Capturing Human Factors to Optimize Crowdsourced Label Acquisition through Active Learning

Senjuti Basu Roy
New Jersey Institute of Technology
senjutib@njit.edu

Abstract

The goal of this article is to propose an optimization framework by acknowledging human factors to enable label acquisition through active learning . *In particular, we are interested to investigate tasks, such as, providing (collecting or acquiring) and validating labels, or comparing data using active learning techniques. Our basic approach is to take a set of existing* active learning techniques for a few well known supervised and unsupervised algorithms, but study them in the context of crowdsourcing, especially considering worker-centric optimization (i,e., human factors). *Our innovation lies in designing optimization functions that appropriately capture these two fundamental yet complementary facets, performing systematic investigation to understand the complexity of such optimization problems, and designing efficient solutions with theoretical guarantees.*

1 Introduction

Human workers or crowd can be used as a building block in data-intensive applications, especially for acquiring and validating data. One such fundamental crowdsourcing task is *labeling*, where a human worker is asked to provide one or multiple labels for the underlying observation. A plethora of applications in cyber human system directly benefit from such efforts, where the objective is to design automated classification/prediction algorithms but require *labeled data* for that. Even though human intelligence provides substantial benefit to computation, incorporating humans in the computational loop incurs *additional burden* - it becomes time-consuming, monetarily expensive, or both in such cyber-human systems.

Active learning principles [34, 36] are proposed to optimize system-centric criteria in classification problems, by employing human workers judiciously only for a few tasks. When crowd is involved in data acquisition, additional challenges emerge: (1) contribution from crowd is potentially noisy, (ii) to ensure higher engagement and productivity, one has to understand *worker-centric criteria* [39], such as, *worker skill, motivation*, that are referred to as *human factors* in the literature [2, 48, 31, 8]. In a hybrid human-machine computational environment, an opportunity exists in *laying a scientific foundation for predictive analytics* that combines system-centric optimization derived from active learning [34, 36] principles and worker-centric optimization through human factors modeling.

Imagine that a computational ecologist wants to design a binary classifier [15] that accurately predicts the presence or the absence of species given environmental covariates (such as geographical coordinates, elevation, soil type, average precipitation, etc). In order to learn the classifier, the ecologist needs annotated data (i.e., identify the presence or absence of a species in a given location). Indeed, as shown in Figure 1, there exists

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

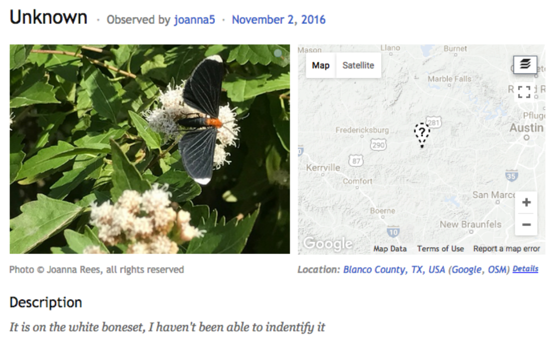

Figure 1: Unidentified Species at iNaturalist Website

unidentified species at large-scale citizen science based platforms like iNaturalist that are to be crowdsourced to be labeled by their registered workers. While doing that, we however need to find the most "suitable" worker taking into account worker-centric criteria, such as, her geographic distance from the observation location, her expertise and/or motivation to identify insects, by considering her past activities/tasks. The goal is to focus on applications, where the workers (e.g., registered citizen scientists of iNaturalist) are involved in completing tasks that are repetitive in characteristics and do not evolve much over the time. We propose to develop an *optimized human-machine intelligence framework for such cyber-human systems for single and multi-label classification problems [27, 55] through active learning*.

Our effort will be investigate and adapt existing *active learning techniques for a few well known supervised algorithms for single-label and multi-label classification in the context of crowdsourcing, especially considering worker-centric optimization through human factors modeling*. The general idea behind active learning is to select the instances that are likely to be most informative. Then the selected instances are annotated by human, and the computation loop is repeated. Our innovations lie in appropriately capturing these two fundamental yet complementary facets in a single optimization function, performing systematic investigation to these problems and designing innovative solutions.

Example 1: Motivating Example: The iNaturalist platform contains photo vouchered independently verified species occurrence records by the citizen scientists across the world and is one of the fastest growing source of contributions to the global biodiversity information facility. We focus on such platforms, where tasks are repetitive in nature, do not change over the time, and one can retain all the past tasks undertaken by the registered workers (citizen scientists). A computational ecologist makes use of such platforms to develop species distribution models using environmental covariates. Thus, she aims to design a crowdsourcing effort to judiciously obtain single and/or multiple label(s) to annotate some of the unidentified images (refer to Figure 1). A single label acquisition is about identifying the insect (which will then be augmented with the location information), whereas, multiple labels will require identifying the Kingdom, Phylum, Class, Order, Sub-Order, Family, Sub-family of the insect in the image.

For these examples, label acquisition involves human workers. On the other hand, domain expert may not have many human workers at her disposal who are qualified for the task - even if there are workers, they may not be *motivated* to undertake these tasks. Furthermore, workers may even have constraints (e.g., only likes to watch birds, can not travel more than 25 miles from her location). Therefore, which task(s) are to be selected and assigned to which worker(s) remain to be a problem.

Desirable properties: To generalize, the above scenarios call out for the following desirable characteristics: (1) An "activized" label acquisition is desirable - i.e., acquire more data only when it optimizes the underlying computational task considering system-centric criteria. (2) At the same time, select workers and assign tasks to enable *worker-centric optimization*. (3) We argue the necessity of studying these two facets together as a single *optimization problem*, as a staged solution (first select sub-tasks based on active learning, then assign those to the workers to enable worker-centric optimization) may be inadequate, as tasks selected by active learning techniques may end up having a very low worker-centric optimization, resulting in poor outcome overall.

High level approach: (1) We propose *worker-centric optimization by characterizing human factors* in the crowdsourcing context by adapting well-known theories of behavioral psychology [48, 31, 8]. (2) We propose system-centric optimization by adapting a set of well-known *active learning techniques* for supervised algorithms [34, 36, 51, 52, 10, 21, 17, 23, 55, 24, 35, 61] and augment them by combining worker-centric optimization through human factors modeling. (3) We propose systematic investigations on how the two *aforementioned optimization problems could be combined* and propose effective solutions.

Additionally, we will design both retrospective as well as prospective studies. We will perform these evaluations using publicly available citizen science datasets.

Novelties: To the best of our knowledge, no existing work has studied what we propose. The closest related works for active learning through crowdsourcing are for single label acquisition [64, 62, 41, 16, 18]. Worker-centric optimization is not considered there. Active learning research for multi-labels remains in a rather nascent state [55, 24, 35, 61]. These aforementioned studies do not investigate the problem in the context of crowdsourcing. Therefore, the necessity of worker-centric optimization does not even arise there. Our designed prototypes on iNaturslist platform will bear a long-lasting effect to understand global bio-diversity.

1.1 Research Objectives

Our long term vision is to optimize knowledge discovery processes for cyber-human systems. We are interested in designing effective solutions and support both worker and system criteria through active learning and human factors modeling. The research proposed here puts us on track to achieving this vision by addressing a first series of concrete challenges on a very important application domain.

(1) Optimized *single-label acquisition* involving crowd: In this research aim, we strive to propose optimization guided hybrid human-machine algorithms considering active learning for single label acquisition. Active learning is popular in *single label supervised algorithms*, where the key idea is that a machine learning algorithm can achieve greater accuracy with fewer training labels, if it is allowed to choose the data from which it learns [34, 36, 51, 52], where each data has a single label.

We will study and adapt a set of well-known active learning techniques, such as, *uncertainty sampling [34], query-by-committee [52, 42], or expected-error reduction[47]* that are popularly used in well-known classification algorithm, such as, *Naive Bayes' Classifier, Support Vector Machine [58, 50], Decision Trees [37], or ensemble classification algorithms [32]*. Similarly, we will characterize *human factors* of the workers [48, 31, 8], such as *skill, motivation* and then design principled optimization functions that combines task-centric and worker-centric optimization. These complex optimization functions will guide the selection of the right training sample (i.e, task) for further labeling and request the appropriate workers to undertake that task. Using Example 1, this is akin to selecting the most appropriate observation site and select the most appropriate workers to observe the presence or absence of the species there. When multiple workers with varying level of expertise are involved to undertake the same labeling task, we will study how to aggregate their annotations to infer the truth considering *weighted majority voting or iterative approach* [20, 25]. We will formalize *stopping conditions* - i.e., *when to terminate this labeling process by exploiting the confidence [63] of the classification tasks, available budget, or availability of the human workers.* We will investigate effective scalable algorithms to solve these problems by exploiting discrete optimization techniques.

(2) Optimized *multi-labels acquisition* involving crowd : In this aim, we will investigate how to enable active learning principles for *multiple-labels classification tasks* involving crowd (Recall 1). Multi-labels classification is different from multi-class classification, where only a single label needs to be predicted per data point for the latter, albeit there are more than two possible labels. Unlike its single-label counterpart, multi-labels classification using active learning is far less studied, except for a few recent works [55, 24, 35, 61]. In fact, to acquire multiple labels, we are unaware of any related work that attempts to design active learning like techniques involving crowd.

Akin to previous aim, we will adapt a few known active learning algorithms for multi-labels classifications using Support Vector Machine (SVM), Naive Bayes, or Ensemble classifiers [55, 24, 35, 61]. Using this, our objective is to select tasks that will be maximally informative for the classifier. Alternatively, task selection can be guided by a version space analysis such that it will give rise to maximum reduction in the version space of the classifier [57]. We will then augment them with *worker-centric optimization through human factors modeling*, such as worker skill or motivation and design a combined optimization function. This function will dictate which task is to be selected for which worker. Using Example 1, this is akin to selecting the most appropriate unidentified image of the species and select the most appropriate workers to label it. Since a task could be labeled by multiple workers, we will study how to aggregate multiple responses and infer the correct labels (truth inference problem) of a task. We will design an iterative algorithm to effectively infer each task's correct labels. We will also explore the use of correlations among different labels to improve the inference quality. Finally, we will investigate the stopping condition of multi-labels acquisition tasks based on various *convergence criteria*.

We first introduce and characterize different variables (Section 2) pertinent to workers and tasks to describe human factors, then propose worker-centric optimization (Section 3). Both of these are pivotal to investigate crowdsourced single and multi-label tasks through active learning (Sections 3.3 and 4).

2 Data Model

We introduce different variables and characterize human factors [48, 40, 26, 31, 8, 54, 11]. A crowdsourcing platform typically comprises of workers and tasks that serve as the foundation of the framework we propose. We also note that not all the variables are pertinent to every application domain (for example, citizen science applications are usually voluntary contributions). Our effort is to propose a generalization nevertheless.

Domains/types: A set $D = \{d_1, d_2, \ldots, d_l\}$ of given domains is used to describe the different types of tasks in an application. Using Example 1, a particular species may construe a domain.

Workers: A set of m human workers $\mathcal{U} = \{u_1, u_2, \ldots, u_m\}$ are available in a crowdsourcing platform.

Tasks and sub-tasks: A task $\mathcal{T}$ is a hybrid human-machine computational task (classification for example), with a quality condition $Q^{\mathcal{T}}$ and an overall monetary budget $B^{\mathcal{T}}$ that decide its termination. Using Example 1, $\mathcal{T}$ is a classification task which terminates, when $Q^{\mathcal{T}} = 80\%$ accuracy is achieved, or $B^{\mathcal{T}} = \$100$ is exhausted.

Without loss of generality, $\mathcal{T}$ comprises of a set of n subtasks, i.e., $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$. These sub-tasks are of interests to us, as workers will be involved to undertake these sub-tasks. Each sub-task can either be performed by human workers or computed (inferred) by machine algorithms. We consider *pool based active learning*, where a finite pool of sub-tasks exists and given.

Sub-tasks: For single label, a sub-task is an unlabeled instance of the data that requires labeling. Considering Example 1, this is analogous to confirming the presence or absence of a species in a particular geographic location. For multi-label scenario, a sub-task requires multiple labels to be obtained. Using Example 1, this is analogous to obtaining Kingdom, Phylum, Class, Order, etc of the insect.

Worker Response: We assume that a worker u's *response to a particular sub-task* t *may be erroneous, which is used by the machine algorithm in one or more rounds of interactions*. Our framework may ask multiple workers to undertake the same task to reduce the error probability, and may decide which questions to ask in the next round to whom based on the answers obtained in the previous round.

Human Factors: These are the variables that characterize the behavior of the workers in a crowdsourcing platform [48, 40, 26, 31, 8, 54, 11].

Skill (Expertise/Accuracy): Worker's skill in a domain is her expertise/accuracy. Skill of a worker in a domain d is quantified in a continuous $[0, 1]$ scale (to allow a probabilistic interpretation). A worker u may have skills in one or more domains (e.g., different species observation accuracy).

Wage: A worker u may have a fixed wage w_u, or may have to accept the wage a particular task offers. u's may have different wage for different types of tasks.

Motivation: Motivation aims at capturing the worker's willingness to perform a task. A related work [31] proposes a theoretical foundation in motivation theory in crowdsourcing platform and characterizes them in two different ways:

(a) Intrinsic motivation: Intrinsic motivation exists if an individual works for fulfillment generated by the activity (e.g. working just for fun). Furthermore, related works [31, 8, 54] have identified that intrinsic motivation emerges in the following ways: (1) skill variety (refers to the extent to which a worker can utilize multiple skills), (2) task identity (the degree to which an individual produces a whole, identifiable unit of work, versus completion of a small unit which is not an identifiable final product), (3) task significance (the degree to which the task has an influence over others), (4) autonomy (the degree to which an individual holding a job is able to schedule his or her activities), (5) feedback (the extent to which precise information about the effectiveness of performance is conveyed).

(b) Extrinsic motivation: Extrinsic motivation is an instrument for achieving a certain desired outcome (e.g. making money).

The challenge however is, either the values of these factors have to be explicitly given or they have to be estimated. Related works, including our own, have proposed solutions to estimate skill [46, 30] by analyzing historical data. Nevertheless, we are not aware of any effort that models motivational factors or design optimization involving them.

Worker specific constraints: Additionally, a worker may specify certain constraints (e.g., can not work more than 6 hours, or travel farther than 10 miles from her current location).

Characterizing sub-tasks considering human factors: It is easy to notice that the motivational factors described above are actually related to tasks (i.e, sub-tasks).

Formally, we describe that a set A of attributes or meta-data is available to characterize each sub-task t. They are its required skill-domain[1] s^t , cost/wage w^t, duration $time^t$, location $location^t$, significance sig^t, identity $iden^t$, autonomy $auto^t$, task feedback fb^t. Each t, if performed correctly, contributes by a quantity q^t to $Q^{\mathcal{T}}$. These contributions are purely dictated by the active learning principles, such as how much it reduces the uncertainty.

3 Worker-Centric Optimization through Human Factors Modeling

Recall Section 1 and note that worker-centric optimization is a common theme across single and multi-labels tasks, which we first examine here.

Objectives: Our objective here is to explore mathematical models for worker-centric optimization in crowdsourcing platforms. Specifically, given an available pool of tasks and workers where workers perform repetitive tasks, we first obtain human factors of the workers by analyzing their past tasks and then study the problem of task assignment to enable worker-centric optimization. A recent work performs an ethnomethodological study at *Turker Nation*[2] and argues [39] that it is critical to enable worker-centric optimization. Our effort here is to make a formal step towards that goal, independent of any specific system-centric optimization (i.e., the active learning principles). Therefore, such a study has a much broader applicability that goes beyond active learning. Of course, our framework will ultimately combine both system and worker-centric criteria.

Challenges: While the significance of human factors is well-acknowledged in crowdsourcing, the challenge is to be able to estimate them effectively and propose appropriate models that could capture them during task assignment. Added to the fact is the dependence of the underlying crowdsourcing domain, which makes some of these factors more important than the rest (e.g., unlike AMT, there is no monetary pay-offs in citizen science activity, but skill variety is acknowledged to be critical to avoid monotony).

3.1 Proposed Directions

First, we propose how to model and estimate human factors [48, 40, 26, 31, 8, 54, 11] that are pertinent to capture motivation using the variables that are described in Section 2. Then, we describe mathematical models that

[1]for simplicity, we assume that each sub-task requires one skill, whereas, in reality, multiple skills may be needed for a sub-task. The latter assumption is trivially extensible by our framework.

[2]http://www.turkernation.com/

leverages these estimated human factors to explicitly assign tasks to workers.

Estimating human factors: We leverage the past task completion history of the workers as well as the new tasks to compute a Boolean task completion matrix T, where the rows are the workers and the columns are the (sub)-tasks. If a worker u has completed a (sub)-task t successfully in the past, the corresponding entry gets a 1, it gets a 0 otherwise. We assume that the factors that capture intrinsic motivation, i.e., skill variety, task identity, task significance, autonomy, feedback are independent yet latent variables. The second matrix we consider is the task factor matrix $\mathcal{F}$, where the rows are the tasks and the columns are the motivation related latent variables. The final matrix is the user factor matrix $\mathcal{U}$ where rows are the factors and columns are the users. This matrix could be fully latent or observed. In case it is latent, we minimize the error function, as defined below:

$$\sum_{i,j}(t_{ij} - \mathcal{U}_i F_j)^2 + \lambda(\|\mathcal{U}\|^2 + \|\mathcal{F}\|^2) \tag{1}$$

Here, λ is the regularization parameter. The goal is to find $\mathcal{U}$ and $\mathcal{F}$ such that it minimizes the error. For any new worker and new task, the predicted task completion score is calculated by multiplying U_i with F_j. Here, the important thing is to notice that the optimization function only minimizes the error for which ratings are present. We apply the alternating least square approach [56] to solve this problem. This is an iterative approach, where at each iteration, we fix the tasks' latent factor matrix $\mathcal{F}$ in order to solve for $\mathcal{U}$ and vice versa. We have designed a similar solution for predicting tasks to workers considering implicit workers' feedback[45].

Worker-Centric Task Assignment: The solution above only estimates the intrinsic motivational factors, but does not describe how to aggregate them together or combine with extrinsic motivation to perform worker-centric task assignment.

Psychologists Hackman and Oldham [19] have combined factors associated to intrinsic motivations defined *motivating potential score (MPS)* :

$$MPS = \frac{\text{skill-variety} + \text{task-identity} + \text{task-significance}}{3} * \text{ autonomy } * \text{ feedback} \tag{2}$$

Considering this aforementioned formulation, we study the worker-centric task assignment as a global optimization problem to maximize the *aggregated intrinsic and extrinsic motivation*. For a given set of tasks S^{t_u}, $V(S^{t_u})$ represents the overall motivation for worker u, by combining her extrinsic motivation (EXTM) (recall Section 2 that EXTM could be modeled using wage w^t) and intrinsic motivation, i.e, *motivating potential score (MPS)*(refer to Equation 2) [19]. In our initial effort, we combine them linearly, as that allows us to design efficient algorithms. Assigning a set of tasks per worker is reasonable as well as desirable from worker's perspective, because workers in a typical crowdsourcing platform intend to undertake multiple tasks as opposed to a single task. Workers may also have constraints, such as, not spend more than X^u hours, or the aggregated wage must at least be b^u dollars.

Technically, we want to assign tasks to the workers *to maximize the aggregated motivation, such that the assignment satisfies each worker-specific constraints*. One such optimization function is described in Equation 3 (Recall Section 2 where $time^t$ and w^t are the duration and wage of sub-task t, respectively).

$$\text{Maximize} \sum_{u \in \mathcal{U}} [V(S^{t_u}) = EXTM(S^{t_u}) + MPS(S^{t_u})] \tag{3}$$

$$V(S^{t_u}) = \begin{cases} \\ \text{if } \sum_{t \in S^{t_u}} time^t \leq X^u \text{ and } \sum_{t \in S^{t_u}} w^t \geq b^u \\ 0 \quad \text{otherwise} \end{cases}$$

As a simple example, given two tasks i and j, we can add the individual significance $sig^i + sig^j$, identity $iden^i + iden^j$, autonomy $auto^i + auto^j$, or feedback $fb^i + fb^j$. Similarly, the wage of two tasks could also be added and normalized to compute EXTM. Alternative problem formulation is explored below.

3.2 Open Problems

Solving the optimization problem: How to design an effective solution to maximize worker motivation based on the aforementioned objective function formulation is challenging. We observe that the proposed optimization problem is NP-hard [14], using a reduction from the assignment problems [49]. In a recent work, we have modeled motivation using *only skill-variety* and we have proved that the problem is NP-hard using a reduction from the Maximum Quadratic Assignment Problem [3]. For our problem, we note that an integer programming based solution is simply not scalable. We will explore greedy heuristic strategies that are effective and efficient. For example, we will assign tasks to the workers greedily based on the marginal gain [49].

Complex modeling for estimating intrinsic motivation & task assignment: In our preliminary direction, we have assumed that variables associated with intrinsic motivations are independent and could be combined as suggested by Hackman and Oldham [19], or intrinsic and extrinsic motivation could be combined linearly. In reality, that may not be the case. In this open problem, we will study the feasibility of a probabilistic model [67], namely a *hierarchical Bayesian framework* [38] for this problem. If the worker is completely new in the platform, we will bootstrap to collect a small set of evidence. We will consider each of the variables associated with worker motivation as a random variable and present a model using hierarchical Bayesian Networks [29] by encoding a joint distribution of these variables over a multi-dimensional space. This model will first establish the relationship among the intrinsic motivational variables themselves and then between intrinsic and extrinsic motivation to capture a workers' "preference" to a given task. We will apply Constraint Based, Score-Based, and Hybrid methods to learn the structure of the network [59]. We will leverage *Bayesian Parameter Estimation as well as Maximum Likelihood Estimation techniques* to learn the parameters of the constructed network. For efficient parameter estimation considering this complex joint distribution, we will use Gibbs sampling [7].

3.3 Optimized Single-Label Acquisition Involving Crowd

We now investigate our proposed optimization framework for single-label acquisition. This problem is examined by augmenting active learning principles with worker-centric optimization (refer to Section 3).

Objectives: We are assuming a setting where single-label acquisition is difficult, expensive, and time consuming (such as, Example 1). We adapt a set of popular as well as well-known active learning principles[34, 52, 42, 47] that are proposed to optimize system-centric cirteria, such as, *minimizing uncertainty or maximizing expected error-reduction* that are known to be effective in supervised (classification) algorithms [58, 50, 37, 32]. We augment these active learning principles with worker-centric optimization. Given a pool of unlabeled instances (of sub-tasks) and an available set of workers, the objective is to select sub-tasks for further labeling and assign workers for annotations, such that, the assignment optimizes both system and workers. The same sub-task may be annotated by multiple workers.

Challenges: An oracle, who knows the ground truth, no longer exists in crowdsourcing; instead, multiple workers, with varying expertise (skill), are available. Under this settings, how to realign traditional active learning goals that are system-centric (i.e., optimizes underlying computational task) requires further investigations. How to systematically design *optimization function*, i.e., one that combines worker-centric optimization in traditional active learning settings [64, 62] is the second important challenge. An equally arduous challenge is the efficiency issue which is mostly overlooked in the existing research. Finally, when to terminate further label acquisition also needs to be examined.

3.4 Proposed Directions

Our overall approach is iterative, where, in each round a set of sub-tasks are selected for annotation and a set of workers are chosen. Once annotations are received, the underlying classification model is retrained. After that, either the process terminates or we repeat. It has three primary directions: (1) *in a given round, which sub-tasks are to be selected for annotation and assigned to which workers?* (2) *how to aggregate multiple annotations to obtain the "true" label?* (3) *when to stop?*

Which sub-tasks are to be selected and assigned to which workers? We take a set of well-known active learning techniques, such as, *uncertainty sampling [34], query-by-committee [52, 42], or expected-error reduction [47],* *used in popular classification algorithms, such as, Naive Bayes [53], SVM [58, 50], Decision Trees [37], or ensemble classification[32]* and study them in crowdsourcing.

When a single classifier with a binary classification task is involved and the classifier is probabilistic (such as Naive Bayes), we consider existing uncertainty sampling [34] techniques. We use entropy [53] to model uncertainty to choose that sub-task for labeling whose posterior probability of being positive is closest to 0.5. For non-probabilistic classifiers (such as SVM or Decision Tree), we explore *heterogeneous approach* [33], in which a probabilistic classifier selects sub-tasks for training the non-probabilistic classifier. We also study existing expected-error reduction [47] techniques that select the sub-tasks to minimize the expected future error of the supervised algorithm, considering *log-loss or* $0/1$*-loss*. We study the query-by-committee[52, 42] technique, we choose that sub-task for further labeling which has the *highest disagreement*.

Active learning principles mentioned above are too *ideal* to be useful in a crowdsourcing platform. A simple alternative is to design a *staged solution*, where we first select the tasks and then the workers [64]. For us, we can take the task-selection solution from [64] and then plug in our worker-centric optimization (Section 3) to compose tasks for the workers. We, however, argue that such a staged solution is *sub-optimal*, simply because, tasks selected by *active learning* techniques may end up having a very low worker-centric optimization, resulting in poor outcome overall. We therefore propose a global optimization that combines (1) worker-centric goals (recall Equation 3). (2) active learning principles considering workers with varying expertise.

Recall Section 2 and note that q^t represents sub-task t's contribution towards a given active learning goal (for example, how much t reduces uncertainty or expected-error) at a given iteration. Let S^{t_u} represent the sub-tasks assigned to u with value $V(S^{t_u})$ (recall Equation 3). Considering worker's skill s^{u_t} as a probability, u's *expected contribution* to t is $s^{u_t} * q^t$ [9]. One possible way to combine them is as a multi-objective global optimization function where the objective is to select sub-tasks and workers that maximize a weighted linear aggregation of worker and task-centric optimization (Equation 4, where W_1, W_2 are specific weights). While linear aggregation is not the only way, it is more likely to admit efficient solutions, where the weights are tunable by domain experts (by default, $W_1 = W_2 = 0.5$).

$$\text{Maximize } \mathcal{V} = \sum_{u \in \mathcal{U}} [W_1 * V(S^{t_u}) + W_2 * \sum_{t \in S^{t_u}} (s^{u_t} * q^t)] \tag{4}$$

Additionally, if a task has a cost budget associated that could be assigned either as a constraint to this optimization problem, or we could use cost as another objective as part of the optimization function, akin to one of our recent works [49]. Nevertheless, we acknowledge that designing the "ideal" optimization model that suffices the need of every application is practically impossible. We address this in the open problems.

Aggregating multiple annotations: Another challenge is how to combine annotations from multiple workers with varying expertise to obtain the "true" label. We apply weighted majority voting types of approach [20], where the weights are chosen according to the skills of the workers. We also consider iterative algorithm for this purpose. Examples of iterative techniques include EM or Expectation Maximization[25]. The main idea behind EM is to compute in the E step the probabilities of possible answers to each task by weighting the answers of workers according to their current expertise, and then to compute in the M step re-estimates of the expertise of workers based on the current probability of each answer. The two steps are iterated until convergence. We explore Bayesian solution [9] to probabilistically obtain the true label, i.e., given workers' annotations and skill, compute $Pr(t = 0)$ and $Pr(t = 1)$ and choose the one which has the higher probability.

3.5 Open Problems

Solving the optimization problem Solving the optimization problem described above is challenging. In a very recent work, we have formalized task assignment as a linear combination of task relevance (based on a Boolean match between worker expertise and the skill requirements of a task) and skill-diversity [43] and proved the problem to be NP-Complete [13, 12]. We use Maximum Quadratic Assignment Problem (MAXQAP in short) [3] to design an efficient algorithm with approximation factor $1/4$. For our problem, we will examine if it is at all possible to design an objective function (perhaps as a special case) to exploit its nice structural properties, such as, *sub-modularity or cancavity*. Such an effort is made for active learning problems recently [22] without considering human workers. We will also study the possibility of staged algorithms and heuristic solutions, as described above. To make the algorithm computationally efficient, we will examine how to design incremental active learning strategies [44], such as finding the new classification model that is most similar to the previous one, under a set of constraints.

Complex function design and stopping condition We note that the formulation described in Equation 4 is rather *simple* - a linear function may not be adequate to combine worker and task-centric optimization. We will explore non-linear multiplicative functions. Another possible way is to formalize this as a bi-criteria optimization problem and design pareto-optimal solution that does not require us to assign any specific weight to the individual functions [6, 2, 4]. Finally, we will examine *when to terminate this iterative process*. For the overall classification task $\mathcal{T}$, when quality threshold is not reached or budget is not exhausted (these are two hard stopping conditions), we will design stopping condition by measuring the confidence [63] of the classification model, or availability of suitable workers.

Develop a number of optimization models that are likely to cover a variety of scenarios We realize that what constitutes the "ideal" optimization model is an extremely difficult problem and highly application dependent (e.g., Which factors are important? Should we add or multiply different human factors? In the case of linear weighting, what should be the weighting coefficients?). Even a domain expert who is very knowledgeable about the specific application may not be able to shed enough light on this. We hope to develop a rich set of different models that will cover the various types of applications. This idea of developing a set of optimization models draw parallels from Web Search and Information Retrieval - where a set of alternative criteria, such as relevance, diversity, and coverage, are considered [5]. In our case, this is analogous to developing models that only consider workers skills/expertise, or cost, or motivation, or includes a subset of human factors that we are interested to study in this project.

4 Optimized Multi-Labels Acquisition Involving Crowd

We now investigate the multi-labels acquisition scenario. We are unaware of any related work that performs multi-label acquisition in an active learning settings involving crowd. Although one can transform a multi-label task to several single-label tasks, this simple approach can generate many tasks, incurring a high cost and latency. Akin to the previous section, our effort is to design solutions that adapt a few recent active learning works [55, 24, 35, 61] for multi-label acquisition and combine that with worker-centric optimization, described in Section 3.

Objectives: We will adapt a few known active learning algorithms for multi-label classifications using Support Vector Machine (SVM), Naive Bayes, or Ensemble classifiers [55, 24, 35, 61]. We will combine and augment them with *worker-centric optimization through human factors modeling*. Using Example 1, this is akin to selecting the most appropriate unidentified image of the species and select the most appropriate workers to provide multiple labels. Since a task could be labeled by multiple workers, we will study how to aggregate multiple responses and infer the correct labels (truth inference problem) of a task. We will also explore the use of correlations among different labels to improve the inference quality. Finally, we will investigate the stopping condition or *convergence criteria*.

Challenges: Workers may exhibit different characteristics in multi-label tasks: a conservative worker would only select labels that the worker is certain of, while a casual worker may select more labels. To determine the multi-label tasks' results, the key is to devise the so-called "worker model" to accurately express the behavior of the worker in answering multi-labels. Furthermore, different from single-label tasks, correlations among labels inherently exist in multi-label tasks. For Example 1, consider one pairwise label dependency: if the insect in the image is labeled as Papilionidae (Family name) , then it is highly probable that it also has label Swallowtail (Sub-family name). Therefore, how to understand and leverage label correlation is another challenge. Finally, how to systematically design *optimization function*, i.e., one that combines worker-centric optimization in active learning settings [55, 24, 35, 61] is the final important challenge.

4.1 Proposed Directions

Our overall approach is iterative here as well, where, in each round a set of sub-tasks are selected to be annotated with multi-labels and a set of workers are chosen. Once multiple labels are acquired, the underlying classification model is retrained. After that, either the process terminates or we repeat. It has three primary directions: (1) *Task assignment* (2) *Truth Inference, i.e., aggregate multiple annotations to obtain the "true" labels.* (3) *Label Correlation*.

Task Assignment: In our preliminary investigation, we have studied the active learning problem for the multi-label scenario considering the widely popular SVM classifier using the *Maximum-Margin Uncertainty Sampling*. Uncertainty sampling [34] is one of the simplest and most effective active learning strategies used for single-label classification. The central idea of this strategy is that the active learner should query the instance which the current classifier is most uncertain about. For binary SVM classifiers, the most uncertain instance can be interpreted as the one closest to the classification boundary by selecting the sample with the smallest classification margin. Multi-label active learning methods simply extend this binary uncertainty concept into the multi-label learning scenarios by integrating the binary uncertainty measures associated with each individual class in independent manners, such as taking the minimum over all classes, and taking the average over all classes.

In our initial direction, given the active learning principle, we combine that with worker-centric optimization and design an objective function akin to Equation 4, as described in Section 3.3. Obviously, exploring alternative optimization models, or how to design a set of optimization functions that can handle a variety of scenarios, or when to stop the iterative process are additional challenges. Once we understand these challenges for the single-label acquisition problem in Section 3.3, we believe they will extend for the multi-label scenarios.

Truth Inference Problem: The truth inference problem, i.e, how to aggregate the annotations provided by multiple workers and generate the actual set of labels requires deeper attention for the multi-label scenario. As the correct set of labels associated with each sub-task is unknown (ground-truth is unknown), the accuracy or expertise of a worker can only be estimated based on the collected answer. To model worker expertise, we compute the following two measures, *True Positive (TP)* and *False Positive (FP)*. TP is the number of labels that a worker selected correctly and FP is the number of labels she selected incorrectly. Unlike a prior work [67], False Negative and True Negative are not relevant, if the workers annotate the labels. In the case where workers validate the given labels, these latter two measures are also relevant. Once these measures are computed, we design a worker's contingency table and calculate her expertise. After that, we design an iterative approach, which can jointly infer the correct labels associated with the tasks and the expertise of the workers. Our iterative solution is motivated by the Expectation Maximization (EM) algorithms and comprises of the following two steps: (step 1), we assume that the worker expertise is known and constant, and infer the probabilistic truth of each object and label pair. (step 2), based on the computed probabilistic truth of each object and label pair, we re-estimate workers expertise.

Label correlation: Since the annotated labels of an object are not independent (Recall Example 1 and note that Papilionidae (Family name) and Swallowtail (Sub-family name) are highly correlated), we study how label correlations can be inferred and facilitate truth inference. In our initial direction, we leverage the existing label correlation techniques [65, 66] to generate the *pairwise label correlations* and regard them as prior input to our problem. For example, the conditional dependency of two labels defines the probability that one label is correct for an object under the condition that the other label is correct. Capturing the higher order label correlations requires computing the joint probability which could be computationally expensive. Once label correlation is computed, we shall explore how to use that information for improved truth inference.

4.2 Open Problems

Alternative Active Learning Strategy Design In our initial direction, we have discussed how to adapt uncertainty sampling to design active learning strategies for SVM classifier for multi-label scenario. The average number of correct labels assigned to each instance in a multi-label data set is called its label cardinality. Thus the number of predicted labels of an unlabeled instance is expected to be consistent with the label cardinality computed on the labeled data. For an unlabeled instance, this inconsistency measure could be defined as the distance between the number of correctly predicted labels so far and the label cardinality of the labeled data. We will study this **label cardinality inconsistency** [60] to select that sub-task where the label inconsistency is highest. Additionally, we will also study the active learning strategies known for other classifiers, such as Naive Bayes and Ensemble methods could be adapted to our problem [55, 24, 35, 61]. Alternatively, task selection can be guided by a version space analysis such that it will give rise to maximum reduction in the version space of the classifier [57].

Truth Inference with Label Correlation We will study how to use the information obtained from label correlation to improve the truth inference. Intuitively, our truth inference problem could benefit from label correlation in the following way: using Example 1, if label correlation infers high correlation among two labels, let's say, Papilionidae and Swallowtail (family and sub-family of butterflies), it is likely that Papilionidae and Mimic Sulfurs (which is a sub-family of butterflies, but Mimic Sulfurs belong to a different family (Pieridae) will have a very low correlation. Therefore, the probabilistic truth of the labels which have Mimic Sulfurs should be downgraded to reflect that fact. It has been shown in Information Retrieval that the more frequent two words occur together in text corpus, the more similar their vectors are [5]. We will regard each label as a word and compute the similarity (e.g., cosine similarity) between the vectors of two labels. We will explore widely popular Sigmoid function [28] to map a probability value to a real value, re-scale the value based on label correlation, and then revert the re-scaled correlation back to a probability score using the Sigmoid function again.

5 Conclusion

The goal of this article is to propose an *an optimized human-machine intelligence framework for single and multi-label tasks through active learning*. We conceptualize an iterative framework that judiciously employs human workers to collect single or multiple labels associated with such tasks, which, in turn are used by the supervised machine algorithms to make intelligent prediction. Our basic approach is adapt a few existing *active learning techniques for single and multi-label classification, but study them in the context of crowdsourcing, especially considering worker-centric optimization, i,e., human factors*. Our innovation lies in systematically characterizing variables to model human factors, designing optimization models that appropriately combine system and worker-centric goals, and designing effective solutions.

6 Acknowledgment

The work of Senjuti Basu Roy is supported by the National Science Foundation under Grant No.: 1814595 and Office of Naval Research under Grant No.: N000141812838.

References

[1] S. Amer-Yahia and S. B. Roy. Human factors in crowdsourcing. *Proceedings of the VLDB Endowment*, 9(13):1615–1618, 2016.

[2] A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, and S. Leonardi. Online team formation in social networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 839–848. ACM, 2012.

[3] E. M. Arkin, R. Hassin, and M. Sviridenko. Approximating the maximum quadratic assignment problem. *Information Processing Letters*, 77(1):13–16, 2001.

[4] A. Asudeh, G. Zhang, N. Hassan, C. Li, and G. V. Zaruba. Crowdsourcing pareto-optimal object finding by pairwise comparisons. *arXiv preprint arXiv:1409.4161*, 2014.

[5] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[6] V. Bilò, M. Flammini, and L. Moscardelli. Pareto approximations for the bicriteria scheduling problem. In *Parallel and Distributed Processing Symposium, 2004. Proceedings. 18th International*, page 83. IEEE, 2004.

[7] C. K. Carter and R. Kohn. On gibbs sampling for state space models. *Biometrika*, 81(3):541–553, 1994.

[8] D. Chandler and A. Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123–133, 2013.

[9] R. T. Clemen and R. L. Winkler. Aggregating probability distributions. *Advances in decision analysis: From foundations to applications*, pages 154–176, 2007.

[10] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[11] E. Estellés-Arolas and F. González-Ladrón-de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200, 2012.

[12] T. Feo, O. Goldschmidt, and M. Khellaf. One-half approximation algorithms for the k-partition problem. *Operations Research*, 40(1-supplement-1):S170–S173, 1992.

[13] T. A. Feo and M. Khellaf. A class of bounded approximation algorithms for graph partitioning. *Networks*, 20(2):181–195, 1990.

[14] M. R. Garey and D. S. Johnson. Computers and intractability: a guide to np-completeness, 1979.

[15] M. B. Garzón, R. Blazek, M. Neteler, R. S. De Dios, H. S. Ollero, and C. Furlanello. Predicting habitat suitability with machine learning models: the potential area of pinus sylvestris l. in the iberian peninsula. *ecological modelling*, 197(3):383–393, 2006.

[16] C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. Shavlik, and X. Zhu. Corleone: hands-off crowdsourcing for entity matching. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 601–612. ACM, 2014.

[17] N. Grira, M. Crucianu, and N. Boujemaa. Active semi-supervised fuzzy clustering for image database categorization. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 9–16. ACM, 2005.

[18] D. Haas, J. Wang, E. Wu, and M. J. Franklin. Clamshell: Speeding up crowds for low-latency data labeling. *Proceedings of the VLDB Endowment*, 9(4):372–383, 2015.

[19] J. R. Hackman and G. R. Oldham. Motivation through the design of work: Test of a theory. *Organizational behavior and human performance*, 16(2):250–279, 1976.

[20] C.-J. Ho, S. Jabbari, and J. W. Vaughan. Adaptive task assignment for crowdsourced classification. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 534–542, 2013.

[21] T. Hofmann and J. M. Buhmann. Active data clustering. *Advances in Neural Information Processing Systems*, pages 528–534, 1998.

[22] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424. ACM, 2006.

[23] Y. Huang and T. M. Mitchell. Text clustering with extended user feedback. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 413–420. ACM, 2006.

[24] C.-W. Hung and H.-T. Lin. Multi-label active learning with auxiliary learner. In *ACML*, pages 315–332, 2011.

[25] N. Q. V. Hung, N. T. Tam, L. N. Tran, and K. Aberer. An evaluation of aggregation techniques in crowdsourcing. In *International Conference on Web Information Systems Engineering*, pages 1–15. Springer, 2013.

[26] P. G. Ipeirotis. Demographics of mechanical turk. 2010.

[27] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM, 2010.

[28] Y. Ito. Approximation of continuous functions on r d by linear combinations of shifted rotations of a sigmoid function with and without scaling. *Neural Networks*, 5(1):105–115, 1992.

[29] F. V. Jensen. *An introduction to Bayesian networks*, volume 210. UCL press London, 1996.

[30] M. Joglekar, H. Garcia-Molina, and A. G. Parameswaran. Evaluating the crowd with confidence. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 686–694, 2013.

[31] N. Kaufmann, T. Schulze, and D. Veit. More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk. In *AMCIS*, volume 11, pages 1–11, 2011.

[32] C. Körner and S. Wrobel. Multi-class ensemble-based active learning. In *Machine Learning: ECML 2006*, pages 687–694. Springer, 2006.

[33] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the eleventh international conference on machine learning*, pages 148–156, 1994.

[34] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[35] X. Li and Y. Guo. Active learning with multi-label svm classification. In *IJCAI*. Citeseer, 2013.

[36] D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

[37] C. X. Ling, Q. Yang, J. Wang, and S. Zhang. Decision trees with minimal costs. In *Proceedings of the twenty-first international conference on Machine learning*, page 69. ACM, 2004.

[38] R. Liu, K. Zolfaghar, S.-c. Chin, S. B. Roy, and A. Teredesai. A framework to recommend interventions for 30-day heart failure readmission risk. In *2014 IEEE International Conference on Data Mining*, pages 911–916. IEEE, 2014.

[39] D. Martin, B. V. Hanrahan, J. O'Neill, and N. Gupta. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 224–235. ACM, 2014.

[40] W. Mason and D. J. Watts. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108, 2010.

[41] B. Mozafari, P. Sarkar, M. Franklin, M. Jordan, and S. Madden. Scaling up crowd-sourcing to very large datasets: a case for active learning. *Proceedings of the VLDB Endowment*, 8(2):125–136, 2014.

[42] K. Nigam and A. McCallum. Employing em in pool-based active learning for text classification. In *Proceedings of ICML-98, 15th International Conference on Machine Learning*, 1998.

[43] J. Pilourdault, S. Amer-Yahia, D. Lee, and S. B. Roy. Motivation-aware task assignment in crowdsourcing. In *EDBT Conference*, 2017.

[44] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang. Two-dimensional multilabel active learning with an efficient online adaptation model for image classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(10):1880–1897, 2009.

[45] H. Rahman, L. Joppa, and S. B. Roy. Feature based task recommendation in crowdsourcing with implicit observations, 2016.

[46] H. Rahman, S. Thirumuruganathan, S. B. Roy, S. Amer-Yahia, and G. Das. Worker skill estimation in team based tasks. *PVLDB*, 2015.

[47] N. Roy and A. McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.

[48] S. B. Roy, I. Lykourentzou, S. Thirumuruganathan, S. Amer-Yahia, and G. Das. Crowds, not drones: Modeling human factors in interactive crowdsourcing. In *DBCrowd 2013-VLDB Workshop on Databases and Crowdsourcing*, pages 39–42. CEUR-WS, 2013.

[49] S. B. Roy, I. Lykourentzou, S. Thirumuruganathan, S. Amer-Yahia, and G. Das. Task assignment optimization in knowledge-intensive crowdsourcing. *The VLDB Journal*, pages 1–25, 2015.

[50] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *ICML*, pages 839–846. Citeseer, 2000.

[51] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

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

[53] C. E. Shannon. A note on the concept of entropy. *Bell System Tech. J*, 27:379–423, 1948.

[54] A. D. Shaw, J. J. Horton, and D. L. Chen. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 275–284. ACM, 2011.

[55] M. Singh, E. Curran, and P. Cunningham. Active learning for multi-label image annotation. In *Proceedings of the 19th Irish Conference on Artificial Intelligence and Cognitive Science*, pages 173–182, 2009.

[56] S. M. Stigler. Gauss and the invention of least squares. *The Annals of Statistics*, pages 465–474, 1981.

[57] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[58] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[59] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[60] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *Dept. of Informatics, Aristotle University of Thessaloniki, Greece*, 2006.

[61] D. Vasisht, A. Damianou, M. Varma, and A. Kapoor. Active learning for sparse bayesian multilabel classification. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 472–481. ACM, 2014.

[62] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. *International Journal of Computer Vision*, 108(1-2):97–114, 2014.

[63] A. Vlachos. A stopping criterion for active learning. *Computer Speech & Language*, 22(3):295–312, 2008.

[64] Y. Yan, G. M. Fung, R. Rosales, and J. G. Dy. Active learning from crowds. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1161–1168, 2011.

[65] Y. Yu, W. Pedrycz, and D. Miao. Multi-label classification by exploiting label correlations. *Expert Systems with Applications*, 41(6):2989–3004, 2014.

[66] M.-L. Zhang and K. Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 999–1008. ACM, 2010.

[67] B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment*, 5(6):550–561, 2012.