

# Integrating Informativeness, Representativeness and Diversity in Pool-Based Sequential Active Learning for Regression

Ziang Liu, Dongrui Wu

Ministry of Education Key Laboratory of Image Processing and Intelligent Control,  
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China  
Email: {ziangliu, drwu}@hust.edu.cn

**Abstract**—In many real-world machine learning applications, unlabeled samples are easy to obtain, but it is expensive and/or time-consuming to label them. Active learning is a common approach for reducing this data labeling effort. It optimally selects the best few samples to label, so that a better machine learning model can be trained from the same number of labeled samples. This paper considers active learning for regression (ALR) problems. Three essential criteria – informativeness, representativeness, and diversity – have been proposed for ALR. However, very few approaches in the literature have considered all three of them simultaneously. We propose three new ALR approaches, with different strategies for integrating the three criteria. Extensive experiments on 12 datasets in various domains demonstrated their effectiveness.

**Index Terms**—Active learning, ridge regression, greedy sampling

## I. INTRODUCTION

In many real-world machine learning applications, unlabeled samples are easy to obtain, but it is expensive and/or time-consuming to label them. For example, in emotion estimation from speech signals, speech samples are easy to obtain; however, labeling them requires multiple human experts to listen carefully to score the emotion primitives, e.g., 6-17 experts were used in the VAM corpus [1], and at least 110 experts in IADS-2 [2].

Active learning (AL) [3] is an effective approach for reducing this data labeling effort. It optimally selects the best few samples to label, so that a better machine learning model can be trained from the same number of labeled samples, compared with random sampling. However, mainly of them considered only classification problems [4]–[9].

This paper considers only regression problems, i.e., active learning for regression (ALR). There are two scenarios in ALR: *population-based* and *pool-based* [10]. This paper considers the latter, where a pool of unlabeled samples is given, and we need to optimally select few of them to label, so that a regression model trained from them can be used to label other samples. The samples are selected iteratively, and only one sample is selected in each iteration. This situation is common in off-line machine learning applications [11].

This research was supported by Hubei Technology Innovation Platform under Grant 2019AEA171 and National Natural Science Foundation of China under Grant 61873321.

Three essential criteria for ALR have been proposed in our previous work [11]:

- 1) *Informativeness*, which means that the selected samples should contain rich information, so labeling them would contribute more to the regression function.
- 2) *Representativeness*, which means the sample density around each selected sample should be high, so that each selected sample can represent more neighbouring samples and less likely to be an outlier.
- 3) *Diversity*, which means the selected samples should scatter across the entire input space, instead of concentrating on a small region of it.

We [11] have also demonstrated empirically that each individual criterion contributes to the overall ALR performance.

Traditional ALR approaches, e.g., query-by-committee (QBC) [3], [7], [12]–[15] and expected model change maximization (EMCM) [16], [17], consider only the informativeness. We [11] proposed a representativeness-diversity (RD) approach to consider simultaneously the representativeness and the diversity. RD can also be integrated with QBC or EMCM to consider the informativeness, representativeness and diversity simultaneously. At the same time, [18] proposed an improve greedy sampling (iGS) approach to consider both the informativeness and diversity, and demonstrated that iGS outperformed QBC and EMCM.

However, no one has investigated if integrating RD and iGS can improve the performance of both, and also outperform RD-QBC and RD-EMCM. This paper fills this gap by proposing three new strategies to integrate RD and iGS. Our main contributions are:

- 1) We consider ALR in two separate processes: the initialization process and the iteration process. The initialization process selects few samples from the unlabeled sample pool to label and builds the initial regression model. The sample selection is completely unsupervised, thus only the representativeness and the diversity can be considered. An iteration process selects more samples to label, using the previously built regression model, and updates the regression model. The iteration process is supervised, and hence all three criteria can be considered.

- 2) We propose three new strategies to integrate RD and iGS, and hence the informativeness, representativeness and diversity are considered simultaneously.
- 3) We performed experiments on various datasets from diverse application domains to verify the effectiveness of our proposed ALR approaches. Particularly, two of them performed much better than the state-of-the-art ALR approaches.

The remainder of this paper is organized as follows: Section II introduces five state-of-the-art ALR approaches, which will be compared with our proposed ALR approaches. Section III proposes three new approaches for integrating the informativeness, representativeness, and diversity in ALR. Section IV compares the performances of our proposed ALR approaches with another six approaches. Finally, Section V draws conclusions.

## II. EXISTING ALR STRATEGIES

This section briefly introduces five state-of-the-art pool-based sequential ALR approaches. Each considers one or more of the three criteria. We assume the pool consists of  $N$  unlabeled samples  $\{\mathbf{x}_i\}_{i=1}^N$ , the initialization process selects  $M_0$  samples to label, and each iteration processes selects one more sample to label.

QBC [14], proposed in 1995, is probably the first ALR approach. It uses random sampling in the initialization process. In each sequential iteration, it first bootstraps the labeled samples and builds a committee of regression models. Then, for each unlabeled sample in the pool, it predicts its labels by the committee and computes the variance. Finally, it selects the sample with maximum variance to label. Clearly, QBC considers the informativeness only.

GSx (Greedy Sampling in the Input Space) [18], [19], first proposed in 2010, is a passive sampling approach that considers the diversity only. It selects the sample closest to the centroid of the whole sample pool as the first sample to label, and then in each sequential iteration a sample furthest away from all existing selected samples to label. The selected samples can be labeled at once after all of them are determined, because the selection process does not need the label information.

More specifically, let  $\{\mathbf{x}_m\}_{m=1}^M$  be the already selected samples. For each of the  $N - M$  unlabeled samples, GSx computes its closest distance to the  $M$  labeled samples, and then selects the sample  $\mathbf{x}^*$  with the maximum distance to label:

$$d_n^x = \min_m \|\mathbf{x}_n - \mathbf{x}_m\|, \quad m = 1, \dots, M; n = M + 1, \dots, N \quad (1)$$

$$\mathbf{x}^* = \arg \max_{\mathbf{x}_n} d_n^x, \quad (2)$$

EMCM [16], proposed in 2013, also considers the informativeness only. In the initialization process, it also randomly selects  $M_0$  samples to label, and train a regression model from them. In each sequential iteration, it selects the sample that may change the model parameters the most to label. More specifically, it uses all  $M$  already labeled samples to train a

regression model, whose prediction for the  $n$ th sample  $\mathbf{x}_n$  is  $\hat{y}_n$ . It also uses bootstrap on the  $M$  already labeled samples to construct  $P$  linear regression models. Let the  $p$ th model's prediction for the  $n$ th sample  $\mathbf{x}_n$  be  $y_n^p$ . Then, for each of the  $N - M$  unlabeled samples, it computes

$$g(\mathbf{x}_n) = \frac{1}{P} \sum_{p=1}^P \|(y_n^p - \hat{y}_n)\mathbf{x}_n\|, \quad n = M + 1, \dots, N \quad (3)$$

EMCM then selects the sample with the maximum  $g(\mathbf{x}_n)$  to label.

RD-EMCM [11], proposed in 2019, integrates RD and EMCM, and is the best-performing ALR in [11]. Its initialization process uses RD only, i.e., it performs  $k$ -means ( $k = M_0$ ) clustering of all  $N$  samples and selects the one closest to each cluster centroid to label. In each sequential iteration, it performs  $k$ -means clustering ( $k = M + 1$ , where  $M$  is the number of labeled samples) on the  $N$  samples, finds the largest cluster that does not contain an already labeled sample, and performs EMCM in that cluster to select one sample to label.

iGS [18], proposed in 2019, improves GSx by using greedy sampling in both the feature space and the label space. GSx is used in the initialization process. In each sequentially iteration, for all unlabeled samples, it first uses (1) to compute  $d_n^x$  in the input space, and then computes  $d_n^y$  in the label space:

$$d_n^y = \min_m \|\hat{y}_n - y_m\|, \quad m = 1, \dots, M; n = M + 1, \dots, N \quad (4)$$

It next computes  $d_n^{xy}$  to integrate  $d_n^x$  and  $d_n^y$ :

$$d_n^{xy} = d_n^x \cdot d_n^y, \quad n = M + 1, \dots, N \quad (5)$$

Finally, it selects the sample  $\mathbf{x}^*$  with the maximum  $d_n^{xy}$  to label:

$$\mathbf{x}^* = \arg \max_{\mathbf{x}_n} d_n^{xy} \quad (6)$$

Our previous work [18] has shown that iGS outperforms QBC, EMCM and GSx.

## III. THREE NEW STRATEGIES TO INTEGRATE RD AND IGS

In this section, we propose three new strategies to integrate RD and iGS, so that the informativeness, representativeness and diversity are considered simultaneously in ALR.

### A. RD-iGS

RD-iGS integrates RD and iGS, similar to the way of integrating RD and EMCM in [11]. Its initialization process uses RD only, i.e., it performs  $k$ -means ( $k = M_0$ ) clustering of all  $N$  samples and selects the one closest to each cluster centroid to label. In each sequential iteration, it performs  $k$ -means clustering ( $k = M + 1$ , where  $M$  is the number of labeled samples) on the  $N$  samples, finds the largest cluster that does not contain an already labeled sample, and performs iGS in that cluster to select one sample to label.

Comparing against the three criteria for ALR, RD considers the diversity and the representativeness, and iGS considers the diversity and the informativeness. Thus, RD-iGS considers

all three criteria. The pseudo-code of RD-iGS is shown in Algorithm 1.

---

**Algorithm 1:** The proposed RD-iGS algorithm.

---

**Input:** A pool of  $N$  unlabeled samples,  $\{\mathbf{x}_i\}_{i=1}^N$ ;  
 $M_{\max}$ , the maximum number of samples to label.

**Output:** The regression model  $f(\mathbf{x})$ .

// Initialization process  
 Perform  $k$ -means ( $k = M_0$ ) clustering on  $\{\mathbf{x}_i\}_{i=1}^N$ ;  
 Select from each cluster one sample closest to its centroid, query for its label, and add it to the training set  $T$ ;

Construct the initial regression model  $f(\mathbf{x})$ ;  
 // Iteration process  
**for**  $M = M_0 + 1, \dots, M_{\max}$  **do**  
   Perform  $k$ -means ( $k = M$ ) clustering on  $\{\mathbf{x}_i\}_{i=1}^N$ ;  
   Identify the largest cluster  $C_{\max}$  that does not already contain any labeled sample;  
   Compute  $d^{xy}$  in (5) for each sample in  $C_{\max}$  ;  
   Select the sample with the maximum  $d^{xy}$ , query for its label, and add it to the training set  $T$ ;  
   Update the regression model  $f(\mathbf{x})$  using  $T$ ;  
**end**

---

### B. RDIGS

In RDIGS, RD is used for initialization and iGS for iteration. It performs unsupervised RD in the initialization process, as in RD-iGS, and performs iGS in the iteration process.

The pseudo-code of RDIGS is shown in Algorithm 2. It also considers all three criteria for ALR.

---

**Algorithm 2:** The proposed RDIGS algorithm.

---

**Input:** A pool of  $N$  unlabeled samples,  $\{\mathbf{x}_i\}_{i=1}^N$ ;  
 $M_{\max}$ , the maximum number of samples to label.

**Output:** The regression model  $f(\mathbf{x})$ .

// Initialization process  
 Perform  $k$ -means ( $k = M_0$ ) clustering on  $\{\mathbf{x}_i\}_{i=1}^N$ ;  
 Select from each cluster one sample closest to its centroid, query for its label, and add it to the training set  $T$ ;

Construct the initial regression model  $f(\mathbf{x})$ ;  
 // Iteration process  
**for**  $M = M_0 + 1, \dots, M_{\max}$  **do**  
   Compute  $d^{xy}$  in (5) for each unlabeled sample  $\{\mathbf{x}_n\}_{n=M}^N$ ;  
   Select the sample with the maximum  $d^{xy}$ , query for its label, and add it to the training set  $T$ ;  
   Update the regression model  $f(\mathbf{x})$  using  $T$ ;  
**end**

---

### C. RDIGSr

RDIGSr performs RD in the initialization process, and iGSr in the iteration process. iGSr first uses (1) to compute  $d_n^x$  in the input space and (4) to compute  $d_n^y$  in the label space. It then uses (5) to compute  $d_n^{xy}$ , just as iGS. Finally, it computes the representativeness  $R_n$  for all unlabeled samples:

$$R_n = \sum_{i=1}^N \|\mathbf{x}_n - \mathbf{x}_i\| \quad (7)$$

and selects the one with the maximum  $d_n^{xy}/R_n$  to label:

$$\mathbf{x}^* = \arg \max_{\mathbf{x}_n} (d_n^{xy}/R_n), \quad n = M + 1, \dots, N \quad (8)$$

The pseudo-code of RDIGSr is shown in Algorithm 3. It also considers all three criteria for ALR.

---

**Algorithm 3:** The proposed RDIGSr algorithm.

---

**Input:** A pool of  $N$  unlabeled samples,  $\{\mathbf{x}_i\}_{i=1}^N$ ;  
 $M_{\max}$ , the maximum number of samples to label.

**Output:** The regression model  $f(\mathbf{x})$ .

// Initialization process  
 Perform  $k$ -means ( $k = M_0$ ) clustering on  $\{\mathbf{x}_i\}_{i=1}^N$ ;  
 Select from each cluster one sample closest to its centroid, query for its label, and add it to the training set  $T$ ;

Construct the initial regression model  $f(\mathbf{x})$ ;  
 // Iteration process  
**for**  $M = M_0 + 1, \dots, M_{\max}$  **do**  
   Compute  $d^{xy}$  in (5) for each unlabeled sample  $\{\mathbf{x}_n\}_{n=M}^N$ ;  
   Compute  $R$  in (7) for each unlabeled sample  $\{\mathbf{x}_n\}_{n=M}^N$ ;  
   Use (8) to select one sample, query for its label, and add it to the training set  $T$ ;  
   Update the regression model  $f(\mathbf{x})$  using  $T$ ;  
**end**

---

### D. Summary

Table I summarizes how the informativeness, representativeness and diversity are considered in different ALR approaches.

## IV. EXPERIMENTS AND RESULTS

Extensive experiments are performed in this section to demonstrate the advantages of our proposed three ALR approaches over five state-of-the-art ALR approaches.

### A. Datasets

Table II shows the summary of the 12 datasets used in our experiments. They cover a wide variety of application domains. More of them are from the UCI Machine Learning Repository<sup>1</sup> and the CMU StatLib Datasets Archive<sup>2</sup>, which

<sup>1</sup><http://archive.ics.uci.edu/ml/index.php>

<sup>2</sup><http://lib.stat.cmu.edu/datasets/>

TABLE I  
SUMMARY OF HOW THE INFORMATIVENESS (I), REPRESENTATIVENESS (R) AND DIVERSITY (D) ARE CONSIDERED IN DIFFERENT ALR APPROACHES. RS MEANS RANDOM SAMPLING.

	Initialization				Iteration			
	Alg	I	R	D	Alg	I	R	D
QBC	RS	-	-	-	QBC	✓	-	-
GSx	GSx	-	-	✓	GSx	-	-	✓
EMCM	RS	-	-	-	EMCM	✓	-	-
RD-EMCM	RD	-	✓	✓	RD-EMCM	✓	✓	✓
iGS	GSx	-	-	✓	iGS	✓	-	✓
RD-iGS	RD	-	✓	✓	RD-iGS	✓	✓	✓
RDIGS	RD	-	✓	✓	iGS	✓	-	✓
RDIGSr	RD	-	✓	✓	iGSr	✓	✓	✓

had also been used in many previous ALR experiments [11], [16]–[19]. We also used an affective computing dataset *Vera am Mittag* (VAM; *Vera at Noon* in English) [1]), which has been used in many previous studies [20]–[24]. Only arousal in VAM was used as the regression output.

Two datasets (autoMPG and CPS) contained both numerical and categorical features. For them, we first used one-hot coding to covert the categorical values into numerical values, which increased their dimensionality. We then used principal component analysis to reduce their dimensionality back to their original number of dimensions. For each dataset, we normalized each dimension of the input to mean zero and standard deviation one.

TABLE II  
SUMMARY OF THE 12 REGRESSION DATASETS.

Dataset	No. of samples	No. of raw features	No. of numerical features	No. of categorical features	No. of total features
Concrete-CS	103	7	7	0	7
Yacht	308	6	6	0	6
autoMPG	392	7	6	1	9
NO2	500	7	7	0	7
PM10	500	7	7	0	7
Housing	506	13	13	0	13
CPS	534	10	7	3	19
VAM-Arousal	947	46	46	0	46
Concrete	1,030	8	8	0	8
Airfoil	1,503	5	5	0	5
Wine-Red	1,599	11	11	0	11
Wine-White	4,898	11	11	0	11

### B. Performance Evaluation Process

For each dataset, we randomly selected 80% samples as the training pool, and the remaining 20% as the test set. We used the mean and the variance of the training samples to normalize the test samples, because in practice the test samples are unknown.

We compared the performances of our proposed RD-iGS, RDIGS and RDIGSr with six approaches: random sampling (RS), QBC, GSx, EMCM, RD-EMCM, and iGS. For each sampling approach, we selected  $M \in [d + 1, d + 21]$  samples,

where  $d$  is the feature dimensionality after data preprocessing<sup>3</sup>. We then labeled them and built a ridge regression model<sup>4</sup> with the L2 regularization coefficient  $r = 0.01$ . We used the root mean squared error (RMSE) and the correlation coefficient (CC) as performance measures. The above process was repeated 100 times to obtain statistically meaningful results.

### C. Results

Fig. 1 shows the performances of the nine approaches, each averaged across the 100 runs. RDIGSr achieved the smallest RMSE on most datasets and for most  $M$ .

To compare the performances of different algorithms more directly, we computed the area under the curves (AUCs) of the mean RMSEs and the mean CCs. Because the AUCs from different datasets varied a lot, we normalized them w.r.t. the AUC of RS on each dataset; thus, the AUC of RS was always 1 on each dataset. Fig. 2 shows the results, and Table III shows the improvements of the eight ALR approaches over the RS baseline. On average, all ALR approaches had smaller RMSE and larger CC than RS. The differences between the RMSEs of different ALR approaches were more obvious than the differences between their CCs. On average, our proposed RDIGSr had the smallest RMSE.

TABLE III  
PERCENTAGE IMPROVEMENTS OF THE AUCs OF THE MEAN RMSEs AND THE MEAN CCs OVER RS. THE BEST PERFORMANCES ARE MARKED IN BOLD.

		GSx	QBC	EMCM	RD-EMCM	iGS	RD-iGS	RDIGS	RDIGSr
RMSE	Mean	7.1	6.2	6.7	13.1	12.5	13.0	15.9	<b>17.0</b>
	Var	61.9	35.8	41.2	70.4	47.5	69.1	75.3	<b>75.4</b>
CC	Mean	6.4	5.5	3.9	<b>8.6</b>	6.6	8.4	3.3	7.0
	Var	39.5	11.7	11.1	39.2	15.6	<b>39.9</b>	30.0	34.5

To determine whether the performance improvements of our proposed three ALR approaches over the other two best performing ALR approaches (RD-EMCM and iGS) were statistically significant, we used two-sample one-tailed  $t$ -test (MATLAB2019 function *ttest2*) with the significance level  $\alpha = 0.05$  on the average AUCs between each pair of ALR approaches. The results are shown in Table IV. Both RDIGS and RDIGSr had significantly smaller RMSEs than RD-EMCM and iGS.

Comparing RDIGSr with RDIGS, RDIGSr had smaller RMSE ( $p = 0.0690$ , close to the threshold  $p = 0.05$ ) and larger CC ( $p = 0.0001$ ) than RDIGS, which demonstrated the advantages of iGSr (Section 3) over iGS (Section II), both of which are sequential sampling approaches used in the iteration process of ALR.

<sup>3</sup>A reliable linear regression model requires at least  $d + 1$  training samples, where  $d$  corresponds to the  $d$  coefficients of the  $d$  features, and 1 corresponds to the bias term.

<sup>4</sup>We used the function *firlinear* in MATLAB 2019, which performs better than the function *ridge* used in our previous work [11], [18], especially when  $M$  is small. The performance improvements of all ALR approaches decreased as the performance of RS improved a lot.

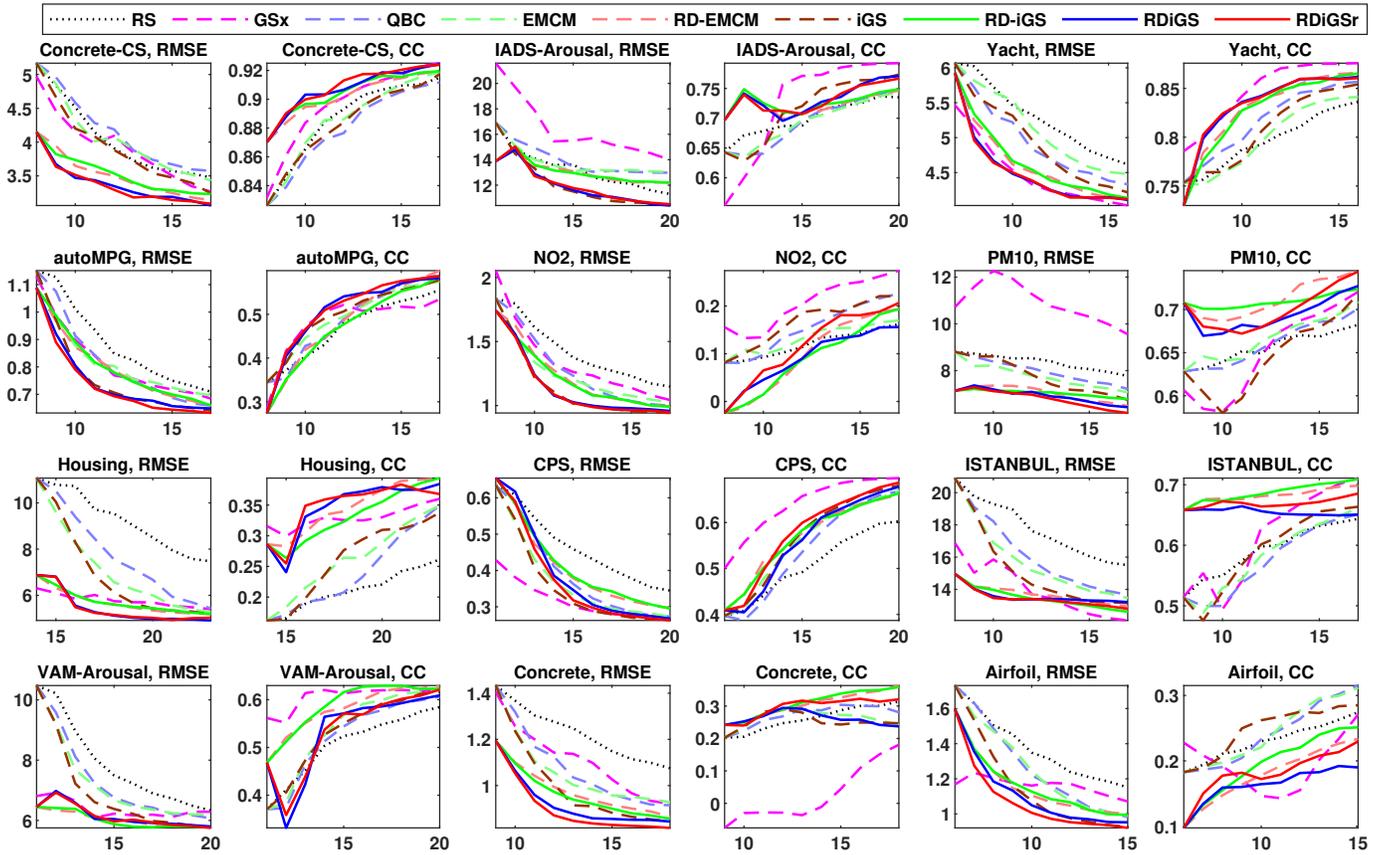


Fig. 1. Mean RMSEs and mean CCs of the nine sampling approaches on the 12 datasets, averaged over 100 runs. The horizontal axis represents  $M$ , the number of samples to be labeled.

TABLE IV

$p$ -VALUES OF THE  $t$ -TESTS. THE ALTERNATIVE HYPOTHESIS IS THAT THE ROW-WISE ALR APPROACH PERFORMED BETTER (HAD SMALLER RMSE OR LARGER CC) THAN THE COLUMN-WISE ALR APPROACH. A NUMBER IN BOLD MEANS THE ALTERNATIVE HYPOTHESIS IS ACCEPTED.

		RD-EMCM	iGS
RD-iGS	RMSE	.4904	<b>.0000</b>
	CC	.5839	<b>.0000</b>
RDIGS	RMSE	<b>.0000</b>	<b>.0000</b>
	CC	1.0000	.6713
RDIGSr	RMSE	<b>.0000</b>	<b>.0000</b>
	CC	.9707	<b>.0010</b>

In summary, all of our three proposed ALR approaches, RD-iGS, RDIGS and RDIGSr, performed better than or comparable with the state-of-the-art RD-EMCM and iGS approaches in the literature. Among the three, RDIGSr may be the best.

#### D. Sensitivity to the Regularization Coefficient

To determine whether our proposed approaches always perform better with different regularization coefficient, we repeated the experiments for different ridge regression models with different L2 regularization coefficient  $r = 0.001$  and  $r = 0.1$ . The AUCs of the eight ALR approaches on the

12 datasets are shown in Figs. 3 and 4, for  $r = 0.001$  and  $r = 0.1$ , respectively. On average, RDIGSr and RDIGS still had smaller RMSE than RD-EMCM and iGS, and RDIGSr still had smaller RMSE and larger CC than RDIGS, for both  $r = 0.001$  and  $r = 0.1$ . These results confirm again that RDIGSr was the best ALR approach among the eight.

## V. CONCLUSIONS

ALR is a frequently used machine learning approach for reducing the data labeling effort in regression problems. This paper considers pool-based sequential ALR, where a pool of unlabeled samples is given; we need to iteratively select some samples to label, from which a regression model can be trained to label other samples. Three essential criteria have to be considered in ALR: informativeness, representativeness, and diversity. We decompose ALR into two processes (initialization process and iteration process) and study how to incorporate these three criteria in each process separately. We proposed three new ALR approaches (RD-iGS, RDIGS and RDIGSr), corresponding to three different strategies for integrating the three criteria. RDIGS and RDIGSr performed much better than the state-of-the-art ALR approaches, and on average RDIGSr achieved the best performance.

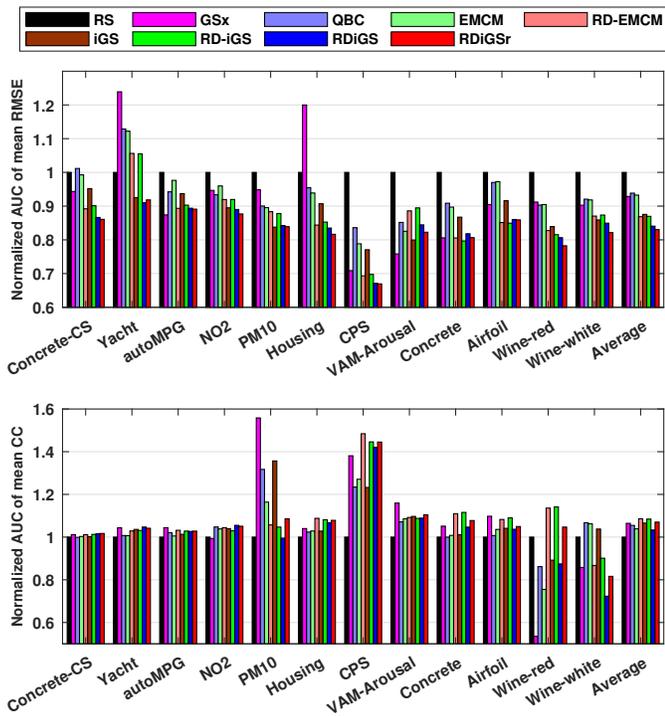


Fig. 2. Normalized AUCs of the mean RMSEs and the mean CCs on the 12 datasets. Ridge regression ( $r = 0.01$ ) was used as the regression model.

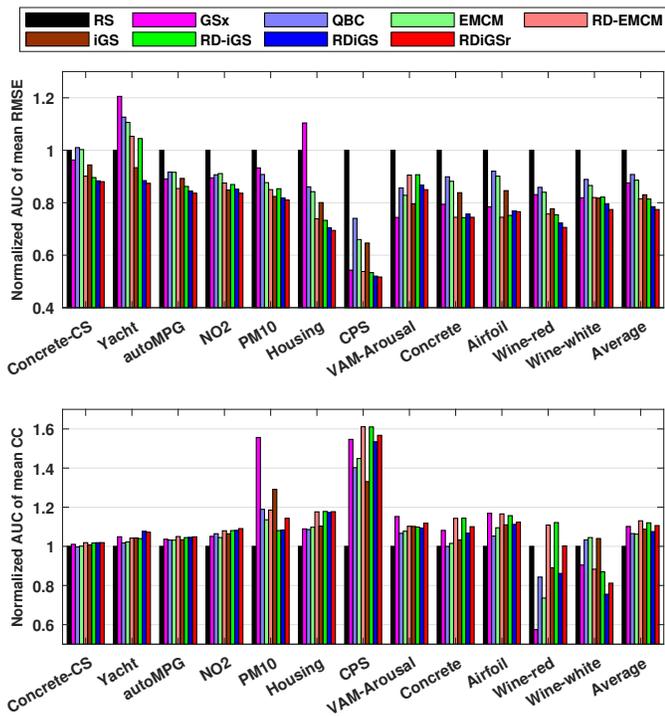


Fig. 3. Normalized AUCs of the mean RMSEs and the mean CCs on the 12 datasets. Ridge regression ( $r = 0.001$ ) was used as the regression model.

## REFERENCES

[1] M. Grimm, K. Kroschel, and S. S. Narayanan, "The Vera Am Mittag German audio-visual emotional speech database," in *Proc. Int'l Conf. on*

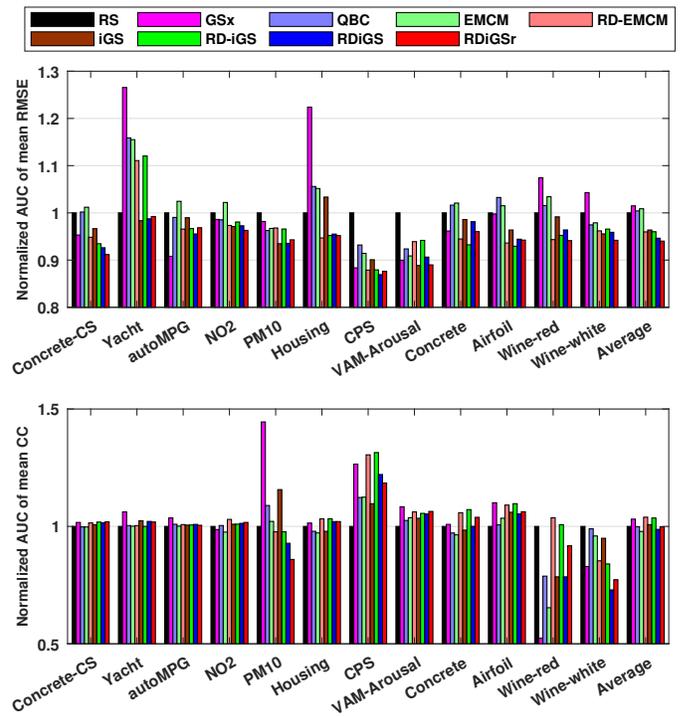


Fig. 4. Normalized AUCs of the mean RMSEs and the mean CCs on the 12 datasets. Ridge regression ( $r = 0.1$ ) was used as the regression model.

*Multimedia & Expo (ICME)*, Hannover, German, June 2008, pp. 865–868.

- [2] M. M. Bradley and P. J. Lang, "The international affective digitized sounds (2nd edition; IADS-2): Affective ratings of sounds and instruction manual," University of Florida, Gainesville, FL, Tech. Rep. B-3, 2007.
- [3] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.
- [4] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in *Proc. 15th Int'l Conf. on Machine Learning*, Madison, WI, July 1998, pp. 1–9.
- [5] W. Cai, Y. Zhang, S. Zhou, W. Wang, C. Ding, and X. Gu, "Active learning for support vector machines with maximum model change," *Lecture Notes in Computer Science*, vol. 8724, pp. 211–216, 2014.
- [6] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *Proc. 34th Int'l. Conf. on Machine Learning*, vol. 70, Sydney, Australia, Aug. 2017, pp. 1183–1192.
- [7] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Proc. Neural Information Processing Systems*, Denver, CO, November 1995, pp. 231–238.
- [8] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, December 2008, pp. 1289–1296.
- [9] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proc. Conf. on Empirical Methods in Natural Language Processing*, Honolulu, HI, October 2008, pp. 1069–1078.
- [10] M. Sugiyama and S. Nakajima, "Pool-based active learning in approximate linear regression," *Machine Learning*, vol. 75, no. 3, pp. 249–274, 2009.
- [11] D. Wu, "Pool-based sequential active learning for regression," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1348–1359, 2019.
- [12] R. Burbidge, J. J. Rowland, and R. D. King, "Active learning for regression based on query by committee," *Lecture Notes in Computer Science*, vol. 4881, pp. 209–218, 2007.
- [13] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with

- statistical models,” *Journal of Artificial Intelligence Research*, vol. 4, no. 1, pp. 129–145, 1996.
- [14] T. RayChaudhuri and L. G. Hamey, “Minimisation of data collection by active learning,” in *Proc. IEEE Int’l. Conf. on Neural Networks*, vol. 3, Perth, Australia, November 1995, pp. 1338–1341.
- [15] B. Demir and L. Bruzzone, “A multiple criteria active learning method for support vector regression,” *Pattern Recognition*, vol. 47, pp. 2558–2567, 2014.
- [16] W. Cai, Y. Zhang, and J. Zhou, “Maximizing expected model change for active learning in regression,” in *Proc. IEEE 13th Int’l. Conf. on Data Mining*, Dallas, TX, December 2013.
- [17] W. Cai, M. Zhang, and Y. Zhang, “Batch mode active learning for regression with expected model change,” *IEEE Trans. on Neural Networks and Learning Systems*, vol. 28, no. 7, pp. 1668–1681, July 2017.
- [18] D. Wu, C.-T. Lin, and J. Huang, “Active learning for regression using greedy sampling,” *Information Sciences*, vol. 474, pp. 90–105, 2019.
- [19] H. Yu and S. Kim, “Passive sampling for regression,” in *Proc. IEEE Int’l. Conf. on Data Mining*, Sydney, Australia, December 2010, pp. 1151–1156.
- [20] M. Grimm and K. Kroschel, “Emotion estimation in speech using a 3D emotion space concept,” in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. Vienna, Austria: I-Tech, 2007, pp. 281–300.
- [21] M. Grimm, K. Kroschel, E. Mower, and S. S. Narayanan, “Primitives-based evaluation and estimation of emotions in speech,” *Speech Communication*, vol. 49, pp. 787–800, 2007.
- [22] D. Wu, T. D. Parsons, E. Mower, and S. S. Narayanan, “Speech emotion estimation in 3D space,” in *Proc. IEEE Int’l Conf. on Multimedia & Expo (ICME)*, Singapore, July 2010, pp. 737–742.
- [23] D. Wu, T. D. Parsons, and S. S. Narayanan, “Acoustic feature analysis in speech emotion primitives estimation,” in *Proc. InterSpeech*, Makuhari, Japan, September 2010.
- [24] D. Wu and J. Huang, “Affect estimation in 3D space using multi-task active learning for regression,” *IEEE Trans. on Affective Computing*, 2020, in press.