Active Learning for Regression With Correlation Matching and Labeling Error Suppression

Xiaohua Li, *Senior Member, IEEE*, and Jian Zheng

***Abstract*—In this letter, we develop an active learning algorithm to optimize the selection of training data for robust linear regression. This algorithm selects training data based on the principle of correlation matching between the training dataset and the overall data pool. Considering the inevitable and potentially heavy human labeling errors, we model the probability of labeling errors based on the item response theory (IRT) and develop data screening techniques to control the error sparsity. Compressive sensing theory is then exploited for human labeling error suppression. This algorithm is robust even in the case of short training dataset with nonsparse labeling errors. Its performance is verified by simulations with both artificial data and real benchmark data. Experiments are also conducted to demonstrate the validity of the IRT-based human labeling error model and the superior performance of the algorithm in practical applications.**

***Index Terms*—Active learning, compressive sensing, human labeling error, item response theory (IRT), linear regression.**

I. Introduction

THE objective of active learning is to minimize the amount of training data by looking for the most informative ones. It is useful when the training data are costly to label or when they have to be transmitted through bandwidth/power limited wireless networks [1]. Active learning is becoming increasingly important nowadays because many modern machine learning problems have large dimensions and need an enormous amount of training data, which has made data labeling a severe bottleneck.

For regression, many effective active learning algorithms have been developed. In [2], active regression was conducted based on output variance minimization, which was shown equivalent to minimizing the generalization error, i.e., the error when applying the regression results to the test data. An active regression algorithm that directly minimizes the expected generalization error was developed in [3]. More recently, the algorithm in [4] was based on the principle of maximizing the expected model change. The algorithms in [5] and [6] used the stratification and vector norm maximization techniques, respectively. Sequential active regression was studied in [7] and [8] under a concept of integrated human and machine learning.

Manuscript received April 26, 2016; revised June 23, 2016; accepted June 24, 2016. Date of publication June 28, 2016; date of current version July 08, 2016. This work was supported by the National Science Foundation under Grant CNS-1443885. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Magno T. M. Silva.

X. Li and J. Zheng are with the Department of Electrical and Computer Engineering, State University of New York at Binghamton, Binghamton, NY 13902 USA (e-mail: xli@binghamton.edu; jzheng65@binghamton.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2016.2585496

It has been shown that active learning can outperform the conventional passive learning, as seen in [5], [6], and [9]. Nevertheless, active learning usually suffers from a severe performance fluctuation in practice. Active learning tends to select the most ambiguous and noisy data for training, which unfortunately causes not only high regression errors but also heavy human labeling errors. It is more error-prone for human labelers to work mostly on highly ambiguous and noisy data.

Modeled as outliers, human labeling errors can be mitigated by robust regression techniques, including both conventional techniques such as random sample consensus [10] and new techniques such as compressive sensing [11]–[15]. However, in active learning, outliers may aggregate to a high level within the limited training dataset, which will render these techniques ineffective.

The human labeling error problem has not been studied sufficiently in active regression [1]. In this letter, we develop a new active regression algorithm to address this problem. We apply correlation matching for fast active learning and compressive sensing for efficient outlier suppression. In order to resolve the nonsparse outlier problem, we exploit the item response theory (IRT) [16] to describe the outlier probability and develop data screening techniques for outlier sparsity control.

The organization of this letter is as follows. In Section II, we give the regression model. In Section III, we develop the new algorithm. Simulation and experiment results are presented in Section IV, and conclusion is given in Section V.

II. Regression and Human Labeling Error Model

We consider the linear regression problem where the unknown response r_i to a data vector $\mathbf{x}_i$ is to be predicted via a linear regression vector $\boldsymbol{\theta}$. The true response is

$$r_i = \mathbf{x}_i'\boldsymbol{\theta} + \epsilon_i, \quad i = 1, \ldots, I \tag{1}$$

where $\mathbf{x}_i$ and $\boldsymbol{\theta}$ are both $N \times 1$ dimensional vectors, and the modeling error $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is modeled as a Gaussian noise with zero mean and variance σ_ϵ^2. We consider real data, so $(\cdot)'$ denotes transpose.

We need to estimate $\boldsymbol{\theta}$ by training. If $\mathbf{x}_i$ is selected as training data, a labeler generates a label y_i, where

$$y_i = \mathbf{x}_i'\boldsymbol{\theta} + \epsilon_i + v_i \tag{2}$$

and v_i is the human labeling error (outlier). From the pool of I data vectors, we need to select and label T of them to construct a training dataset $\{(\mathbf{x}_{t_\ell}, y_{t_\ell}) | t_\ell \in \{1, \ldots, I\}, \ell = 1, \ldots, T\}$ and use them to estimate $\boldsymbol{\theta}$.

It is challenging to model and analyze quantitatively the human labeling error v_i in linear regression. In this letter, we model

$v_i = H_i \tilde{v}_i$, where $H_i \in \{0, 1\}$ is a Bernoulli random variable describing whether there is significant human error over the regular noise ϵ_i, and $\tilde{v}_i$ denotes the value of the human error. If there is no significant human error, then $H_i = 0$ and the label is correct. Based on IRT, a theory of standardized test design developed in psychology [16], we assume that a labeler with skill $\alpha \in (-\infty, \infty)$ has probability

$$\mathbb{P}[H_i = 0] = \frac{1}{1 + e^{\beta_i - \alpha}} \tag{3}$$

where $\beta_i \in (-\infty, \infty)$ denotes the difficulty level of the labeling task. Large β_i means that the labeling task is difficult, while large α means that the labeler is skillful. If $\alpha = \beta_i$, there is a probability of 0.5 that an error will be made. Note that the distribution of $\tilde{v}_i$ is unknown. Similar models with slightly different parameters have been used in [17] and [18] to study the image labeling errors in crowd-sourcing experiments.

The cause of human error is complex, and may depend on noise, workload, labeler skill, difficulty level of the labeling task, etc. We consider noise only and assume $\beta_i = |\epsilon_i|$. Obviously, in the model (1), the noise is the major factor describing the difficulty level of the labeling task.

III. Active Learning for Linear Regression

A. Correlation Matching and Sparse Outlier Suppression

We select the training data iteratively in our scheme. Consider the jth iteration, in which we need to select a new data vector $\mathbf{x}_{t_j}$ to label. Note that in the previous iterations we have already labeled $j-1$ training data and formed the training dataset $(\mathbf{X}_{j-1}, \mathbf{y}_{j-1})$, where $\mathbf{X}_{j-1} = [\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_{j-1}}]'$ and $\mathbf{y}_{j-1} = [y_{t_1}, \ldots, y_{t_{j-1}}]'$.

The new data $\mathbf{x}_{t_j}$ is to be selected from the data pool $\mathcal{X}_j = \{\mathbf{x}_i | 1 \le i \le I,\ \mathbf{x}_i \ne \mathbf{x}_{t_\ell},\ 1 \le \ell \le j-1\}$. We can simply evaluate each of the $I - j + 1$ data vectors in $\mathcal{X}_j$ and choose

$$\mathbf{x}_{t_j} = \arg \min_{\mathbf{z} \in \mathcal{X}_j} \left\| \frac{1}{j} \left(\mathbf{X}'_{j-1} \mathbf{X}_{j-1} + \mathbf{z}\mathbf{z}' \right) - \mathbf{R} \right\|^2 \tag{4}$$

where $\mathbf{R} = \frac{1}{I} \sum_{i=1}^{I} \mathbf{x}_i \mathbf{x}_i'$ is the sample correlation matrix of the overall data pool. The objective of the optimization (4) is to rapidly match the correlation of the training dataset with that of the overall data pool.

With the selected $\mathbf{x}_{t_j}$, we acquire its label y_{t_j} and append $(\mathbf{x}_{t_j}, y_{t_j})$ into the set $(\mathbf{X}_{j-1}, \mathbf{y}_{j-1})$ to get $(\mathbf{X}_j, \mathbf{y}_j)$. From (2) we have

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\theta} + \boldsymbol{\epsilon}_j + \mathbf{v}_j \tag{5}$$

where $\boldsymbol{\epsilon}_j = [\epsilon_{t_1}, \ldots, \epsilon_{t_j}]'$ and $\mathbf{v}_j = [v_{t_1}, \ldots, v_{t_j}]'$ are noise and outlier (labeling error) vectors, respectively.

If $\mathbf{v}_j$ is sparse, we can use the compressive sensing theory to estimate $\boldsymbol{\theta}$ and $\mathbf{v}_j$ jointly as

$$\{\hat{\boldsymbol{\theta}}(j), \hat{\mathbf{v}}_j\} = \arg \min_{\{\boldsymbol{\theta}, \mathbf{v}\}} \|\mathbf{y}_j - \mathbf{v} - \mathbf{X}_j \boldsymbol{\theta}\|, \quad s.t., \ \|\mathbf{v}\|_0 \le \|\mathbf{v}_j\|_0 \tag{6}$$

where $\|\cdot\|_0$ is ℓ_0 norm and $\|\mathbf{v}_j\|_0$ is the sparsity.

A common practice of compressive sensing is to replace the ℓ_0 norm with the convex ℓ_1 norm and solve

$$\{\hat{\boldsymbol{\theta}}(j), \hat{\mathbf{v}}_j\} = \arg \min_{\{\boldsymbol{\theta}, \mathbf{v}\}} \|\mathbf{y}_j - \mathbf{v} - \mathbf{X}_j \boldsymbol{\theta}\| + \lambda_1 \|\mathbf{v}\|_1 \tag{7}$$

where the weighting coefficient λ_1 is adjusted to match the sparsity $\|\mathbf{v}_j\|_0$. Following [13], we can find the solution

$$\hat{\boldsymbol{\theta}}(j) = \left(\mathbf{X}'_j \mathbf{X}_j \right)^{-1} \mathbf{X}'_j (\mathbf{y}_j - \hat{\mathbf{v}}_j) \tag{8}$$

where the outlier vector can be found as

$$\begin{aligned} \hat{\mathbf{v}}_j &= \arg \min_{\mathbf{v}} \|\mathbf{y}_j - \mathbf{v} - \mathbf{X}_j \hat{\boldsymbol{\theta}}(j)\| + \lambda_1 \|\mathbf{v}\|_1 \\ &= \arg \min_{\mathbf{v}} \left\| \left(\mathbf{I}_j - \mathbf{X}_j (\mathbf{X}'_j \mathbf{X}_j)^{-1} \mathbf{X}'_j \right) (\mathbf{y}_j - \mathbf{v}) \right\| + \lambda_1 \|\mathbf{v}\|_1 . \end{aligned} \tag{9}$$

Note that $\mathbf{I}_j$ is a $j \times j$ identity matrix. Pseudoinverse will be used when $\mathbf{X}'_j \mathbf{X}_j$ is rank deficient or when $j < N$.

Proposition 1. Assume the outlier sparsity $\|\mathbf{v}_j\|_0 < (j - N)/2$. Consider the optimization $\hat{\mathbf{v}}_j = \arg\min_{\mathbf{v}} \|(\mathbf{I}_j - \mathbf{X}_j (\mathbf{X}'_j \mathbf{X}_j)^{-1} \mathbf{X}'_j)(\mathbf{y}_j - \mathbf{v})\|$, s.t., $\|\mathbf{v}\|_0 < (j - N)/2$. The optimal solution satisfies $\lim_{\sigma_\epsilon \to 0} \hat{\mathbf{v}}_j = \mathbf{v}_j$.

Proof. Let $\mathbf{y}_j - \mathbf{v} = \mathbf{X}_j \boldsymbol{\theta} + \boldsymbol{\epsilon}_j + \Delta\mathbf{v}$ where $\Delta\mathbf{v} = \mathbf{v}_j - \mathbf{v}$ is the residue error of the outlier subtraction. Since the matrix $\mathbf{L} = \mathbf{I}_j - \mathbf{X}_j (\mathbf{X}'_j \mathbf{X}_j)^{-1} \mathbf{X}'_j$ is idempotent with rank $j - N$, we have $\mathbf{L}(\mathbf{y}_j - \mathbf{v}) = \mathbf{L}(\boldsymbol{\epsilon}_j + \Delta\mathbf{v})$. Consider the singular value decomposition $\mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{U}'$, where $\mathbf{D}$ is the diagonal singular value matrix with all the nonzero singular values only and $\mathbf{U}$ is the $j \times (j - N)$ singular vector matrix. The minimization problem $\min \|\mathbf{L}(\boldsymbol{\epsilon}_j + \Delta\mathbf{v})\|$ is equivalent to $\min \|\mathbf{D}\mathbf{U}'(\boldsymbol{\epsilon}_j + \Delta\mathbf{v})\|$. When $\sigma_\epsilon \to 0$, the optimization is reduced to $\min \|\mathbf{D}\mathbf{U}'\Delta\mathbf{v}\|$. If $\|\mathbf{v}_j\|_0 < (j - N)/2$ and $\|\mathbf{v}\|_0 < (j - N)/2$, then we have $\|\Delta\mathbf{v}\|_0 < j - N$. This leads to $\lim_{\sigma_\epsilon \to 0} \Delta\mathbf{v} = \mathbf{0}$, which means $\lim_{\sigma_\epsilon \to 0} \hat{\mathbf{v}}_j = \mathbf{v}_j$.

Proposition 1 shows that at high signal-to-noise ratio, all the outliers can be estimated and cancelled perfectly if the fraction of outliers $\|\mathbf{v}_j\|_0 / j$ in the training dataset satisfies

$$\frac{\|\mathbf{v}_j\|_0}{j} < \frac{1}{2} - \frac{N}{2j}. \tag{10}$$

This explains the important observation that the spectrum sensing method becomes ineffective when the fraction of outliers (a measure of sparsity) is over 0.5.

From the proof of Proposition 1, we can also see that the variance of the residue error $\Delta v_{t_\ell} = v_{t_\ell} - \hat{v}_{t_\ell}$, denoted as $\sigma^2_{\delta v}$, is comparable in size to noise variance σ^2_ϵ. Outliers as small as noise cannot be removed. Furthermore, the optimization (9) with ℓ_1 norm introduces such residue error to all the elements of $\hat{\mathbf{v}}_j$. The estimator (8) thus suffers from a noise-plus-residue error with variance $\sigma^2_\epsilon + \sigma^2_{\delta v}$. To mitigate this noise amplification effect, we can let

$$\hat{v}_{t_\ell} = 0, \quad \text{if } |\hat{v}_{t_\ell}| < a\sigma_\epsilon \tag{11}$$

for some constant a (e.g., $a = 2$), and replace (8) by the weighted least squares estimator [3]

$$\hat{\boldsymbol{\theta}}(j) = \left(\mathbf{X}'_j \mathbf{W}_j \mathbf{X}_j \right)^{-1} \mathbf{X}'_j \mathbf{W}_j (\mathbf{y}_j - \hat{\mathbf{v}}_j) \tag{12}$$

where $\mathbf{W}_j = \text{diag}\{w_{t_1}, \ldots, w_{t_j}\}$ with $w_{t_\ell} = 1$ if $\hat{v}_{t_\ell} = 0$ and $w_{t_\ell} = \sigma_\epsilon^2/(\sigma_\epsilon^2 + \sigma_{\delta v}^2)$ otherwise.

B. Training Data Screening for Nonsparse Human Errors

In active learning with a relatively small amount of training data, the constraint (10) can be easily violated by heavy human labeling errors. This makes the outlier vector $\mathbf{v}_j$ nonsparse and the compressive sensing method ineffective. To resolve this issue, we develop data screening techniques based on the human labeling error model (3) and $\beta_i = |\epsilon_i|$. Note that i equals t_ℓ since we work on the labeled training data only.

Because the noise ϵ_i is real and Gaussian, $|\epsilon_i|$ has folded normal distribution $f(|\epsilon_i|) = \frac{2}{\sigma_\epsilon\sqrt{2\pi}} e^{-\frac{|\epsilon_i|^2}{2\sigma_\epsilon^2}}$ with mean $\sigma_\epsilon\sqrt{2/\pi}$. The expected fraction of outliers is then

$$\begin{aligned}\overline{P} &= \int_0^\infty (1 - \mathbb{P}[H_i = 0]) f(|\epsilon_i|) d|\epsilon_i| \\ &= 1 - \int_0^\infty \frac{1}{1 + e^{|\epsilon_i| - \alpha}} \frac{2}{\sigma_\epsilon\sqrt{2\pi}} e^{-\frac{|\epsilon_i|^2}{2\sigma_\epsilon^2}} d|\epsilon_i|. \end{aligned} \quad (13)$$

Proposition 2. The fraction of outliers $\overline{P} \in [0, 1]$ and

$$\overline{P} \geq 1 - \frac{1}{1 + e^{\sqrt{2/\pi}\sigma_\epsilon - \alpha}}. \quad (14)$$

Proof. By Jensen's inequality, we have $\overline{P} = 1 - E[1/(1 + e^{|\epsilon_i| - \alpha})] \geq 1 - 1/(1 + e^{E[|\epsilon_i|] - \alpha})$, where $E[\cdot]$ denotes expectation. With the mean of $|\epsilon_i|$, we can obtain (14).

Numerical evaluation indicates that the two sides of (14) are in fact quite close to each other. The average of the outlier sparsity $\|\mathbf{v}_j\|_0$ is $j\overline{P}$, which can be estimated from σ_ϵ and α, and vice versa.

For each labeled training data $(\mathbf{x}_i, y_i)$, let $u_i = y_i - \mathbf{x}_i'\boldsymbol{\theta} = \epsilon_i + v_i$. Without outlier ($H_i = 0$), $u_i = \epsilon_i$ follows the noise distribution. With outlier ($H_i = 1$), $u_i = \epsilon_i + \tilde{v}_i$ and the distribution equals the convolution of the distributions of ϵ_i and $\tilde{v}_i$. Even though $\tilde{v}_i$ has an unknown distribution, the difference between the two cases can still be exploited for data screening.

Proposition 3. By keeping those labeled data that satisfy $|u_i| < \gamma$ as valid training data, where the threshold γ satisfies

$$\frac{\mathbb{P}[|\epsilon_i + \tilde{v}_i| < \gamma]}{\mathbb{P}[|\epsilon_i| < \gamma]} \leq \frac{\eta(1 - \overline{P})}{1 - \eta\overline{P}} \quad (15)$$

we can reduce the fraction of outliers in the training dataset from $\overline{P}$ to a level below $\eta\overline{P}$, where $0 < \eta < 1$.

Proof. We can reduce the fraction of outliers to a level below $\eta\overline{P}$ if $\mathbb{P}[H_i \neq 0 | |u_i| < \gamma] \leq \eta\overline{P}$. Because $\mathbb{P}[H_i \neq 0 | |u_i| < \gamma] = \mathbb{P}[|u_i| < \gamma | H_i \neq 0] \quad \mathbb{P}[H_i \neq 0]/\mathbb{P}[|u_i| < \gamma]$, and considering the fact that $\mathbb{P}[|u_i| < \gamma | H_i \neq 0] = \mathbb{P}[|\epsilon_i + \tilde{v}_i| < \gamma]$, we can derive $\mathbb{P}[|\epsilon_i + \tilde{v}_i| < \gamma]/\mathbb{P}[|u_i| < \gamma] \leq \eta$. Furthermore, utilizing the total probability formula $\mathbb{P}[|u_i| < \gamma] = \mathbb{P}[|\epsilon_i + \tilde{v}_i| < \gamma]\mathbb{P}[H_i \neq 0] + \mathbb{P}[|\epsilon_i| < \gamma]\mathbb{P}[H_i = 0]$, we can obtain (15) with straightforward deductions.

In practice, although the distribution of $\tilde{v}_i$ is hard to specify, its mean μ_v and variance σ_v^2 can be estimated from $\hat{\mathbf{v}}_{j-1}$ that is obtained in the previous iteration of active learning. Then $\epsilon_i + \tilde{v}_i$ has mean μ_v and variance $\sigma_\epsilon^2 + \sigma_v^2$. We can approximate $\epsilon_i + \tilde{v}_i$ with a Gaussian distribution to calculate γ. Such approximate γ is good enough for the purpose of data screening.

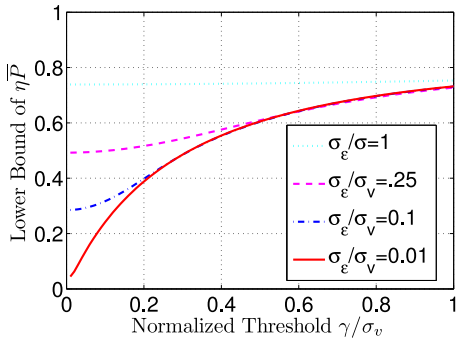

Fig. 1. Lower bound of $\eta\overline{P}$ versus normalized threshold γ/σ_v when $\overline{P} = 0.8$, under various noise levels.

From (15), we can see that the upper bound (i.e., $\eta\overline{P}$) of the new outlier fraction satisfies

$$\eta\overline{P} \geq \frac{h(\gamma)}{h(\gamma) - 1 + \overline{P}^{-1}} \quad (16)$$

where $h(\gamma) = \mathbb{P}[-\gamma \leq \epsilon_i + \tilde{v}_i \leq \gamma]/\mathbb{P}[-\gamma \leq \epsilon_i \leq \gamma]$. For example, assuming $\epsilon_i + \tilde{v}_i \sim \mathcal{N}(\mu_v, \sigma_\epsilon^2 + \sigma_v^2)$, the lower bound of $\eta\overline{P}$ under various threshold γ is illustrated in Fig. 1. The outlier fraction in the valid training dataset can be made sufficiently low if $\sigma_\epsilon/\sigma_v < 0.25$. On the other hand, γ should not be too small considering the constraint (10) on the number of valid training data j.

C. Active Regression Algorithm and Performance

Combining correlation matching, data screening, and outlier mitigation, we have the following new algorithm.

New Active Regression Algorithm

1) Input: Data pool $\{\mathbf{x}_i\}$, $\mathbf{R}$, λ_1, T
2) For iteration $j = 1, 2, \cdots, T$, do
   a) Select new training data $\mathbf{x}_{t_j}$ by (4) and get label y_{t_j}
   b) Update γ, u_{t_ℓ}, and screen the labeled data
   c) Estimate $\hat{\mathbf{v}}_j$ and $\hat{\boldsymbol{\theta}}(j)$ (9)(11)(12)

Output: $\hat{\boldsymbol{\theta}}(T)$.

In the jth iteration, for each labeled data $(\mathbf{x}_{t_\ell}, y_{t_\ell})$, $\ell = 1, \ldots, j$, we calculate u_{t_ℓ} by using $\hat{\boldsymbol{\theta}}(j-1)$ estimated in the previous iteration

$$u_{t_\ell} = y_{t_\ell} - \mathbf{x}_{t_\ell}'\hat{\boldsymbol{\theta}}(j-1), \quad \ell = 1, \ldots, j. \quad (17)$$

Similarly, we use $\hat{\mathbf{v}}_{j-1}$ to update γ based on (15). Then, with data screening, we use all the data $(\mathbf{x}_{t_\ell}, y_{t_\ell})$ that satisfy $|u_{t_\ell}| < \gamma$ as valid training data to estimate $\hat{\boldsymbol{\theta}}(j)$ and $\hat{\mathbf{v}}_j$.

Instead of using a fixed T, we can also use $\|\hat{\boldsymbol{\theta}}(j) - \hat{\boldsymbol{\theta}}(j-1)\|$ as the stopping criteria. In addition, in each iteration we need to run a convex optimization. To reduce the computational complexity, we can select $T_j > 1$ training data in each iteration to reduce the number of iterations to T_t, where $\sum_{j=1}^{T_t} T_j = T$.

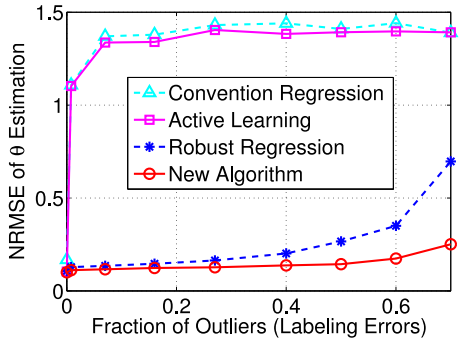

Fig. 2. NRMSE of the estimation of $\boldsymbol{\theta}$ over a toy dataset.

The coefficient λ_1 can be determined according to $\|\mathbf{y}_j - \mathbf{v}_j - \mathbf{X}_j\boldsymbol{\theta}\| \approx \lambda_1\|\mathbf{v}_j\|_1$, which gives

$$\lambda_1 \approx \frac{\sigma_\epsilon}{\eta\overline{P}\sqrt{j(\mu_v^2+\sigma_v^2)}}. \tag{18}$$

The objective of the robust regression is mainly to mitigate the detrimental effect of outliers. An interesting question is whether the outlier-contaminated data can benefit linear regression instead. If so, then the data screening procedure should keep as many training data as possible. With T training data, if all the outliers can be detected, the weighted least squares estimator (12) is unbiased and achieves the minimum error $J_T = \min\|\hat{\boldsymbol{\theta}}(T) - \boldsymbol{\theta}\|^2 = \sigma_\epsilon^2 \mathrm{tr}[(\mathbf{X}_T'\mathbf{W}_T\mathbf{X}_T)^{-1}]$, where tr denotes trace. Using the matrix inversion lemma, it can be shown that $J_T \le J_{T-1}$ for all $T > N$. Therefore, the estimation error decreases monotonically with T and converges to J_I. More training data are always better, even though some of them are contaminated. This also justifies our correlation matching principle since it pursues fast convergence to the global minimum $\sigma_\epsilon^2\mathrm{tr}(I\mathbf{R})^{-1}$.

IV. Simulations and Experiments

We compared our algorithm (*New Algorithm*) with three other algorithms: *Convention Regression* which implemented the least squares estimator without outlier suppression, *Active Learning* which implemented [3] without outlier suppression, and *Robust Regression* which implemented [13] without active learning.

First, we simulated a toy example with $I = 500$, $N = 10$, $T = 200$, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$, and $\epsilon_i \sim \mathcal{N}(0, 0.25)$. $\boldsymbol{\theta}$ was generated randomly and normalized. Outliers were generated with Laplacian distribution $v_i \sim \mathcal{L}(0, 10^3)$ and added according to (3). We evaluated the normalized root mean square error (NRMSE) $\sqrt{E[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2]/E[\|\boldsymbol{\theta}\|^2]}$. Simulation results are shown in Fig. 2. We can see that our new algorithm is more robust to outliers, even when the outlier fraction is high.

Next, we compared these algorithms over the benchmark dataset "redwine" of the UCI machine learning database [19]. The performance metric was the NRMSE of generalization, i.e., $\sqrt{E[|\hat{y}_i - y_i|^2]/E[|y_i|^2]}$. Simulation results are shown in Table I. Since outliers were rare or insignificant, outlier suppression did not play a major role. Instead, the correlation matching technique made our algorithm outperform the other algorithms. Significance test (T-test) was performed on the results, which verified that our algorithm was statistically better ($p < 0.05$) for short training datasets.

TABLE I
NRMSE of Generalization Over Benchmark Dataset

Training length T	50	150	250	350
Convention Regression	0.93	0.86	0.87	0.83
Active Learning	1.2	0.84	0.81	0.81
Robust Regression	0.93	0.86	0.87	0.83
New Algorithm	**0.90**	**0.84**	**0.81**	**0.81**

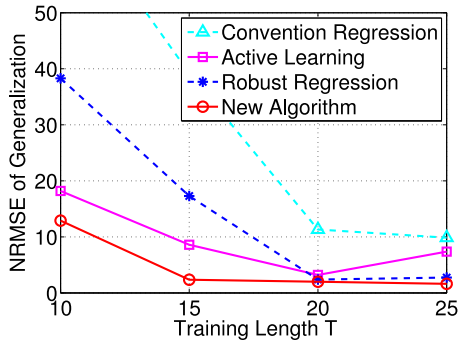

Fig. 3. NRMSE of generalization over experiment dataset.

Finally, we conducted real experiments to compare the performance of these algorithms in short training datasets with heavy outliers [20]. We did mock teacher evaluation surveys in a junior class where students were asked to score faked teachers (y_i) based on their grades in $N = 6$ different teaching evaluation questionnaires ($\mathbf{x}_i$) such as class preparation, instruction effectiveness, explaining complex ideas, etc. For five faked teacher prototypes, we collected $I = 162$ data records $(\mathbf{x}_i, y_i)$. From this dataset, we estimated $\sigma_\epsilon = \{0.09, 0.16, 0.13, 0.29, 0.32\}$ and $\overline{P} = \{0.07, 0.09, 0.13, 0.17, 0.12\}$ for the five prototypes. We can see that the fraction of outliers increases with the noise variance. With $\alpha = 2.2$, we found that (13) predicts $\overline{P}$ with a very small bias of 0.006 and standard deviation of 0.046. This verified the effectiveness of the IRT-based error model.

Then, we did another survey during the end of a class when students were eager to leave. This was actually the common situation when teaching surveys were really conducted. It led to heavy labeling errors. With just one faked teacher prototype, we collected 66 valid data records, from which we estimated $\overline{P} = 0.26$, $\alpha = 1.3$, $\sigma_v = 80$, and $\sigma_\epsilon/\sigma_v = 0.004$. According to Fig. 1, we set $\gamma = 0.2\sigma_v = 16$. We calculated the NRMSE of generalization, and the results in Fig. 3 clearly show that our algorithm outperformed all the other algorithms with the fastest convergence and the most superior outlier mitigation capability.

V. Conclusion

In this letter, we develop a new active regression algorithm that integrates correlation matching and compressive sensing for training optimization and human labeling error suppression. Based on the IRT, we model quantitatively human labeling errors and develop data screening techniques to mitigate nonsparse errors. Simulations and experiments were conducted to show its robustness to short training dataset and heavy labeling errors.

REFERENCES

[1] B. Settles, *Active Learning*. San Rafael, CA, USA: Morgan & Claypool, 2012.

[2] D. A. Chon, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artif. Intell. Res.*, vol. 4, pp. 129–145, 1996.

[3] M. Sugiyama and S. Nakajima, "Pool-based active learning in approximate linear regression," *Mach. Learn.*, vol. 75, no. 3, pp. 249–274, Jun. 2009.

[4] W. Cai, Y. Zhang, and J. Zhou, "Maximizing expected model change for active learning in regression," in *Proc. IEEE Int. Conf. Data Mining*, Dallas, TX, USA, Dec. 2013, pp. 51–60.

[5] S. Sabato and R. Munos, "Active regression by stratification," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, Canada, Dec. 2014, pp. 469–477.

[6] C. Riquelme, B. Zhang, and R. Johari, "Online active linear regression via thresholding," 2016.

[7] X. Li, Y. Chen, and K. Zeng, "Integration of machine learning and human learning for training optimization in robust linear regression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 2613–2617.

[8] X. Li and J. Zheng, "Joint machine learning and human learning design with sequential active learning and outlier detection for linear regression problems," in *Proc. IEEE Annu. Conf. Inform. Sci. Syst.*, Princeton, NJ, USA, Mar. 2016, pp. 407–411.

[9] R. Castro, R. Willet, and R. Nowak, "Faster rates in regression via active learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, Canada, Dec. 2005, pp. 179–186.

[10] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. Hoboken, NJ, USA: Wiley, 2005.

[11] J. J. Fuchs, "An inverse problem approach to robust regression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Phoenix, AZ, USA, Mar. 1999

[12] Y. Jin and B. D. Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 3830–3833.

[13] G. B. Giannakis, G. Mateos, S. Farahmand, V. Kekatos, and H. Zhu, "USPACOR: Universal sparsity-controlling outlier rejection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 2011, pp. 1952–1955.

[14] K. Mitra, A. Veeraraghavan, and R. Chellappa, "Analysis of sparse regularization based robust regression approaches," *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1249–1257, Mar. 2013.

[15] G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang, "Robust computation of linear models by convex relaxation," *Found. Comput. Math.*, vol. 15, no. 2, pp. 363–410, Apr. 2015.

[16] D. Thissen and L. Steinberg, "Item response theory," in *The SAGE Handbook of Quantitative Methods in Psychology*, R. E. Millsap and A. Maydeu-Olivares, Eds., Newbury Park, CA, USA: SAGE, 2009, pp. 148–177.

[17] J. Whitehill, T. Wu, J. Bergsma, J. Movellan, and P. Ruvolo, "Whose vote should count more: Optimal integration of labels form labelers of unknown expertise," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2009, pp. 2035–2043.

[18] P. Welinder, S. Branson, P. Perona, and S. Belongie, "The multidimensional wisdom of crowds," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2010, pp. 2424–2432.

[19] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decis. Support Syst.*, vol. 47, no. 4, pp. 547–553, 2009.

[20] Experiment data. (2016). [Online]. Available: http://www.ws.binghamton.edu/li/expdata.zip