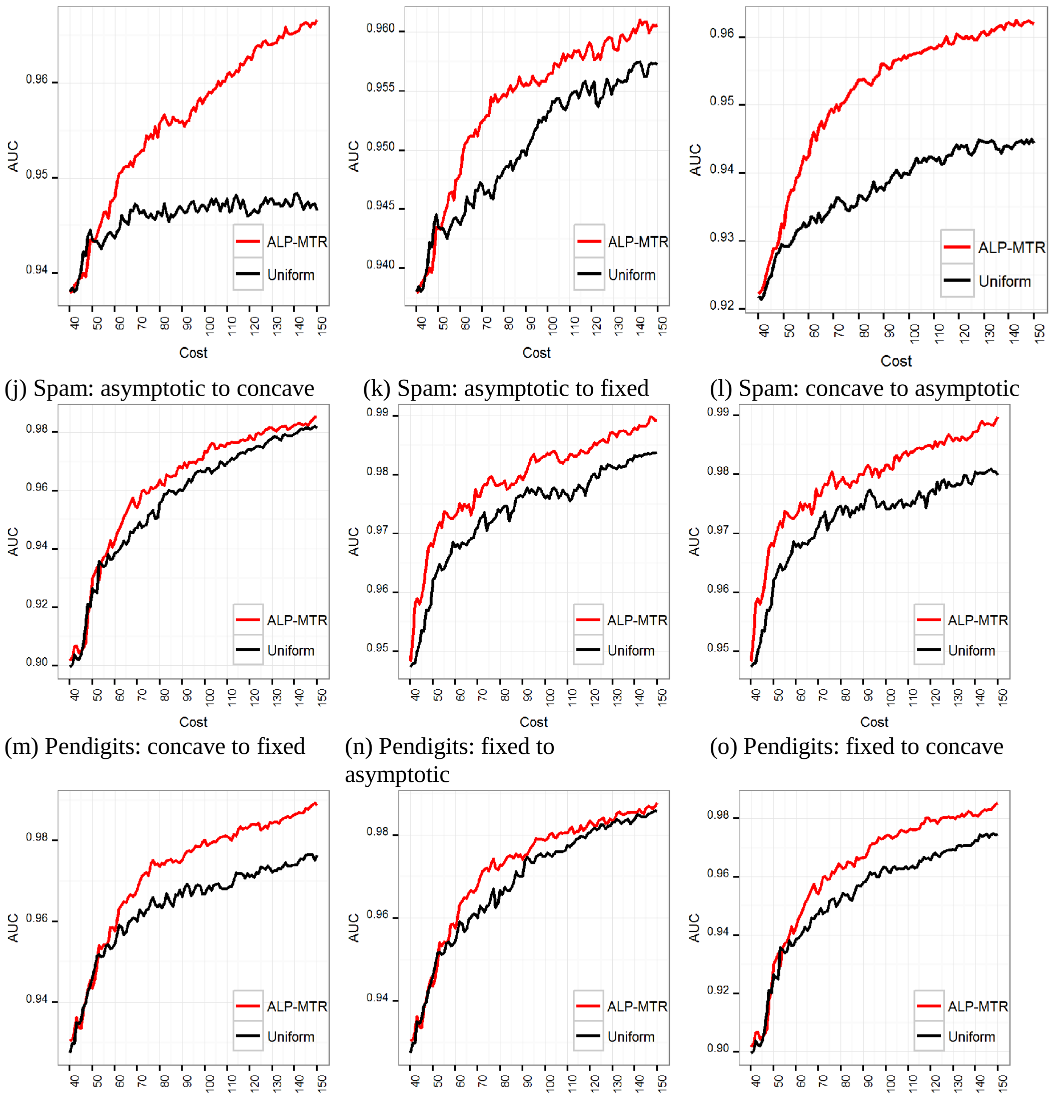
(f) Mushroom, concave to asymptotic

(g) Spam: concave to fixed

(h) Spam: fixed to asymptotic

(i) Spam: fixed to concave

Fig. B1 Comparison of ALP-MTR to UNIFORM under changing tradeoff between payment and quality.



(j) Spam: asymptotic to concave

(k) Spam: asymptotic to fixed

(l) Spam: concave to asymptotic

(m) Pendigits: concave to fixed

(n) Pendigits: fixed to asymptotic

(o) Pendigits: fixed to concave

(p) Pendigits: asymptotic to concave

(q) Pendigits: asymptotic to fixed

(r) Pendigits: concave to asymptotic

Fig. B1 (continued from previous page): Comparison of ALP-MTR to UNIFORM under changing tradeoff between payment and quality.

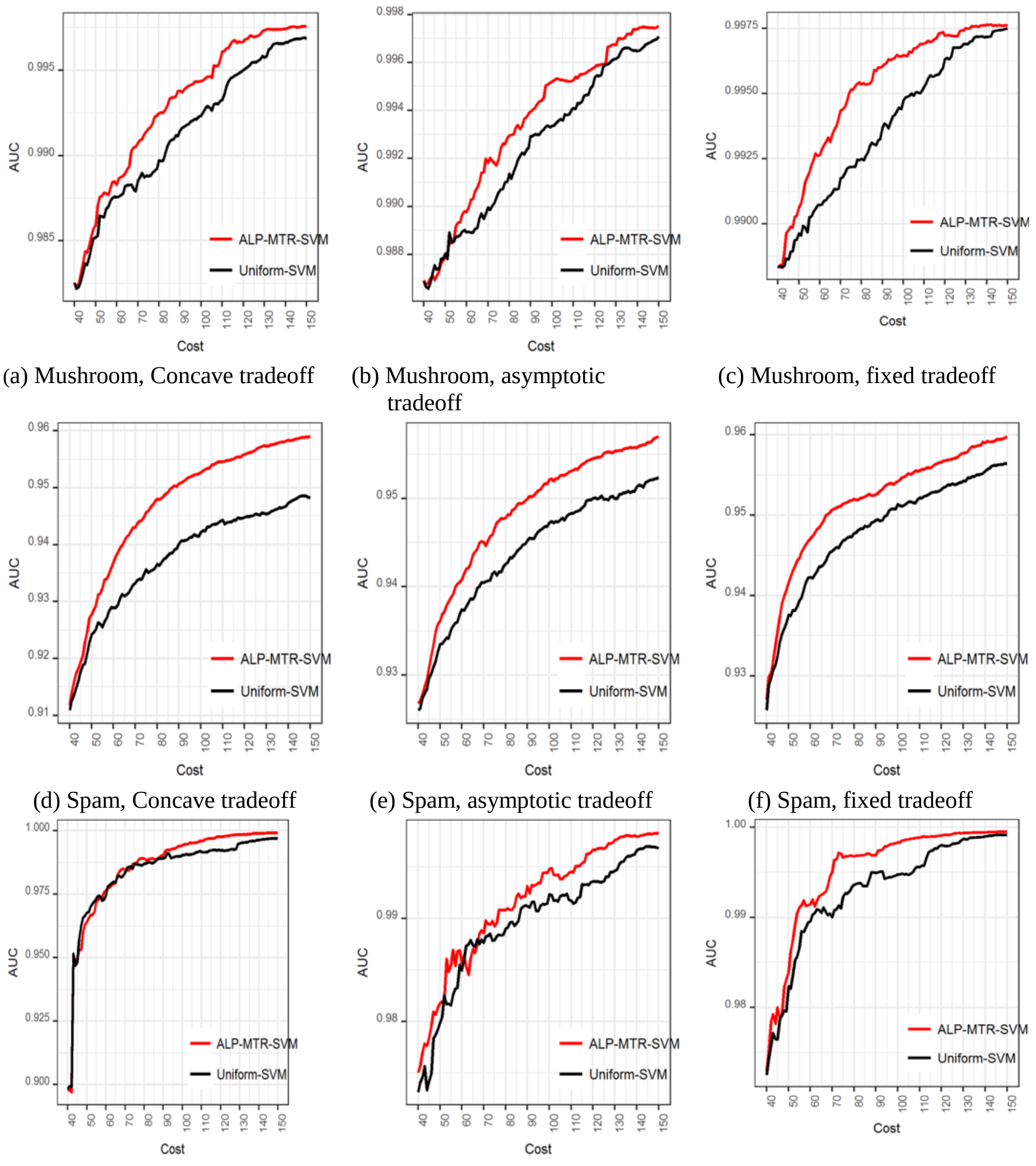
Appendix C: Additional robustness evaluations

We evaluate ALP-MTR using a different popular classification algorithm to derive models from the acquired labels. Specifically, in the experiments below we use Support Vector Machine (SVM) rather than Random Forest as the model inducer. We applied the SVM classifier in R package e1071, using the default settings.

Table C1 Cost savings produced by ALP-MTR relative to the uniform policy when models are induced an SVM inducer.

Dataset	**Tradeoff**	**Cost savings enabled by ALP-MTR**
Mushroom	Asymptotic	11.9%
	Concave	17.1%
	Fixed	15.4%
Spam	Asymptotic	43.8%
	Concave	73.8%
	Fixed	27.1%
Pendigits	Asymptotic	15.0%
	Concave	29.8%
	Fixed	22.1%
Average Savings		**28.4%**

For different data sets and tradeoffs, Table C1 lists the average cost savings for inducing SVM models achieved by ALP-MTR to yield the best SVM model perofrmance exhibited by the UNIFORM policy. As shown, ALP-MTR yields a significant reduction in the costs necessary to yield the best model performance achieved by the UNIFORM payment policy. Specifically, Table 1 shows that, on average, ALP-MTR yield 35.6% savings in labeling costs, across different settings and labeling tasks. Figure C1 shows the AUC performance curves achieved by each approach for different labeling costs. Overall, ALP-MTR is a more advantageous payment policy and it never yields worse models for a given cost than can be achieved by the uniform policy.



(a) Mushroom, Concave tradeoff

(b) Mushroom, asymptotic tradeoff

(c) Mushroom, fixed tradeoff

(d) Spam, Concave tradeoff

(e) Spam, asymptotic tradeoff

(f) Spam, fixed tradeoff

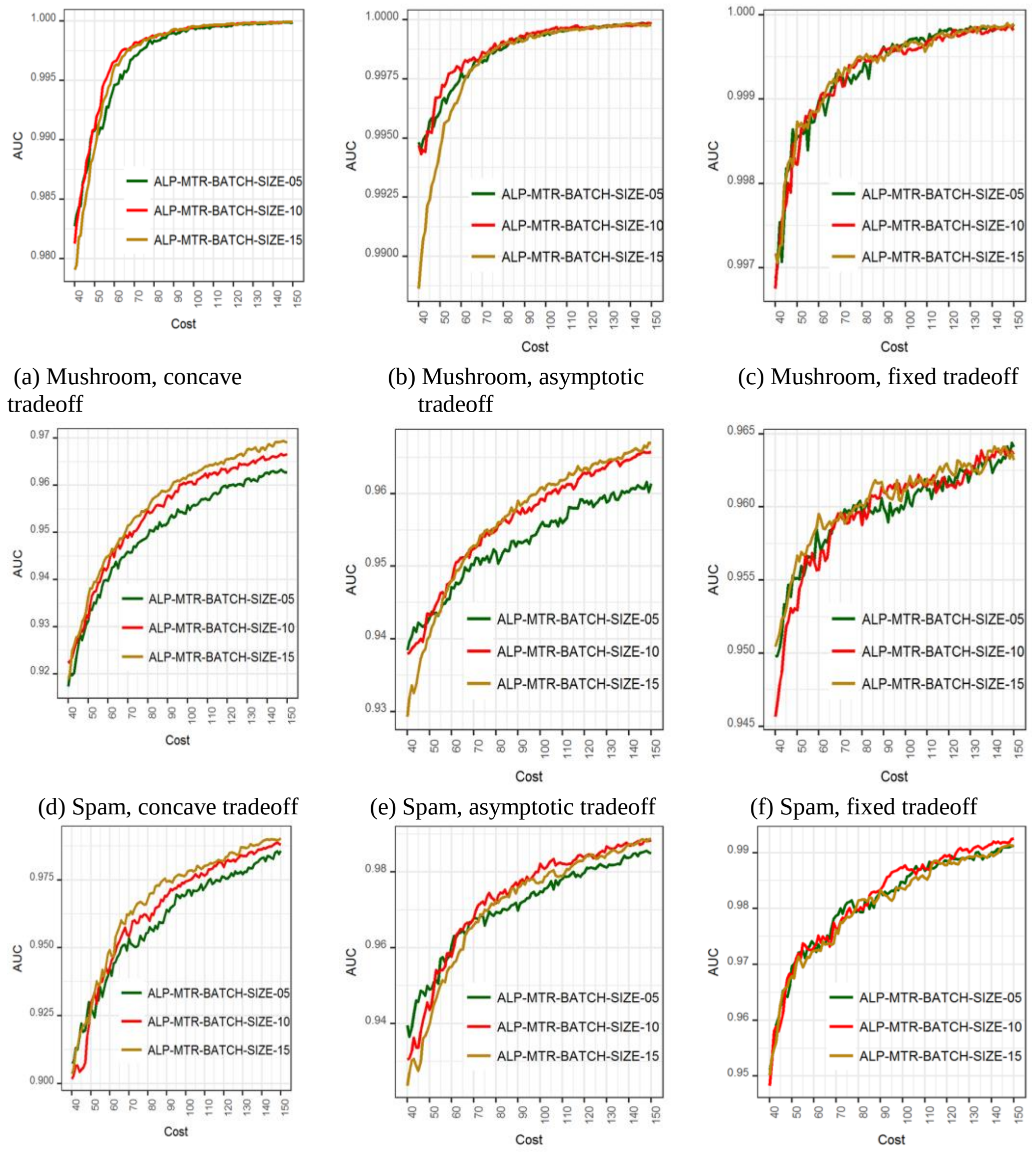
(g) Pendigits, Concave tradeoff

(h) Pendigits, asymptotic tradeoff

(i) Pendigits, fixed tradeoff

Fig. C1 Performance of ALP-MTR compared to the UNIFORM baseline while using the SVM induction algorithm.

We also evaluated the effect of the acquisition batch size on ALP-MTR's performance. In Figure C2 shows the performances of ALP-MTR when at each phase a batch sizes of 5, 10, or 15 instances are acquired. As shown, altering the batch size has no significant effect on performance.



(a) Mushroom, concave tradeoff

(b) Mushroom, asymptotic tradeoff

(c) Mushroom, fixed tradeoff

(d) Spam, concave tradeoff

(e) Spam, asymptotic tradeoff

(f) Spam, fixed tradeoff

(g) Pendigits, concave tradeoff

(h) Pendigits, asymptotic tradeoff

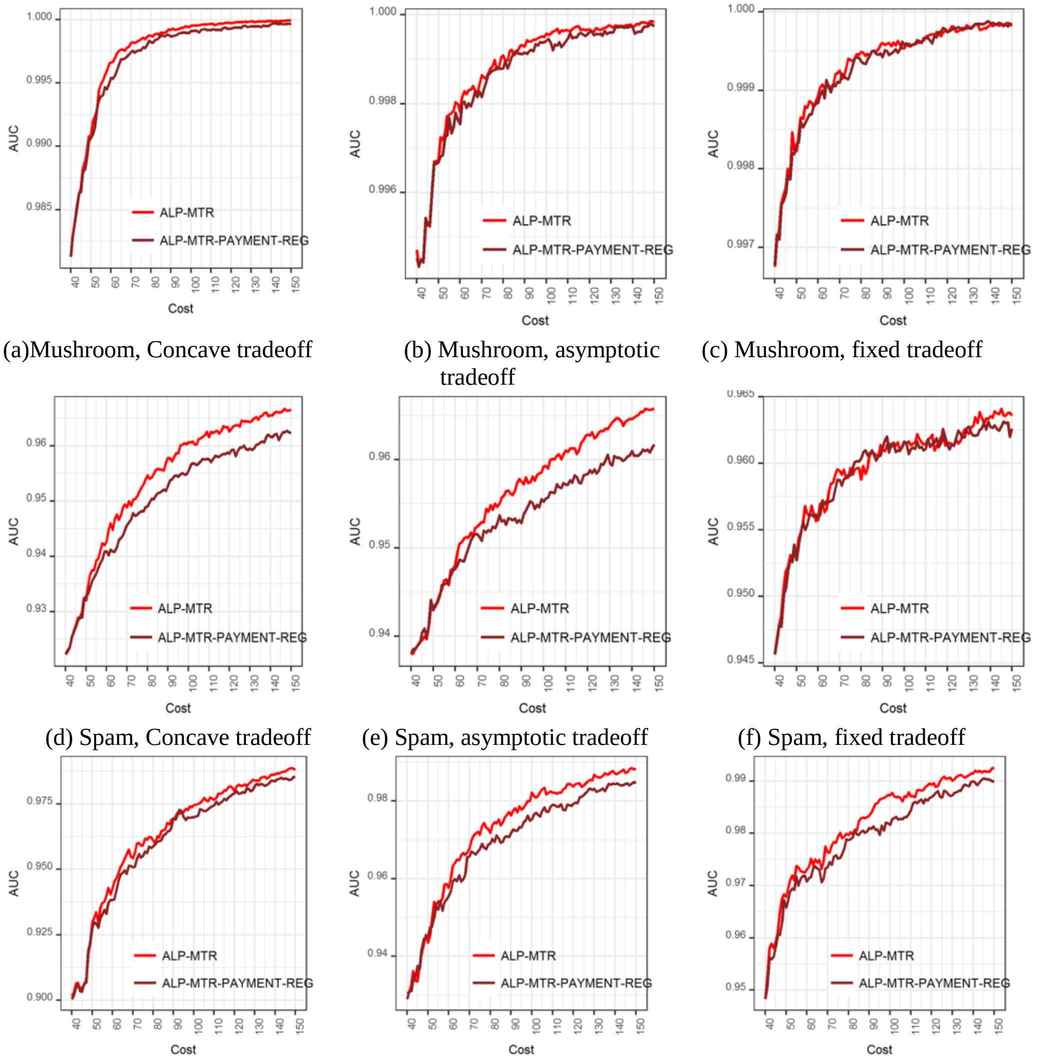
(i) Pendigits, fixed tradeoff

Fig. C2 Performance of ALP-MTR using batch sizes of 5, 10 and 15 instances.

Appendix D: Extending ALP-MTR to include payment regulation

To identify payments that can yield the best model performance at the lowest cost, ALP-MTR uses training instances labeled previously for different payments. Because labels are noisy in our settings, having more acquisitions for a given payment can imporve the estimation of the contribution to induction of labels acquired for the corresponding paymnet to producing a better model. However, it is possible that due to noise, cost-effective payments are deemed undesirable. Further, because such a payment is not selected, the estimation is not improved and the advantageous payment will continue to be ignored. This problem may be particularly significant when the tradeoff between labeling payment and quality varies over time.

In this appendix we extended ALP-MTR to include payment regulation. This ALP-MTR variant, ALP-MTR-PAYMENT-REG, uses ALP-MTR's selection of payments as before; however, it acquires labels at a given payment level if this payment has not been selected in the recent t consecutive batch acquisitions. While ALP-MTR-PAYMENT-REG aims to enhance the estimation of the benefits of different payment levels, it likely also incurs opportunity costs by offering payments on the market that are not cost-effective. We evaluated ALP-MTR-PAYMENT-REG for $t = 10$, $t = 20$ and , $t = 30$, all of which produced similar results. Figure D1 and D2 show the performance of ALP-MTR-PAYMENT-REG for $t = 10$, for different tradeoffs (Figure D1) and when the tradeoff between payment and quality changes (Figure D2) As shown, ALP-MTR-PAYMENT-REG produces similar perormance to that of ALP-MTR, and in most cases it slightly underperforms ALP-MTR. Our results suggest that the standard ALP-MTR approach effectively captures the cost-effectiveness of different payments, and in setting where the tradeoff between payment and quality varies, ALP-MTR effectively adapts to these changes. Thus, enforcing acquisitions at different payment levels, as done by ALP-MTR-PAYMENT-REG, can improve performance in some cases, but often yield inferior performance to that achieved by the standard ALP-MTR.



(a)Mushroom, Concave tradeoff

(b) Mushroom, asymptotic tradeoff

(c) Mushroom, fixed tradeoff

(d) Spam, Concave tradeoff

(e) Spam, asymptotic tradeoff

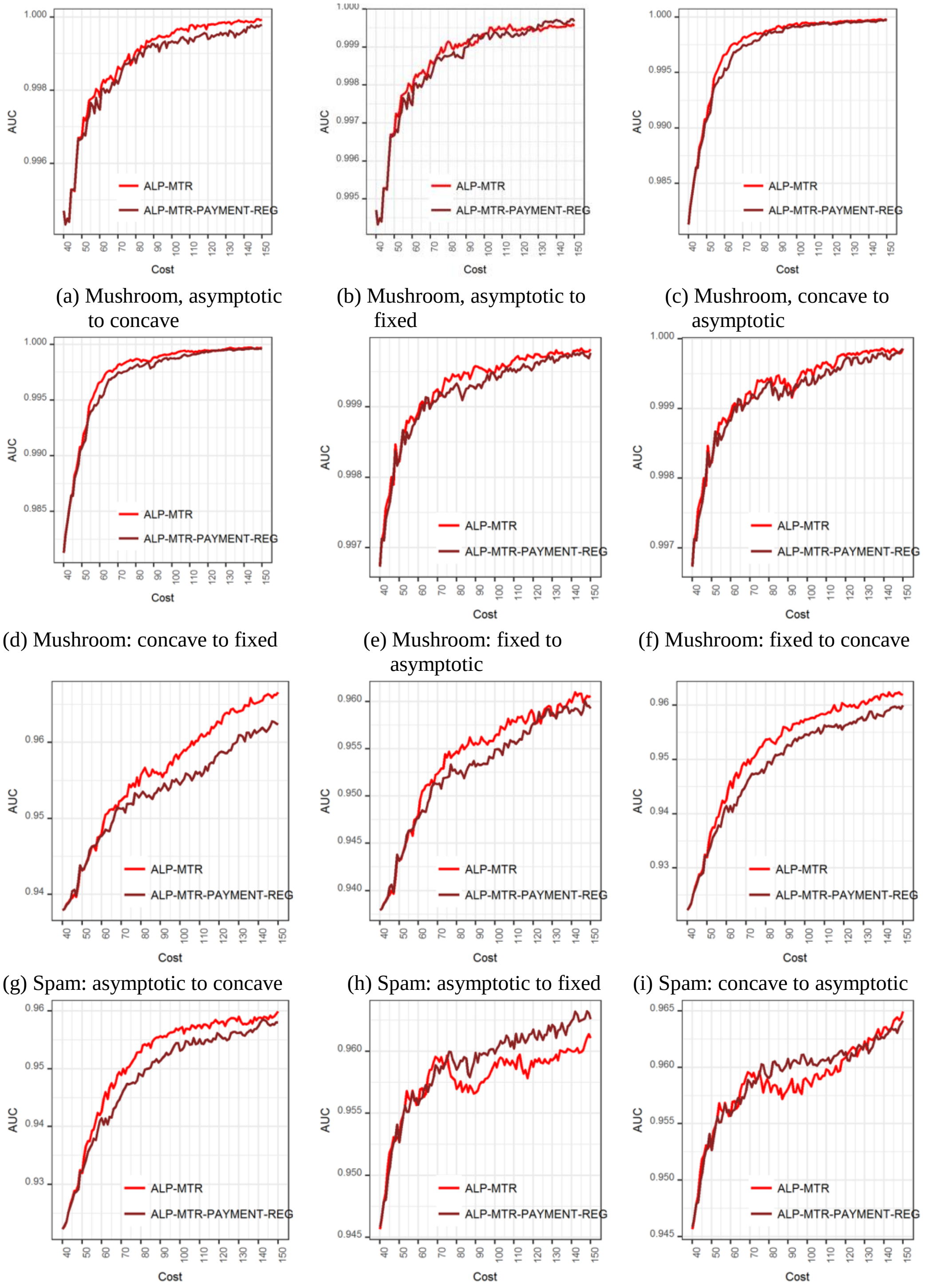
(f) Spam, fixed tradeoff

(g) Pendigits, Concave tradeoff

(h) Pendigits, asymptotic tradeoff

(i) Pendigits, fixed tradeoff

Fig. D Performance of ALP-MTR and ALP-MTR-PAYMENT-REG for different tradeoffs between payment and quality.



(a) Mushroom, asymptotic to concave

(b) Mushroom, asymptotic to fixed

(c) Mushroom, concave to asymptotic

(d) Mushroom: concave to fixed

(e) Mushroom: fixed to asymptotic

(f) Mushroom: fixed to concave

(g) Spam: asymptotic to concave

(h) Spam: asymptotic to fixed

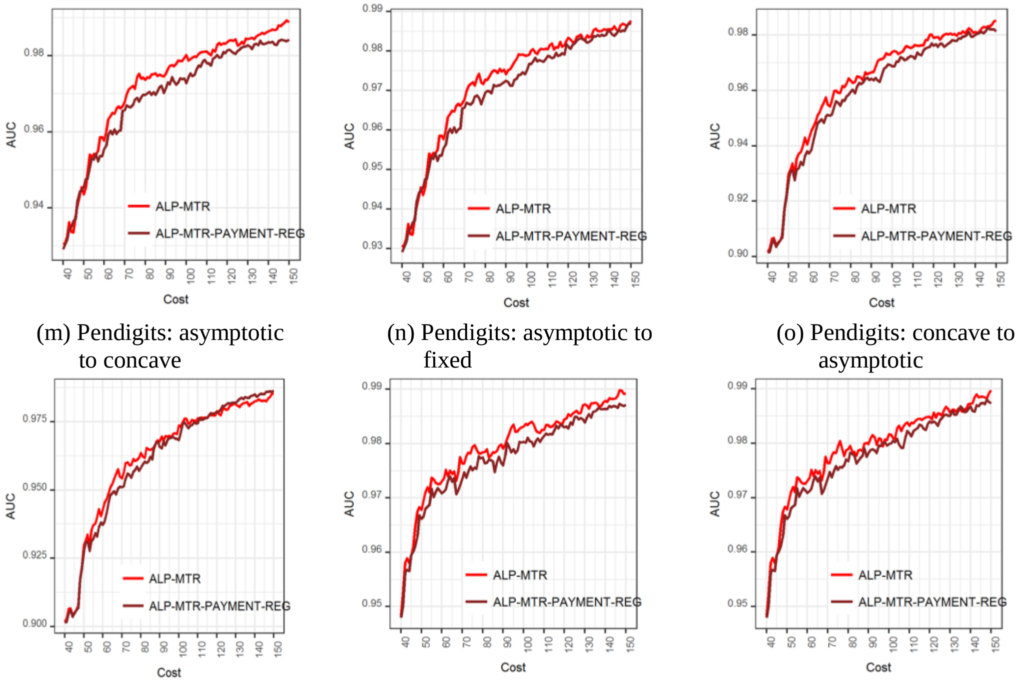
(i) Spam: concave to asymptotic

(j) Spam: concave to fixed

(k) Spam: fixed to asymptotic

(l) Spam: fixed to concave

Fig. D2 Performance of ALP-MTR and ALP-MTR-PAYMENT-REG when the tradeoff between payment and quality changes.



(m) Pendigits: asymptotic to concave

(n) Pendigits: asymptotic to fixed

(o) Pendigits: concave to asymptotic

(p) Pendigits: concave to fixed

(q) Pendigits: fixed to asymptotic

(r) Pendigits: fixed to concave

Fig. D2 (continuing from previous page): Performance of ALP-MTR and ALP-MTR-PAYMENT-REG when the tradeoff between payment and quality changes.

Title page