JOINT MACHINE LEARNING AND HUMAN LEARNING DESIGN WITH SEQUENTIAL ACTIVE LEARNING AND OUTLIER DETECTION FOR LINEAR REGRESSION PROBLEMS

Xiaohua Li and Jian Zheng

State University of New York at Binghamton
Department of ECE, Binghamton, NY 13902
{xli,jzheng65}@binghamton.edu

ABSTRACT

In this paper, we propose a joint machine learning and human learning design approach to make the training data labeling task more efficient and robust in linear regression problems. We consider a sequential active learning scheme which relies on human learning to enlarge training data set, and integrate with it some outlier detection algorithms to mitigate the inevitable human errors during training data labeling. First, we assume sparse human errors and integrate sparse outlier detection within the sequential active learning procedure. Then, we consider non-sparse human errors and exploit the IRT (item response theory) to model the distribution of human errors, based on which appropriate data can be selected to reconstruct a training data set with sparse human errors. Simulations with some typical linear regression data set and human-subject experiment data are conducted to show the desirable performance of the proposed scheme.

Index Terms— machine learning, human learning, item response theory, linear regression, active learning, training

1. INTRODUCTION

Machine learning has found wide applications today due to the rapidly increasing amount of data which is becoming prohibitively demanding for human to process. While machine learning is dominating, the role of human learning should not be overlooked, and it is especially interesting to investigate how to integrate machine learning and human learning together so as to take both of their advantages and to use one of mitigate the challenge of the other.

It is well known that human plays important roles in machine learning design, feature selection, algorithm development, etc [2]. Some machine learning algorithms are developed from the inspiration of human learning principles. In addition, human learning has many important characteristics that can be helpful to resolve many inherent challenges of machine learning, such as case representation, feature selection, over-fitting, generalization, etc. On the other hand, human learning suffers many inherent weaknesses such as the limited data amount capability, errors, bias, etc. Innovative machine learning algorithms can be developed to mitigate such weaknesses while exploiting the benefits of machine learning.

To show the great benefits of integrating machine learning and human learning together, we focus on a typical task that needs both of them, i.e., training data optimization in robust linear regression. Linear regression is one of the important data processing tasks where machine learning has attracted great research effort and has found wide application [1]. Supervised machine learning such as linear regression requires a human-labeled training data set that must be sufficiently long, well case-representative, and correctly labeled. However, considering the high complexity and dimensionality of many practical linear regression tasks, training data may be insufficient, biased, skewed, and error-prone. This causes many problems, such as the well-known over-fitting problem in machine learning [2].

To resolve the insufficient training data issue, one of the popular approaches is active learning [3]-[6]. In contrast, the human labeling error issue is more challenging. The cause of human error is complex, and may depend on the data processing task, noise level, human workload, human cognition capability, etc.

In this paper, we address these issues together within a framework of joint human learning and machine learning design. Specifically, we consider the case when the initial training data labeled by human are both insufficient and error-prone. We will develop two new robust linear regression algorithms based on both the active learning method [6] and the robust linear regression method [7]. In case the error probability is low enough so that the labeling errors become sparse, we integrate the robust linear regression approach of [7] into a sequential active learning scheme adapted from [6] to resolve this issue, which can estimate the sparse labeling errors while learning the linear regression vectors. We apply sequential active learning with human learning to look for extra and better training data under appropriate human workload considerations. Then, more importantly, in case the error probability is not low and the labeling errors are not sparse, we apply the item response theory (IRT) [8] to convert the non-sparse human error cases into the sparse human error case.

The organization of this paper is as follows. In Section 2, we give the linear regression model. In Section 3, we develop the new algorithms with machine learning and human learning tightly coupled together. Simulations are conducted in Section 4, and conclusions are given in Section 5.

2. LINEAR REGRESSION MODEL WITH ACTIVE LEARNING

We consider the classical linear regression problem, where a scalar response y_i is to be predicted using N known (input) data samples $\tilde{\mathbf{x}}_i = [\tilde{x}_{i,1}, \cdots, \tilde{x}_{i,N}]^T$, where $(\cdot)^T$ denotes transpose. The data model of the linear regression problem is

$$\begin{aligned} y_i &= f(\tilde{\mathbf{x}}_i) + \epsilon_i, \\ &= \sum_{\ell=1}^{N} \theta_\ell \phi_\ell(\tilde{\mathbf{x}}_i) + \epsilon_i, \\ &= \mathbf{x}_i^T \boldsymbol{\theta} + \epsilon_i, \quad i = 1, \cdots, I, \end{aligned} \tag{1}$$

where $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_N]^T$ is the $N \times 1$ regression vector, ϵ_i is the noise (or modeling error), and I is the total number of data records. We assume *i.i.d.* Gaussian noise ϵ_i with zero-mean and variance σ_ϵ^2. The functions $\phi_\ell(\tilde{\mathbf{x}}_i)$ are fixed linearly independent functions and $\mathbf{x}_i = [x_{i,1}, \cdots, x_{i,N}]^T$ where $x_{i,\ell} = \phi_\ell(\tilde{\mathbf{x}}_i)$.

Without loss of generality, as in typical supervised linear regression algorithms, we label and use the first L data records $(y_i, \mathbf{x}_i)$, $i = 1, \cdots, L$, as the training data set. The values of y_i are labeled by human. Stacking together all the training data records, we have

$$\mathbf{y}_L = \mathbf{X}_L \boldsymbol{\theta} + \boldsymbol{\epsilon}_L, \tag{2}$$

where $\mathbf{X}_L = [\mathbf{x}_1, \cdots, \mathbf{x}_L]^T$ is the $L \times N$ input data matrix, $\mathbf{y}_L = [y_1, \cdots, y_N]^T$ is the $N \times 1$ labeled output data vector, and $\boldsymbol{\epsilon}_L = [\epsilon_1, \cdots, \epsilon_L]^T$ is the Gaussian noise vector. Assume that the training data $\mathbf{X}_L$ is obtained with distribution $p(\mathbf{x})$ while the testing data (or the overall data) $\mathbf{x}_i$ has distribution $q(\mathbf{x})$. Define the $N \times N$ diagonal matrix $\mathbf{D}$ with the i-th diagonal element as $D_{i,i} = q(\mathbf{x}_i)/p(\mathbf{x}_i)$. A standard weighted least-squares learning gives the optimal estimation of $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{X}_L^H \mathbf{D} \mathbf{X}_L\right)^+ \mathbf{X}_L^H \mathbf{D} \mathbf{y}_L, \tag{3}$$

where $(\cdot)^H$ denotes Hermitian, and $(\cdot)^+$ denotes pseudo-inverse. The use of pseudo-inverse rather than matrix inverse permits us to consider many special training data issues, such as labeling errors, skewed training data, or insufficient amount of training data, etc. These issues may make the matrix $\mathbf{X}_L^H \mathbf{D} \mathbf{X}_L$ singular or ill-conditioned.

The estimated $\hat{\boldsymbol{\theta}}$ can then be used to predict the output $\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\theta}}$ for all the unlabeled data $\mathbf{x}_i$, $i = L+1, \cdots, I$. The performance is measured by the generalization error $\int (\hat{y}_i - y_i)^2 q(\mathbf{x}_i) d\mathbf{x}_i$.

Active learning is a general methodology to deal with the insufficient training data issue. With active learning, linear regression algorithms select some extra data records from $\{\mathbf{x}_i | L+1 \le i \le I\}$ and ask human for labeling. These newly labeled data records will be used together with the initial training data set $(\mathbf{y}_L, \mathbf{X}_L)$. Considering the human workload constraint, the number J of the newly labeled data records can not be too big. Therefore, active learning needs to find the extra data records that contribute the most to the existing training data set. The sequential active learning algorithm adds these extra data records iteratively (sequentially) so as to minimize the expected generalization error.

3. INTEGRATING MACHINE LEARNING WITH HUMAN LEARNING IN SEQUENTIAL ACTIVE LEARNING

3.1. Combining sequential active learning and sparse optimization for sparse human error

Consider first the insufficient training issue in linear regression. We adapt the active learning algorithm of [6] into a sequential active learning algorithm to label J extra training data, where $J \le I - L$. This can be implemented iteratively in J_1 iterations. In each iteration j, where $j = 1, \cdots, J_1$, we need to select some new training data vector $\mathbf{z}_j$ from the set $\mathcal{X}_j = \{\mathbf{x}_\ell | L+1 \le \ell \le I, \ \mathbf{x}_\ell \ne \mathbf{z}_i, \ 1 \le i \le j-1\}$. There are $I - L - (j-1)J/J_1$ data vectors $\mathbf{x}_\ell$ in the set $\mathcal{X}_j$.

Without loss of generality, let us consider the jth iteration. In the beginning this iteration, before selecting $\mathbf{z}_j$, we have the labeled training data set $(\mathbf{y}_{L+j-1}, \mathbf{X}_{L+j-1})$, where $\mathbf{y}_{L+j-1} = [\mathbf{y}_L^T, \ u_1, \ \cdots, \ u_{j-1}]^T$, and u_i is the correct labeling of the data record $u_i = \mathbf{z}_i^T \boldsymbol{\theta} + \epsilon_i$. Note that $(u_i, \mathbf{z}_i)$, $i = 1, \cdots, j-1$, are the extra training data selected and labeled in the previous $j-1$ iterations. With this labeled data set, from (3) we have the estimation

$$\hat{\boldsymbol{\theta}}(j-1) = \left(\mathbf{X}_{L+j-1}^H \mathbf{D} \mathbf{X}_{L+j-1}\right)^+ \mathbf{X}_{L+j-1}^H \mathbf{D} \mathbf{y}_{L+j-1}. \tag{4}$$

We let $\hat{\boldsymbol{\theta}}(0) \stackrel{\triangle}{=} \hat{\boldsymbol{\theta}}$ of (3) as the initial condition.

To select the new training data vector $\mathbf{z}_j$, we solve the following optimization. First, we pre-select a set of probability density functions $\mathcal{P}$ which provides various approximation of the true distribution $q(\mathbf{x})$. For each density function $p(\mathbf{x}) \in \mathcal{P}$, we select a set of new J/J_1 training data following the distribution $p(\mathbf{x})$ and calculate

$$\begin{aligned} F(p(\mathbf{x})) = \mathrm{tr} \Big(\mathbf{U} \left(\mathbf{X}_{L+j-1}^H \mathbf{D} \mathbf{X}_{L+j-1}\right)^+ \mathbf{X}_{L+j-1}^H \mathbf{D} \\ \left[\left(\mathbf{X}_{L+j-1}^H \mathbf{D} \mathbf{X}_{L+j-1}\right)^+ \mathbf{X}_{L+j-1}^H \mathbf{D}\right]^T \Big), \end{aligned} \tag{5}$$

where the matrix $\mathbf{U}$ is $N \times N$ with the (i,j)-th element as

$$U_{i,j} = \int \phi_i(\mathbf{x}) \phi_j(\mathbf{x}) q(\mathbf{x}) d\mathbf{x}. \tag{6}$$

Then, we determine the best distribution $p(\mathbf{x})$ and the associated new data training set by solving the optimization

$$\mathbf{z}_j = \arg \min_{\mathbf{x}\in\mathcal{X}_j} \; J(p(\mathbf{x})), \tag{7}$$

Details are similar to [6], where the difference is that we adapted it into a sequential active learning scheme.

After $\mathbf{z}_j$ is selected, human learning kicks in to label the output u_j. Then we can insert the new training data $(u_j, \mathbf{z}_j)$ into the existing training data set to form $(\mathbf{y}_{L+j}, \mathbf{X}_{L+j})$ where $\mathbf{X}_{L+j} = \left[\begin{array}{cc} \mathbf{X}_{L+j-1}^T & \mathbf{z}_j \end{array}\right]^T$ and $\mathbf{y}_{L+j} = [\mathbf{y}_{L+j-1}^T, u_j]^T$, and calculate $\hat{\boldsymbol{\theta}}(j)$ similarly to (4). After this, the new iteration $j+1$ will start.

We need to address the inevitable human labeling errors in this active learning procedure. Human errors can affect all the training data. Following the robust linear regression formulation of [7] which deal with outliers, we model the labeling error by o_i which changes the true value model (1) to the error labeling model

$$y_i = \mathbf{x}_i^T\boldsymbol{\theta} + \epsilon_i + o_i, \quad i = 1, \cdots, I. \tag{8}$$

Note that the labeled value y_i in (8) may no longer be the true value of (1). However, for notational convenience, we reuse the same variable y_i. Similarly, although o_i exists in the training data set only, we have defined o_i for all $1 \le i \le I$ since each i is the selection and labeling candidate in the sequential active learning procedure.

We assume that the vector $\mathbf{o}_I = [o_1, \cdots, o_I]^T$ is sparse in this subsection. Non-sparse $\mathbf{o}_I$ will be addressed in the next subsection.

Consider again the jth iteration of the sequential active learning procedure. We need to revise (4) so as to estimate $\hat{\boldsymbol{\theta}}(j-1)$ robustly from human labeling errors. This can be conducted by the joint optimization

$$\min_{\boldsymbol{\theta}, \mathbf{o}_{L+j-1}} \|\mathbf{y}_{L+j-1} - \mathbf{o}_{L+j-1} - \mathbf{X}_{L+j-1}\boldsymbol{\theta}\| + \lambda_0 \|\mathbf{o}_{L+j-1}\|_0, \tag{9}$$

which estimates $\hat{\boldsymbol{\theta}}(j-1)$ and the sparse labeling error vector $\mathbf{o}_{L+j-1} = [o_1, \cdots, o_{L+j-1}]^T$ simultaneously. The ℓ_0 norm $\|\mathbf{o}_{L+j-1}\|_0$ is to guarantee the sparsity of the human error vector $\mathbf{o}_{L+j-1}$. By choosing appropriate weighting coefficient λ_0, we can make $\mathbf{o}_{L+j-1}$ to have various sparsity values.

Because ℓ_0 norm is not convex, we can replace it by the convex ℓ_1 norm. Then the optimization (9) is changed to

$$\min_{\boldsymbol{\theta}, \mathbf{o}_{L+j-1}} \|\mathbf{y}_{L+j-1} - \mathbf{o}_{L+j-1} - \mathbf{X}_{L+j-1}\boldsymbol{\theta}\| + \lambda_1 \|\mathbf{o}_{L+j-1}\|_1, \tag{10}$$

which is convex in either $\boldsymbol{\theta}$ or $\mathbf{o}_{L+j-1}$.

As shown in [7], a two-step procedure can find the solution to (10). First, conditioned on $\boldsymbol{\theta}$, we estimate $\mathbf{o}_{L+j-1}$ from the convex optimization

$$\begin{aligned}\hat{\mathbf{o}}_{L+j-1} = &\arg \min_{\mathbf{o}_{L+j-1}} \|\mathbf{y}_{L+j-1} - \mathbf{o}_{L+j-1} - \mathbf{X}_{L+j-1}\boldsymbol{\theta}\| \\ &+ \lambda_1 \|\mathbf{o}_{L+j-1}\|_1. \end{aligned} \tag{11}$$

Second, with the estimated $\hat{\mathbf{o}}_{L+j-1}$, we estimate $\boldsymbol{\theta}(j-1)$ as

$$\begin{aligned}\hat{\boldsymbol{\theta}}(j-1) = &\left(\mathbf{X}_{L+j-1}^H \mathbf{D} \mathbf{X}_{L+j-1}\right)^+ \mathbf{X}_{L+j-1}^H (\mathbf{D}\mathbf{y}_{L+j-1} \\ &- \hat{\mathbf{o}}_{L+j-1}). \end{aligned} \tag{12}$$

Furthermore, we can replace $\boldsymbol{\theta}$ of (11) by $\hat{\boldsymbol{\theta}}(j-1)$ of (12), which changes the optimization (11) to

$$\begin{aligned}&\hat{\mathbf{o}}_{L+j-1} = \\ &\arg \min_{\mathbf{o}_{L+j-1}} \|(\mathbf{I}_{L+j-1} - \mathbf{X}_{L+j-1}(\mathbf{X}_{L+j-1}^H \mathbf{D} \mathbf{X}_{L+j-1})^+ \\ &\times \mathbf{X}_{L+j-1}^H)(\mathbf{D}\mathbf{y}_{L+j-1} - \mathbf{o}_{L+j-1})\| + \lambda_1 \|\mathbf{o}_{L+j-1}\|_1, \end{aligned} \tag{13}$$

where $\mathbf{I}_{L+j-1}$ is an $(L+j-1)\times(L+j-1)$ dimensional identity matrix. The vector $\hat{\mathbf{o}}_{L+j-1}$ now depends on the training data set $(\mathbf{y}_{L+j-1}, \mathbf{X}_{L+j-1})$ only, not on the regression vector $\boldsymbol{\theta}$.

Therefore, to solve the optimization (10), we first solve (13) to obtain $\hat{\mathbf{o}}_{L+j-1}$ via convex optimization, and then use (12) to calculate $\hat{\boldsymbol{\theta}}(j-1)$. This replaces (4) in the sequential active learning procedure.

To conduct the next step of the sequential active learning, i.e., optimizing (7) so as to select the new data record $\mathbf{z}_j$, we need to evaluate $\widetilde{\boldsymbol{\theta}}(j)$. Based on (7) and (12), we have

$$\widetilde{\boldsymbol{\theta}}(j) = \left(\widetilde{\mathbf{X}}_{L+j}^H \widetilde{\mathbf{D}\mathbf{X}}_{L+j}\right)^+ \widetilde{\mathbf{X}}_{L+j}^H \mathbf{D} \left[\begin{array}{c} \mathbf{y}_{L+j-1} - \hat{\mathbf{o}}_{L+j-1} \\ \mathbf{z}^T \hat{\boldsymbol{\theta}}(j-1) \end{array}\right]. \tag{14}$$

The estimation $\hat{\mathbf{o}}_{L+j-1}$ of (13) is still used in (14).

In summary, the algorithm for the sequential selection of the extra training data while mitigating the sparse human labeling errors is given below.

Alg. 1: Sequential active learning & sparse human error
i) Initialize: Learning with training $(\mathbf{y}_L, \mathbf{X}_L)$ (13) (12)
ii) For iteration $j = 1, 2, \cdots, J_1$, do
1) for each $\mathbf{z} \in \mathcal{X}_j$, calculate $\widetilde{\boldsymbol{\theta}}(j)$ (14), 2) select the optimal $\mathbf{z}_j$ that optimizes (7), 3) label u_j for new training data $(u_j, \mathbf{z}_j)$ (human learning), 4) form $(\mathbf{y}_{L+j}, \mathbf{X}_{L+j})$, update linear regression (13) (12).
Output $\hat{\boldsymbol{\theta}}(J_1)$ and linear prediction $\hat{\mathbf{y}}_I = \mathbf{X}_I \hat{\boldsymbol{\theta}}(J_1)$.

Inside this machine learning algorithm, the step (3) involves human learning to label the extra training data. There are J_1 iterations to find J extra training data, and there is a convex optimization in each iteration. The value J can be adjusted according to human workload and human error principles.

3.2. Robust linear regression for non-sparse human errors

One of the major limitations of Algorithm 1 is that the labeling error vector $\mathbf{o}_{L+j}$ has to be sparse in order for the convex sparse optimization to work. There are many cases that human errors are non-sparse. In this subsection, we develop a way to reconstruct a new training data set with sparse human labeling errors. This is conducted by removing the training data that are more likely to have errors. By using those data that are less likely to have errors, we can effectively change the non-sparse error cases into the sparse case so Algorithm 1 can still be used.

We model the human labeling error by the item response theory (IRT). The basic idea of IRT is to use some item response function (IRF) to describe the probability for human to make correct decisions on a task [8]. As a typical IRF, a person with cognition capability (or intelligence) s can label the ith training data correctly with probability

$$q_i = c_i + \frac{1-c_i}{1+e^{-a_i(s-b_i)}}, \tag{15}$$

where the parameter b_i denotes the difficulty level of labeling the ith data, c_i and a_i are systematic parameters regarding the data labeling task. We assume that b_i depends only on the noise magnitude $|\epsilon_i|$, since higher noise makes human labeling more difficult. The probability of error-labeling of the data record $(y_i, \mathbf{x}_i)$ is then

$$p(|\epsilon_i|) = (1-c_i)\left(1 - \frac{1}{1+e^{-a_i(s-|\epsilon_i|)}}\right), \tag{16}$$

which is an increasing function of $|\epsilon_i|$.

From the model (8), if the data are real, then $|\epsilon_i|$ has folded normal distribution with probability density function $g(x) = \frac{2}{\sigma_\epsilon\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_\epsilon^2}}$. It has mean $\sigma_\epsilon\sqrt{2/\pi}$ and variance $(\pi-2)\sigma_\epsilon^2/\pi$. If the data are complex, then $|\epsilon_i|$ has Rayleigh distribution with probability density function $g(x) = \frac{x}{\sigma_\epsilon^2} e^{-\frac{x^2}{2\sigma_\epsilon^2}}$, whose mean and variance are $\sigma_\epsilon\sqrt{\pi/2}$ and $(4-\pi)\sigma_\epsilon^2/2$, respectively.

The percentage of error-labeled training data, or the average probability for a data to be labeled in error, is

$$\overline{P} = \int_0^\infty p(x)g(x)dx. \tag{17}$$

We have $\overline{P} \in [0,1]$. $(L+j)\overline{P}$ is the number of non-zero entries in $\mathbf{o}_{L+j}$. Obviously, a small $\overline{P}$ means sparse labeling error while a large $\overline{P}$ means non-sparse labeling error. Note that the value of the error o_i can have various distributions which are assumed unknown in this paper.

Consider the jth iteration of the sequential active learning with labeling error mitigation. Define $\mathbf{v}_{L+j-1} = \mathbf{y}_{L+j-1} - \mathbf{X}_{L+j-1}\boldsymbol{\theta} = \mathbf{o}_{L+j-1} + \boldsymbol{\epsilon}_{L+j-1}$, which contains all the error and noise information. Replacing $\boldsymbol{\theta}$ with the standard linear regression vector estimation $\hat{\boldsymbol{\theta}}$ of (3), we have

$$\begin{aligned}\mathbf{v}_{L+j-1} = &(\mathbf{I}_{L+j-1} \\ &-\mathbf{X}_{L+j-1}(\mathbf{X}_{L+j-1}^H\mathbf{D}\mathbf{X}_{L+j-1})^+\mathbf{X}_{L+j-1}^H\big)\,\mathbf{D}\mathbf{y}_{L+j-1}.\end{aligned} \tag{18}$$

Then we can use $\mathbf{v}_{L+j-1} = [v_1, \cdots, v_{L+j-1}]^T$ of (18) to determine approximately whether each v_i has error or not.

Without the error o_i, the distribution of $v_i = \epsilon_i$ is $\mathcal{N}(0, \sigma_\epsilon^2)$. With the error o_i, the distribution of $v_i = o_i + \epsilon_i$ is the convolution of the distributions of ϵ_i and o_i which is unknown. Nevertheless, based on the error-labeling distribution $p(|\epsilon_i|)$ of (16), to reduce the percentage of labeling errors from $\overline{P}$ to $\eta\overline{P}$, we just need to find a threshold value γ such that

$$(1-p(|\epsilon_i|))\mathbb{P}[|\epsilon_i| < \gamma] > (1-\eta\overline{P})\mathbb{P}[|v_i| < \gamma]. \tag{19}$$

With the threshold γ, we select all the labeled data that satisfy $|v_i| < \gamma$ to construct the new training data set. All the data with $|v_i| \geq \gamma$ are removed from the new training data set. Some approximated estimation of γ based on the numerical evaluation of v_i's distribution from (18) will be reliable enough for us to create a new training data set that have sparse labeling errors.

To formulate the new learning framework, we introduce a diagonal $(L+j-1)\times(L+j-1)$ weighting matrix $\mathbf{W}$. If o_i can have value $|o_i|$ much larger than $|y_i|$, it is better to apply hard decision when selecting the training data. Therefore, we use $\mathbf{W}$ with diagonal elements

$$w_{i,i} = \begin{cases} 1, & \text{if } |v_i| < \gamma \\ 0, & \text{if } |v_i| \geq \gamma \end{cases} \tag{20}$$

On the other hand, if the value $|o_i|$is mostly comparable to $|y_i|$, then a soft-decision diagonal matrix $\mathbf{W}$ with $w_{i,i} = \mathbb{P}[|v_i| < \gamma]$ can also be used.

With the weighting matrix $\mathbf{W}$, the Algorithm 1 can be easily changed to use just the new training data set. Specifically, the sparse-optimization (10) becomes

$$\min_{\boldsymbol{\theta}, \mathbf{o}_{L+j-1}} \|\mathbf{W}(\mathbf{y}_{L+j-1} - \mathbf{o}_{L+j-1} - \mathbf{X}_{L+j-1}\boldsymbol{\theta})\| + \lambda_1\|\mathbf{W}\mathbf{o}_{L+j-1}\|_1. \tag{21}$$

The solution (13)(12) can be changed to

$$\begin{aligned}&\hat{\mathbf{o}}_{L+j-1} = \\ &\arg\min_{\mathbf{o}_{L+j-1}} \|(\mathbf{I} - \mathbf{X}_{L+j-1}(\mathbf{X}_{L+j-1}^H\mathbf{D}\mathbf{X}_{L+j-1})^+ \\ &\times \mathbf{X}_{L+j-1}^H)\mathbf{D}\mathbf{W}(\mathbf{y}_{L+j-1} - \mathbf{o}_{L+j-1})\| + \lambda_1\|\mathbf{W}\mathbf{o}_{L+j-1}\|_1,\end{aligned} \tag{22}$$

and

$$\begin{aligned}\hat{\boldsymbol{\theta}}(j-1) = &\left(\mathbf{X}_{L+j-1}^H\mathbf{D}\mathbf{X}_{L+j-1}\right)^+\mathbf{X}_{L+j-1}^H\mathbf{D}\mathbf{W}(\mathbf{y}_{L+j-1} \\ &- \hat{\mathbf{o}}_{L+j-1}).\end{aligned} \tag{23}$$

Note that even though $\mathbf{o}_{L+j-1}$ may not be a sparse vector, $\mathbf{W}\mathbf{o}_{L+j-1}$ is sparse. In practical implementations, we can simply remove those data that are not used from the convex optimization.

In summary, we can modify Algorithm 1 into the following Algorithm 2 which can work with training data that have non-sparse labeling errors.

Alg. 2: Sequential active learning with sparsity recovery
i) Initialize: Determine $\mathbf{W}$ based on (18)-(20); Linear regression with training $(\mathbf{y}_L, \mathbf{X}_L)$ (22) (23).
ii) For iteration $j = 1, 2, \cdots, J$, do
1)-4) same as 1)-4) of Algorithm 1, except using (22)(23)
5) Update weight matrix $\mathbf{W}$ using training $(\mathbf{y}_{L+j}, \mathbf{X}_{L+j})$.

4. SIMULATIONS

To verify the performance of Algorithm 1 in resolving problems of insufficient training and sparse labeling errors, we used simulation settings similar to [7]. We let $I = 100$, $N = 10$, $\boldsymbol{\theta} \sim \mathcal{N}(10 \times \mathbf{1}_N, \mathbf{I}_N)$, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}_N, \mathbf{I}_N)$, and $\epsilon_i \sim \mathcal{N}(0, 1)$. We used two human labeling error models: 1) human errors were modeled with Laplacian distribution $o_i \sim \mathcal{L}(0, 10^3)$, and 2) human errors were generated based on a human subject experiment (where we asked students for data labeling in some surveys). The initial training data set size was $L = 15$, and an extra $J = 10$ training data were to be found in active learning.

We compared our new algorithm (**Algorithm 1**) with four other algorithms: **KeepOut:L** which just implemented (3) with L training data; **KeepOut:L+J** which implemented (3) with $L + J$ training data; **KeepOut:Active** which implemented the active learning algorithm of [6]; and **RmvOut:L** which implemented [7] with L training data. Note that the three $\mathbf{KeepOut}$ algorithms did not use any way to mitigate labeling errors. We evaluated NRMSE (normalized root mean square error) of the regression vector estimation $\sqrt{E[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 / \|\boldsymbol{\theta}\|^2]}$ over 100 runs of experiments for each labeling error probability. The simulation results in Fig. 1 clearly show the superior performance of our new Algorithm 1 in robust linear regression.

Next, we evaluated our Algorithm 2 in resolving the problems of insufficient training and non-sparse human labeling errors. We set $I = 500$. While the other algorithms used 10% data for training, our Algorithm 2 used 15 initial training data and searched for more extra training data. Human errors were introduced based on the IRF with appropriate parameters to create various human labeling error probabilities. Simulation results in Fig. 2 clearly show that our algorithm had superior performance because of both using active learning to recruit more training data and using the IRT model to remove those training data with high error probabilities.

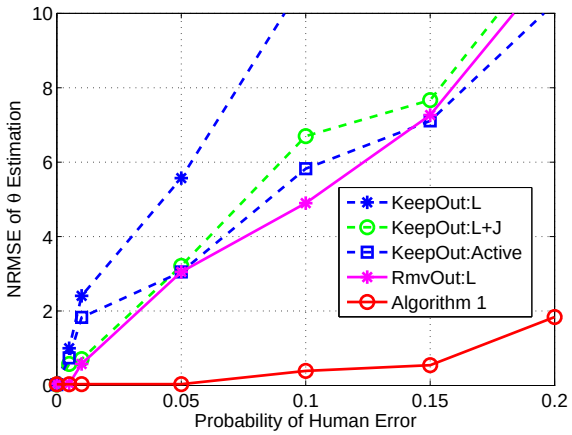

Fig. 1. NRMSE of the estimation of $\boldsymbol{\theta}$ for sparse labeling errors.

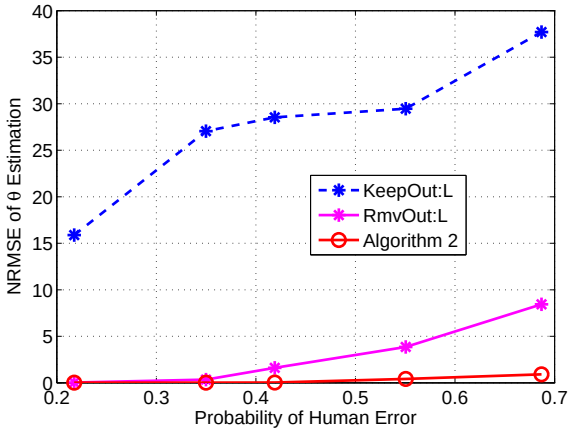

Fig. 2. NRMSE of the estimation of $\boldsymbol{\theta}$ for non-sparse human labeling errors.

5. CONCLUSIONS

In this paper we formulated a joint machine learning and human learning framework for linear regression so as to enhance the robustness to insufficient and error training data. Machine learning is applied to search for more and better training data and to estimate human labeling errors, while human learning is applied to label the extra training data. The IRT (item response theory) model of human errors is applied for removing potentially error-prone data so as to keep the sparsity of the labeling errors. Simulations are conducted to verify the performance of the proposed algorithms.

6. REFERENCES

[1] X. Yan, *Linear Regression Analysis: Theory and Computing*, World Scientific, 2009.

[2] X. Jiang, "Linear subspace learning-based dimensionality reduction," *IEEE Signal Processing Mag.*, vol. 28, no. 2, pp. 16-26, Mar. 2011.

[3] B. Settles, *Active Learning*, Morgan & Claypool, 2012.

[4] R. Castro, R. Willet, and R. Nowak, "Faster rates in regression via active learning," *Proc. Advances in Neural Informaiton Processing Systems*(NIPS'05), Vancouver, Canada, Dec. 2005.

[5] R. Castro, C. Kalish, R. Nowak, R. Qian, T. Rogers and X. Zhu, "Human active learning," *Proc. Advances in Neural Information Processing Systems* (NIPS'08), Vancouver, Canada, Dec. 2008

[6] M. Sugiyama, "Active learning in approximately linear regression based on conditional expectation of generalization errors," *Journal of Machine Learning Research*, vol. 7, pp. 141-166, 2006.

[7] G. B. Giannakis, G. Mateos, S. Farahmand, V. Kekatos, and H. Zhu, "USPACOR: Universal sparsity-controlling outlier rejection," *Proc. ICASSP*, Prague, pp. 1952-1955, May 2011.

[8] D. Thissen and L. Steinberg, "Item response theory," in *The SAGE Handbook of Quantitative Methods in Psychology*, R. E. Millsap and A. Maydeu-Olivares (Edt.), pp. 148-177, SAGE Publications Ltd, 2009.