ACTIVE REGRESSION WITH COMPRESSIVE-SENSING BASED OUTLIER MITIGATION FOR BOTH SMALL AND LARGE OUTLIERS

Jian Zheng and Xiaohua Li

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

ABSTRACT

In this paper, a new active learning scheme is proposed for linear regression problems with the objective of resolving the insufficient training data problem and the unreliable training data labeling problem. A pool-based active regression technique is applied to select the optimal training data to label from the overall data pool. Then, compressive sensing is exploited to remove labeling errors if the errors are sparse and have large enough magnitudes, which are called large outliers. Next, in order to mitigate the non-sparse labeling errors that have relatively small magnitudes, which are called small outliers, a new technique is developed to convert them back into sparse large outliers. With both artificial and real data sets, extensive simulations are conducted to verify the robustness of the proposed scheme in training data selection and outlier suppression.

***Index Terms*—** Robust linear regression, outlier mitigation, compressive sensing, active learning

1. INTRODUCTION

The objective of active regression is to minimize the amount of training data used in regression problems 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 regression 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 significant bottleneck.

Many effective active regression 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] [8] under a concept of integrated human and machine learning.

It has been shown that active regression can outperform the conventional passive regression, as seen in [5] [6] [9]. Nevertheless, active regression usually suffers from a severe performance fluctuation in practice because it tends to select the most ambiguous and noisy data for training, which unfortunately causes not only high regression errors but also heavy labeling errors. Modeled as outliers, if the labeling errors are sparse and have large magnitudes, they can be suppressed by robust regression techniques, including both conventional techniques such as RANSAC (Random Sample Consensus) [10] and new techniques such as compressive sensing [11]- [15].

While large and sparse outliers can be suppressed by robust regression techniques, non-sparse outliers with relatively small magnitudes, which are called small outliers, are another big issue that has not been studied sufficiently. Small outliers comparable to noise can not be mitigated with existing robust regression techniques. Conventionally, they are treated just as noise. However, such an approximation violates the noise assumption, and may fundamentally limit the regression/prediction performance. Detecting and mitigating such small outliers is highly useful for further improving regression performance.

Small and non-sparse outliers are common in practical applications. For example, human labelers usually have small bias or skews when labeling the training data. As another example, in today's Internet of Things (IoT), data are collected from thousands of sensors since IoT is all about data indeed [16]. However, the data may not be reliable enough because sensors may be faulty, or may not be calibrated accurately.

In this paper, we develop an active regression scheme that has the capability of removing both large sparse outliers and small non-sparse outliers. We apply a pool-based active regression technique for reliable training data selection, and employ compressive sensing technique for robust outlier mitigation.

The remainder of this paper is organized as follows. The active regression model is described in Section 2. In Section 3, new outlier mitigation techniques are developed. Simulations and conclusions are given in Sections 4 and 5, respectively.

2. ACTIVE REGRESSION MODEL

We consider the general linear regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\theta} + \epsilon_i + h_i v_i + o_i, \tag{1}$$

where y_i is the data label, $\mathbf{x}_i$ is the $N \times 1$ dimensional data vector, $\boldsymbol{\theta}$ is the $N \times 1$ dimensional regression vector, v_i is the small outlier applied through the scalar factor h_i, o_i is the large outlier, and ϵ_i is the noise with zero-mean and variance σ_ϵ^2.

We have included both large outliers and small outliers in (1). Large outliers have magnitudes much larger than the noise standard deviation σ_ϵ, i.e., $|o_i| \gg \sigma_\epsilon$. They are assumed sparse, or the probability of $o_i \neq 0$ is relatively small. In contrast, small outliers have magnitudes comparable to the noise standard deviation, i.e., $|h_i v_i| \approx \sigma_\epsilon$ or $|h_i v_i| < \sigma_\epsilon$. Small outliers are not assumed sparse.

This work is supported by the National Science Foundation under Grant CNS-1443885.

As a linear regression problem, we need to estimate $\boldsymbol{\theta}$ by training. If $\mathbf{x}_i$ is selected as training data, a labeler generates the label y_i, during which there are possibly large outlier o_i and small outlier $h_i v_i$. From the pool of I data vectors, we select T of them to construct a training data set $(\mathbf{X}_{tr}, \mathbf{y}_{tr})$, where $\mathbf{X}_{tr} = [\mathbf{x}_1, \cdots, \mathbf{x}_T]'$ and $\mathbf{y}_{tr} = [y_1, \cdots, y_T]'$, and use it to estimate $\boldsymbol{\theta}$.

In this paper, we use the pool-based sequential active learning technique of [3] to select the training data. We select the best T input data vectors $\mathbf{x}_i$ iteratively out of the overall I input data vectors with the objective of minimizing the expected prediction error. To be more specific, the input data $\mathbf{X}_{tr}$ would be selected from the pool of I input data with the probability proportional to

$$P_a(\mathbf{X}) = \left(\sum_{i,j=1}^{N} [\hat{\mathbf{U}}^{-1}]_{i,j} \mathbf{X}_i' \mathbf{X}_j \right)^a, \tag{2}$$

where $\mathbf{X}_i$ and $\mathbf{X}_j$ are the ith and jth columns of the overall $I \times N$ input data matrix $\mathbf{X}$, respectively. $\hat{\mathbf{U}}$ is an $N \times N$ matrix, with the (i,j)th element $\hat{U}_{i,j} = E[\mathbf{X}_i' \mathbf{X}_j]$. The parameter a is applied to adjust the tail of the distribution. The optimal training data set $\mathbf{X}_{tr}$ is selected with the best value of a that minimizes the objective function

$$J = \mathrm{tr}\left(\hat{\mathbf{U}} \mathbf{L} \mathbf{L}' \right), \tag{3}$$

where $\mathbf{L} = \left(\mathbf{X}_{tr}' \mathbf{W} \mathbf{X}_{tr} \right)^{-1} \mathbf{X}_{tr}' \mathbf{W}$, $\mathbf{W}$ is a $T \times T$ weighting matrix, and tr stands for the matrix trace operation.

After selecting the training data $\mathbf{X}_{tr}$, the labeled vector $\mathbf{y}_{tr}$ is obtained, and we have

$$\mathbf{y}_{tr} = \mathbf{X}_{tr} \boldsymbol{\theta} + \boldsymbol{\epsilon} + \mathbf{H}\mathbf{v} + \mathbf{o} \tag{4}$$

according to (1). Note that $\boldsymbol{\epsilon}$, $\mathbf{v}$ and $\mathbf{o}$ are derived from the stacking of ϵ_i, v_i, and o_i into vectors, respectively. The matrix $\mathbf{H}$ is resulted from h_i. We need to estimate and suppress the possible outliers $\mathbf{o}$ and $\mathbf{v}$ before estimating $\boldsymbol{\theta}$

3. OUTLIER SUPPRESSION FOR BOTH SMALL AND LARGE OUTLIERS

3.1. Large outlier suppression

The large outlier o_i has much larger magnitude than both the noise and the small outliers. In conventional schemes, the small outliers are treated as noise, and we can use compressive sensing to estimate $\boldsymbol{\theta}$ and $\mathbf{o}$ through the following joint optimization

$$\{\hat{\boldsymbol{\theta}}, \hat{\mathbf{o}}\} = \arg \min_{\{\boldsymbol{\theta}, \mathbf{o}\}} \|\mathbf{y}_{tr} - \mathbf{o} - \mathbf{X}_{tr}\boldsymbol{\theta}\| + \lambda_0 \|\mathbf{o}\|_0. \tag{5}$$

The weighting coefficient λ_0 is adjusted to match the sparsity $\|\mathbf{o}\|_0$, where $\|\cdot\|_0$ denotes ℓ_0 norm.

A common practice of compressive sensing is to replace the ℓ_0 norm with the convex ℓ_1 norm, which changes (5) to

$$\{\hat{\boldsymbol{\theta}}, \hat{\mathbf{o}}\} = \arg \min_{\{\boldsymbol{\theta}, \mathbf{o}\}} \|\mathbf{y}_{tr} - \mathbf{o} - \mathbf{X}_{tr}\boldsymbol{\theta}\| + \lambda_1 \|\mathbf{o}\|_1. \tag{6}$$

Following [13], we can find the solution as

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{X}_{tr}' \mathbf{X}_{tr} \right)^{-1} \mathbf{X}_{tr}' (\mathbf{y}_{tr} - \hat{\mathbf{o}}), \tag{7}$$

where the outlier vector can be obtained from

$$\begin{aligned} \hat{\mathbf{o}} &= \arg \min_{\mathbf{o}} \|\mathbf{y}_{tr} - \mathbf{o} - \mathbf{X}_{tr}\hat{\boldsymbol{\theta}}\| + \lambda_1 \|\mathbf{o}\|_1 \\ &= \arg \min_{\mathbf{o}} \left\| \left(\mathbf{I} - \mathbf{X}_{tr} (\mathbf{X}_{tr}' \mathbf{X}_{tr})^{-1} \mathbf{X}_{tr}' \right) (\mathbf{y}_{tr} - \mathbf{o}) \right\| + \lambda_1 \|\mathbf{o}\|_1, \end{aligned} \tag{8}$$

where $\mathbf{I}$ is an identity matrix.

Proposition 1. Assume $T > N$. If $\sigma_\epsilon^2 \to 0$, $h_i v_i \to 0$ and the outlier sparsity satisfies $\|\mathbf{o}\|_0 < (T-N)/2$, the optimization $\hat{\mathbf{o}} = \arg \min_{\mathbf{o}} \left\| \left(\mathbf{I} - \mathbf{X}_{tr} (\mathbf{X}_{tr}' \mathbf{X}_{tr})^{-1} \mathbf{X}_{tr}' \right) (\mathbf{y}_{tr} - \mathbf{o}) \right\|$ has a unique solution $\hat{\mathbf{o}} = \mathbf{o}$.

Proof. Let $\mathbf{y}_{tr} - \hat{\mathbf{o}} = \mathbf{X}_{tr}\boldsymbol{\theta} + \boldsymbol{\epsilon} + \mathbf{H}\mathbf{v} + \Delta\mathbf{o}$ where $\Delta\mathbf{o} = \mathbf{o} - \hat{\mathbf{o}}$ is the residue error of the outlier subtraction.

Since the matrix $\mathbf{Q} = \mathbf{I} - \mathbf{X}_{tr} (\mathbf{X}_{tr}' \mathbf{X}_{tr})^{-1} \mathbf{X}_{tr}'$ is idempotent with rank $T - N$, we have $\mathbf{Q}(\mathbf{y}_{tr} - \hat{\mathbf{o}}) = \mathbf{Q}(\boldsymbol{\epsilon} + \mathbf{H}\mathbf{v} + \Delta\mathbf{o})$. Consider the singular value decomposition $\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{U}'$, where $\mathbf{D}$ is the diagonal singular value matrix with all the non-zero singular values only and $\mathbf{U}$ is the $T \times (T-N)$ singular vector matrix. The minimization problem $\min \|\mathbf{Q}(\boldsymbol{\epsilon} + \mathbf{H}\mathbf{v} + \Delta\mathbf{o})\|$ is equivalent to $\min \|\mathbf{D}\mathbf{U}'(\boldsymbol{\epsilon} + \mathbf{H}\mathbf{v} + \Delta\mathbf{o})\|$. When $\sigma_\epsilon^2 \to 0$ and $h_i v_i \to 0$, the optimization is reduced to $\min \|\mathbf{D}\mathbf{U}'\Delta\mathbf{o}\|$. Then, if $\|\mathbf{o}\|_0 < (T-N)/2$, we can set $\|\hat{\mathbf{o}}\|_0 < (T-N)/2$ for the optimization. Therefore, $\|\Delta\mathbf{o}\|_0 < T - N$. This leads to the unique solution $\Delta\mathbf{o} = \mathbf{0}$ and thus $\hat{\mathbf{o}} = \mathbf{o}$. ■

Proposition 1 shows that when the noise is small enough, all the large outliers can be estimated and mitigated perfectly if the fraction of outliers $\|\mathbf{o}\|_0/T$ in the training data set satisfies

$$\frac{\|\mathbf{o}\|_0}{T} < \frac{1}{2} - \frac{N}{2T}. \tag{9}$$

This accounts for the important observation that the compressive sensing method becomes ineffective when the fraction of outliers (a measure of sparsity) is over 0.5. In addition, the number of training data T has to be sufficiently larger than N, which is the dimension of the problem, as $T > N + 2\|\mathbf{o}\|_0$.

From the proof we can also see that the variance of the residue error $\Delta o_i = o_i - \hat{o}_i$, denoted as $\sigma_{\delta o}^2$, is comparable in size to noise plus small outlier variance. Outliers as small as or smaller than noise can not be removed.

Therefore, Proposition 1 explains that we need new techniques to address small and non-sparse outliers. In the sequel, we first present small outlier models and then develop small outlier mitigation techniques.

3.2. Small outlier model

Assume that the T training data are labeled by L labelers and each labeler labels $T_L = T/L$ data. Without loss of generality, we assume that the ℓth labeler labels the training data set $(\mathbf{X}_\ell, \mathbf{y}_\ell)$, where $\mathbf{X}_\ell = [\mathbf{x}_{(\ell-1)T_L+1}, \cdots, \mathbf{x}_{\ell T_L}]'$ and $\mathbf{y}_\ell = [y_{(\ell-1)T_L+1}, \cdots, y_{\ell T_L}]'$. In addition, each of the ℓth labeler has a common outlier value v_ℓ, which is added to the labeling values via the weighting vector

$$\mathbf{h}_\ell = [h_{(\ell-1)T_L+1}, \cdots, h_{\ell T_L}]', \quad \ell = 1, \cdots, L. \tag{10}$$

We assume that $\|\mathbf{h}_\ell\| = 1$ and $|v_\ell| \gg \sigma_\epsilon^2$ if $v_\ell \neq 0$. For example, if the ℓth labeler labeled $T_L = 10$ data, and had suffered from small outliers $h_{(\ell-1)T_L+k} v_\ell = \frac{1}{10} v_\ell$, which is as large as the noise standard deviation σ_ϵ and is too small to be detected by conventional outlier mitigation algorithms. In this case, we can also see that v_ℓ is 10 times as large as σ_ϵ. Therefore, if we combine all the 10 small outliers together, we are able to detect v_ℓ.

Based on these assumptions, we consider two outlier models in this paper. For the small outlier model 1, we assume that the weighting vector $\mathbf{h}_\ell$ of each labeler ℓ is unknown, but all the labelers have the same weighting vector, i.e., $\mathbf{h}_\ell = \mathbf{h} = [h_1, \cdots, h_{T_L}]'$. For example, in certain survey experiments, it is likely that all the users have the same gradually increased bias while labeling (answering)

more questions. These users have the same weighting vector $\mathbf{h}$ but different bias values v_ℓ.

For the small outlier model 2, the weighting vectors $\mathbf{h}_\ell$ are different for different labelers, but the vectors are assumed known a priori. For example, the characteristics of the small outliers can be analyzed and determined beforehand.

In the first model, we can estimate $\mathbf{h}$ from the training data set after removing the large outliers o_i firstly with (7) and (8). By collecting all the L users' labeled data $\mathbf{y}_\ell$, where $\ell = 1, \cdots, L$, we can estimate the common weighting vector $\mathbf{h}$ by solving the following maximization

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}} E\left[\|\mathbf{h}'(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell)\|^2\right], \quad \text{s.t.,} \ \|\mathbf{h}\| = 1 \tag{11}$$

where $\hat{\mathbf{o}}_\ell = [\hat{o}_{(\ell-1)T_L+1}, \cdots, \hat{o}_{\ell T_L}]'$

We can find the correlation matrix

$$\mathbf{R}_y = E\left\{(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell)(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell)'\right\} \tag{12}$$

approximately using the sample average $\frac{1}{L}\sum_{\ell=1}^{L}(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell)(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell)'$. The optimization (11) is then

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}} \ \mathbf{h}'\mathbf{R}_y\mathbf{h}, \quad \text{s.t.,} \ \|\mathbf{h}\| = 1. \tag{13}$$

The solution to the optimization (13) is the eigenvector of $\mathbf{R}_y$ corresponding to its maximum eigenvalue.

3.3. Small outlier mitigation

For the small outlier model 1, with the estimated weighting vector $\hat{\mathbf{h}}$, we can construct L new labeled training data

$$z_\ell = \hat{\mathbf{h}}'(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell) = \hat{\mathbf{h}}'\mathbf{X}_\ell\boldsymbol{\theta} + \hat{\mathbf{h}}'\boldsymbol{\epsilon}_\ell + \hat{\mathbf{h}}'\mathbf{h}v_\ell. \tag{14}$$

We can see that

$$\begin{aligned} &E[\|\hat{\mathbf{h}}'(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell)\|^2] \\ &\quad = \hat{\mathbf{h}}' E\left[(\mathbf{X}_\ell\boldsymbol{\theta} + \boldsymbol{\epsilon}_\ell + \mathbf{h}v_\ell)(\mathbf{X}_\ell\boldsymbol{\theta} + \boldsymbol{\epsilon}_\ell + \mathbf{h}v_\ell)'\right]\hat{\mathbf{h}} \\ &\quad = \hat{\mathbf{h}}'\left(E[\mathbf{X}_\ell\boldsymbol{\theta}\boldsymbol{\theta}'\mathbf{X}_\ell']\right)\hat{\mathbf{h}} + \sigma_\epsilon^2\hat{\mathbf{h}}'\hat{\mathbf{h}} \ + \hat{\mathbf{h}}'\mathbf{h}v_\ell^2\mathbf{h}'\hat{\mathbf{h}} \\ &\quad = \hat{\mathbf{h}}'\left(E[\mathbf{X}_\ell\boldsymbol{\theta}\boldsymbol{\theta}'\mathbf{X}_\ell']\right)\hat{\mathbf{h}} + \sigma_\epsilon^2 + v_\ell^2. \end{aligned} \tag{15}$$

In (15), the noise power σ_ϵ^2 stays unchanged, while the outlier power is enhanced from $|h_{(\ell-1)T_L+k}v_\ell|^2$ to $|v_\ell|^2$. A gain of T_L is achieved to boost small outliers. This makes it possible to detect the small outliers which are not detectible with conventional robust regression algorithms. The larger T_L is, the more reliable the small outlier detection will be.

For the small outlier model 2, since the weighting vectors $\mathbf{h}_\ell$ are assumed known, the new labeled data is calculated directly as

$$z_\ell = \mathbf{h}_\ell'(\mathbf{y}_\ell - \hat{\mathbf{o}}_\ell), \quad \ell = 1, \cdots, L. \tag{16}$$

In this way, we obtain L new training data $(\mathbf{h}_\ell'\mathbf{X}_\ell, z_\ell)$, $\ell = 1, \cdots, L$. Appending these new training data to the original training data set, we have $T + L$ training data in total. In these training data, we will have less than L large outliers contained in the data z_ℓ with the magnitude of v_ℓ, which guarantees the sparsity of the large outliers.

Define the new $T + L$ training data set as $(\tilde{\mathbf{X}}, \tilde{\mathbf{y}})$, where $\tilde{\mathbf{X}} = [\mathbf{X}_{tr}', \mathbf{X}_1'\mathbf{h}_1, \cdots, \mathbf{X}_L'\mathbf{h}_L]'$, $\tilde{\mathbf{y}} = [\mathbf{y}_{tr}', z_1, \cdots, z_L]'$. We have

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\theta} + \tilde{\boldsymbol{\epsilon}} + \tilde{\mathbf{v}}, \tag{17}$$

where

$$\begin{aligned} \tilde{\boldsymbol{\epsilon}} &= [\boldsymbol{\epsilon}', \mathbf{h}_1'\boldsymbol{\epsilon}_1, \cdots, \mathbf{h}_L'\boldsymbol{\epsilon}_L]', \\ \tilde{\mathbf{v}} &= [(\mathbf{H}\mathbf{v})', v_1, \cdots, v_L]'. \end{aligned} \tag{18}$$

Since the new outliers in (17) are large enough in magnitude and sparse, they can be detected by the conventional outlier detection algorithms. Note that there are at most L new large outliers among the $T + L$ overall training data.

Therefore, based on (17), we can use the compressive sensing method again to estimate $\boldsymbol{\theta}$ and $\tilde{\mathbf{v}}$ jointly as

$$\{\hat{\boldsymbol{\theta}}, \hat{\mathbf{v}}\} = \arg\min_{\{\boldsymbol{\theta}, \tilde{\mathbf{v}}\}} \|\tilde{\mathbf{y}} - \tilde{\mathbf{v}} - \tilde{\mathbf{X}}\boldsymbol{\theta}\| + \lambda_1\|\tilde{\mathbf{v}}\|_1. \tag{19}$$

Similar to (7)(8), the solution to the joint optimization of (19) is

$$\hat{\boldsymbol{\theta}} = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'(\tilde{\mathbf{y}} - \tilde{\mathbf{v}}), \tag{20}$$

and

$$\hat{\mathbf{v}} = \arg\min_{\tilde{\mathbf{v}}} \left\|\left(\mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\right)(\tilde{\mathbf{y}} - \tilde{\mathbf{v}})\right\| + \lambda_1\|\tilde{\mathbf{v}}\|_1. \tag{21}$$

Note that $\mathbf{I}$ is a $(T + L) \times (T + L)$ identity matrix.

The key point is that the large outliers have been subtracted from $\tilde{\mathbf{y}}$ via $\hat{\mathbf{o}}$, and the small outliers are removed via $\tilde{\mathbf{v}}$ in (20). Therefore, the estimation of the regression vector $\boldsymbol{\theta}$ in (20) is more accurate.

3.4. New active regression scheme

In summary, the algorithm for the new proposed scheme with joint small and large outlier mitigation and pool-based active regression is given below.

New Robust Regression Algorithm
i) Input: Data pool $\{\mathbf{x}_i, y_i, i = 1, 2, \cdots, I\}$, λ_1, T, T_L
ii) Pool-based active learning: Select T training data out of the data pool with (2) and (3);
iii) Large outlier mitigation: Estimate and remove $\hat{\mathbf{o}}$ with (7) and (8);
iv) Small outlier mitigation:
1) Construct new training data with (14) and (16), and form the $T + L$ new training data $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$;
2) Estimate $\hat{\mathbf{v}}$ and $\hat{\boldsymbol{\theta}}$ with (20) and (21);
v) Output: $\hat{\boldsymbol{\theta}}$ for test data prediction.

In this algorithm, we apply convex optimization to estimate and mitigate both the large outliers and the small outliers. The large outlier vector $\mathbf{o}$ is sparse. The non-sparse small outlier vector $\mathbf{v}$ is converted into the sparse large outlier vector $\tilde{\mathbf{v}}$ through the outlier reconstruction technique in (14) and (16).

4. SIMULATIONS

In order to verify the performance of the proposed scheme for the small and large outlier mitigation in pool-based active regression (**SLOM+PB**), extensive simulations with an artificial data set, a UCI benchmark data set and a survey data set were conducted. We compared the new algorithm with the following algorithms: **Conv. R** which implemented only the conventional regression; **LOM** which applied simply large outlier mitigation using compressive

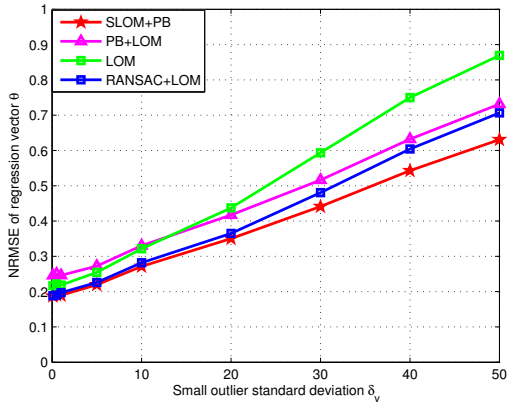

Fig. 1. Regressor estimation performance with the small outlier model 1.

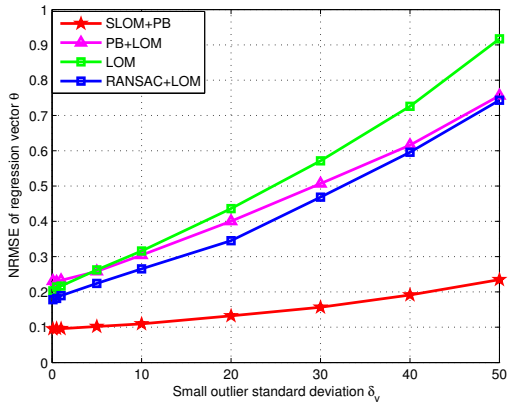

Fig. 2. Regressor estimation performance with the small outlier model 2.

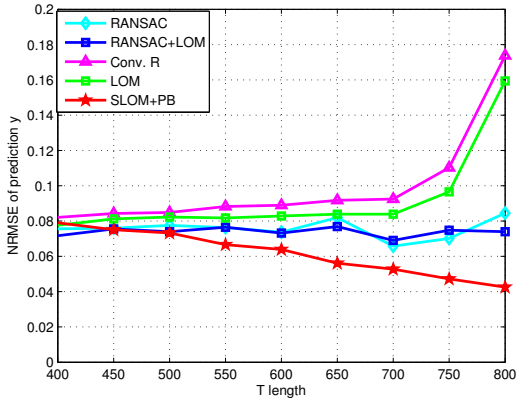

Fig. 3. Prediction performance with the small outlier model 1 in the Air Quality data set.

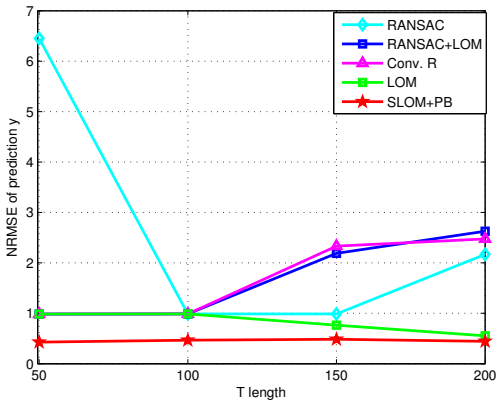

Fig. 4. Prediction performance with the small outlier model 1 in the survey data.

sensing [13]; **RANSAC** which employed Random Sample Consensus [17]; **PB** which implemented [3]; **RANSAC+LOM** which combined RANSAC with LOM; and **PB+LOM** which integrated PB with LOM.

Firstly, we used the artificially generated data for simulation. We let $I = 400$, $N = 10$, $T = 200$, $T_L = 20$, $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}_N, \mathbf{I}_N)$, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}_N, \mathbf{I}_N)$, and $\epsilon_i \sim \mathcal{N}(0, 0.25)$. We modeled large outliers with Laplacian distribution $o_i \sim \mathcal{L}(0, 10^3)$, and small outliers with $v_i \sim \mathcal{L}(0, \delta_v)$. As for the small outlier weighting vectors, in model 1 the vectors followed the Gaussian distribution $\boldsymbol{h} \sim \mathcal{N}(\mathbf{0}_{T_L}, \mathbf{I}_{T_L})$. For model 2, since the weighting vectors for different labelers are different, the set of weighting vectors followed the Gaussian distribution $\mathbf{h}_\ell \sim \mathcal{N}(\mathbf{0}_{T_L}, (0.25 + 0.05a)\mathbf{I}_{T_L})$, where $a \sim \mathcal{N}(\mathbf{0}_L, \mathbf{I}_L)$. We evaluated NRMSE (normalized root mean square error) of the regression vector $\boldsymbol{\theta}$ estimation $\sqrt{E[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2/\|\boldsymbol{\theta}\|^2]}$ with 100 runs of experiments for each small outlier standard deviation δ_v. The simulation results in Fig. 1 clearly show that our new algorithm outperforms the other algorithms with the small outlier model 1. In Fig. 2 with the small outlier model 2, our new algorithm shows even better performance.

Next, we simulated the algorithms by applying the small outlier model 1 to the UCI benchmark data set of Air Quality [18] as well as a mock teacher evaluation survey data set designed and conducted by us. We compared the NRMSE of the prediction of y_i, i.e., $\sqrt{E[|\hat{y}_i - y_i|^2/|y_i|^2]}$ of the five different algorithms. Fig. 3 and Fig. 4 both show that our new algorithm is more robust to outliers.

5. CONCLUSIONS

In this paper, a new robust regression scheme has been developed which integrates active learning with compressive sensing to make the data labeling in linear regression problems more robust to both sparse large outliers and non-sparse small outliers. Non-sparse small outliers were converted to sparse large outliers in order to use the compressive sensing method for outlier mitigation. The robustness of the new algorithm was verified by extensive simulations with artificial data, UCI benchmark data, as well as real survey data.

6. REFERENCES

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

[2] D. A. Chon, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129-145, 1996.

[3] M. Sugiyama and S. Nakajima, "Pool-based active learning in approximate linear regression," *Machine Learning*, 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," *IEEE Int. Conf. Data Mining* (ICDM'13), Dallas, TX, pp. 51-60, Dec. 2013.

[5] S. Sabato and R. Munos, "Active regression by stratification," *Proc. Advances in Neural Information Processing Systems* (NIPS'14), Montreal, Canada, Dec. 2014.

[6] C. Riquelme, B. Zhang, and R. Johari, "Online active linear regression via thresholding," *arXiv preprint arXiv:1602.02845*, 2016.

[7] X. Li, Y. Chen, and K. Zeng, "Integration of machine learning and human learning for training optimization in robust linear regression," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP'16), Shanghai, China, Mar. 2016.

[8] X. Li and J. Zheng, "Joint machine learning and human learning design with sequential active learning and outlier detection for linear regression problems," *Proc. IEEE CISS'2016*, Princeton University, Mar. 2016.

[9] 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.

[10] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, 2005.

[11] J. J. Fuchs, "An inverse problem approach to robust regression," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP'99), Phoenix, AZ, Mar. 1999

[12] Y. Jin and B. D. Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP'10), Dallas, TX, Mar. 2010.

[13] G. B. Giannakis, G. Mateos, S. Farahmand, V. Kekatos, and H. Zhu, "USPACOR: Universal sparsity-controlling outlier rejection," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP'11), Prague, May 2011.

[14] K. Mitra, A. Veeraraghavan, and R. Chellappa, "Analysis of sparse regularization based robust regression approaches," *IEEE Trans. Signal Processing*, 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," *Foundations of Computational Mathematics*, vol. 15, no. 2, pp. 363-410, Apr. 2015.

[16] D Miorandi, S Sicari, F De Pellegrini, "Internet of things: Vision, applications and research challenges," *Ad Hoc Networks*, vol. 10, no. 7, pp.1497-1516, 2012.

[17] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, *Comms. of the ACM*, pp. 381-395, 1981.

[18] S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sensors and Actuators B: Chemical*, Vol. 129, Issue 2, pp. 750-757, 2008.