Final Report

Hongwei Zhang
hongwei@vt.edu

Abstract

The prediction of the ratio of the particle velocity before and after rebound plays a vital role in the particle impact study. Using active learning method can circumvent the massive computational resources needed for the traditional numerical simulations. This project utilizes a pool-based active learning framework to build a regression model to predict the coefficient of restitution with the limited number of labeled data. We adapted a general active learning for regression framework from previous work for our dataset. By comparing the performance of different distance metrics and regression models, we show that having a certain degree of physical understanding towards the problem can greatly improve the model's performance.

1 Statement of the problem

1.1 Problem definition and dataset preparation

Among different particle rebound behaviors, the ratio between the rebound and impact velocities, namely coefficient of restitution (CoR), is vital in aero-engine corrosion. In this study, only normal impact is considered such that CoR can be treated as a scalar. The system illustration is shown in Figure. 1. As the sand particle is one of the most common collision particles, we choose the particle's mechanical properties based on quartz. Different from the particle, the substrate can be made of various materials whose properties vary in a wide range. Therefore, CoR as the description of particle rebound can be seen as a function of impact velocity and the mechanical properties of the impacted substrate.

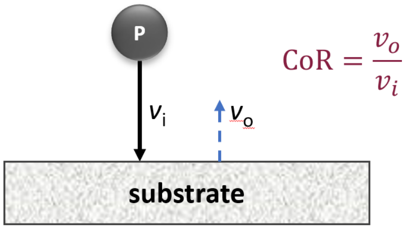


Figure 1: The illustration of the vertical particle rebound on substrate.

To obtain the CoR for certain condition, numerical simulation using LS-DYNA can be performed with high fidelity. However, such simulation takes hours for just one case. The computational cost for developing a reliable prediction model or getting a table for practical usage requires is prohibitively high. From the experimental side, setting the platform and collecting reliable data also has great cost and is limited by the budget. Given that both simulation and experiment take much longer time than training a machine learning model, it's affordable to train a new model whenever we get a new labeled data point.

Pool-based active learning is designed to get reasonable performance using as small data as possible by iteratively requesting a new input point measurement based on the performance of the model built upon the currently available measurements. By using the active learning, the computational or physical cost can be used in the most efficient way to get better model. Tn this study, we construct a dataset to exam this idea by checking the performance of the active learning method. Instead of using the real simulation of experimental data, we use the synthetic data from a greatly simplified model where CoR is only the function of a small number of parameters. $CoR = f(v_i, E, A, \rho)$.

$$CoR = \frac{v_o}{v_i} = \begin{cases} \left(\frac{v_i}{v_{y,w}}\right)^{-0.091} & 0 < \frac{v_i}{v_{y,w}} < 100, \\ 2.08\left(\frac{v_i}{v_{y,w}}\right)^{-0.25} & 100 \leq \frac{v_i}{v_{y,w}} \leq \frac{v^{\star}_{i,w}}{v_{y,w}}, \\ 0.78\left[\frac{v_i}{v_{y,w}} / \frac{E}{Y_w}\right]^{-0.5} & \frac{v^{\star}_{i,w}}{v_{y,w}} < \frac{v_i}{v_{y,w}} < 0.063\left(\frac{E}{Y_w}\right)^2 \end{cases}$$

$$Y_w = A \qquad v_{y,w} = 5.052\left(\frac{Y_w^5}{E^4 \rho_w}\right)^{0.5}$$

$$v^{\star}_{i,w} = 0.02\left(\frac{E}{Y_w}\right)^2$$

where E is Young's modulus of the substrate, A is the yield strength of the substrate, ρ_w is substrate density, and v_i is the particle impact velocity.

The simplified model is used for two reasons. First, using synthetic data could easily have a large number of data without consuming many resources. Despite the great simplification, different intervals corresponding to different input parameter ranges as the most important physical feature of this problem is preserved.[1] Second, though the correlations are pretended to be unknown during the training process, being able to access the ground truth function can greatly help us to elucidate the model performance trained under different algorithms.

Based on the ranges of impact velocity and substrate properties from previous work [1], we sample the combinations of input parameters randomly to form a pool with 4096 data points where 20% are used for testing purpose.

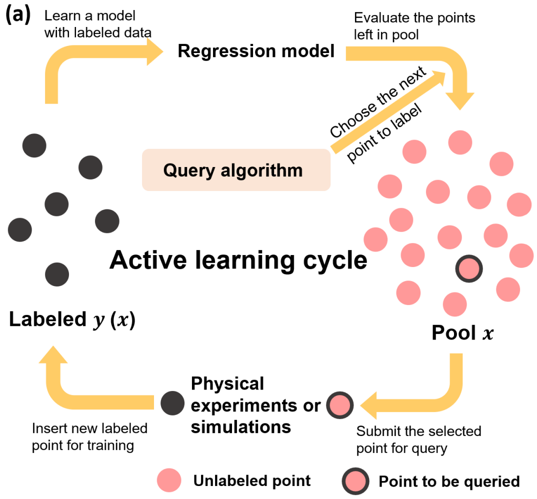
1.2 Pool-based active learning algorithm

The active learning framework is shown in Figure. 2a, where the query algorithm to select the point to label is the critical component for the learning cycle. Traditional algorithms like query-by-committee (QBC) find the point based on the largest variance from the results of all the members. However, limited by the cost to query a data point, we can only label one data point instead of a batch for one cycle, which makes QBC to lost its ability to consider diversity of the point in the input space.

For our specific problem, the query algorithm is adapted from a previous work as shown in Figure. 2b.[2] Their adapted query algorithm considers both the representativeness and diversity (RD) for even a single point by using the k-means clustering method to first identify the points located within the currently largest unlabeled input points cluster before applying the traditional query algorithm.

In order to verify the performance of the RD algorithm, we compare the performance of the random sampling without any algorithm as the baseline, RD without QBC method, and RD combined with QBC as a query algorithm. Here, the greed sampling method (GD), focusing more on the input point distribution, is introduced as a comparison to QBC, which focuses on the output.

As for the regression model, apart from the ridge regression as a widely used parametric model, we choose Gaussian process as a non-parametric model for comparison.


(a)
Learn a model with labeled data
Regression model
Evaluate the points left in pool
Choose the next point to label
Query algorithm
Active learning cycle
Labeled $y(x)$
Pool x
Physical experiments or simulations
Insert new labeled point for training
Submit the selected point for query
Unlabeled point
Point to be queried

(b)

```
Input: $N$ unlabeled samples, $\{\mathbf{x}_n\}_{n=1}^N$, where $\mathbf{x}_n \in \mathbb{R}^d$;
       $M$, the maximum number of labeled samples to query
Output: The regression model $f(\mathbf{x})$.
// Initialize $d$ labeled samples
Perform $k$-means clustering on $\{\mathbf{x}_n\}_{n=1}^N$, where $k = d$;
Select from each cluster the sample closest to its centroid, and query for its label;
// End initialization
for $m = d+1, \dots, M$ do
    Perform $k$-means clustering on $\{\mathbf{x}_n\}_{n=1}^N$, where $k = m$;
    Identify the largest cluster that does not already contain a labeled sample;
    Option 1: Select the sample closest to the cluster centroid for labeling;
    Option 2: Use QBC to select a sample from the cluster for labeling;
    Option 3: Use GS to select a sample from the cluster for labeling;
end
Construct the regression model $f(\mathbf{x})$ from the $M$ labeled samples.
```

Figure 2: (a) The illustration of the active learning cycle. (b) The specific algorithm of the adapted query algorithm.

2 Results and analysis

2.1 Parametric regression model

For active learning, each training cycle corresponds to a regression model trained with a certain amount of labeled data points. The performance of the regression model in that cycle is tested by the test data divided from the entire pool earlier. Here, we show the test error as a function of the number of labeled data points for different query algorithms where the regression model is set to be ridge regression. The relative mean squared root is selected here as the error metric.

$$RMSE = \left[\frac{1}{N} \sum_{n=1}^{N} (f_n - f_n')^2 \right]^{1/2} \tag{1}$$

where f_n and f_n' corresponds to the ground truth and predicted CoR. From Figure. 3a, the baseline

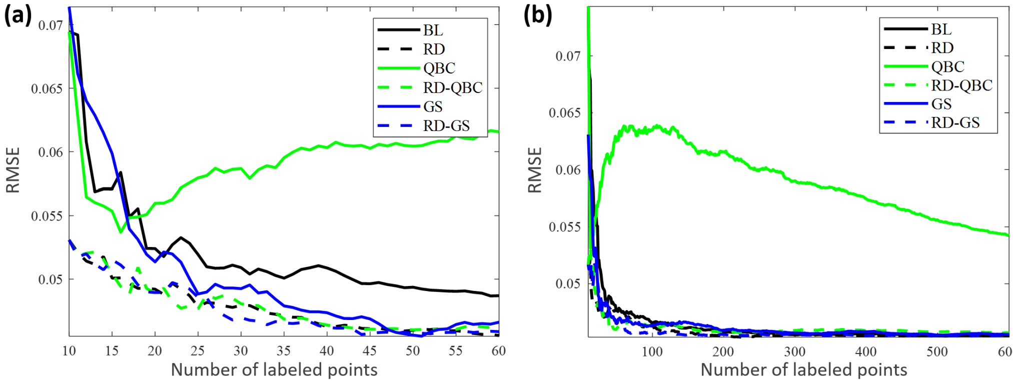


Figure 3: The testing error as function of the number of labeled points from (a) 10 to 60 and (b) 10 to 600.

method generally shows a large error compared to other methods, suggesting the advantage of active learning to achieve higher accuracy with limited data. It can also be found that the solid curves show collectively larger errors than those dashed curves. Given the solid curves correspond to no initialization of the points selected in the first cycle, such results show that proper initialization method could maximize the value of the collected point to give a more reliable model to start with.

In Figure. 3a, the most intriguing part is the curve for the QBC only method, where we see the error first decrease as the number of labeled points increase as expected, but the error later starts to increase again. Such unusual results seem to suggest the trained model become worse when we provide more data, which is counter-intuitive. As comparison, the GS shows no such trend, which suggests the problem comes from the way QBC used to select the next point to label. Also, the QBC combined with RD shows no such trend as well, such that we can further narrow the problem to be the QBC's selection range in the entire pool space.

With the advantage that the actual physical model in this study is fully accessible (considered to be unknown information during the entire active learning cycle), we now rationalize this result from both the function from input to output space and the pool composition. From the function shown earlier, the function consists of three intervals based on the size of $v_i/v_{y,w}$. We can roughly imagine the three intervals consist of a semi-convex shape curve as a function of $v_i/v_{y,w}$ where the first interval is steepest and the third interval is flattest. From the pool composition, we notice that the points lies in the first interval is only about 1% of the points in the entire pool based on the ranges of the parameters selected. Combining the above two pieces of information, the straight line as function of $v_i/v_{y,w}$ is dragged to be very low due to most of the data being drawn from the second and third intervals that are relatively flat.

As QBC starts to work, it will keep selecting the points that are located in the first interval as its curve is steepest, corresponding to great variance for a single point. As the number of points in the first interval increases, their weight increases so that the model is dragged higher towards the optimal location, and the error decreases. However, the model is soon dragged too high and starts to deviate from the optimal location resulting in increasing error. The RD-QBC method circumvented the problem of selecting too many points A by performing the QBC only within the largest cluster without labeled data points. When the model is dragged high enough, the largest variance points finally start to appear in the third and second intervals and drag the model towards the optimal location such that the error should slowly decrease again. Such a decrease can indeed be observed in Figure. 3b.

From Figure. 3b, we found the RMSE seems to converge to an asymptotic value at around 0.045. Considering the mean of the CoR in the dataset is around 0.27, such an RMSE is about 16%, suggesting the model is not reliable. Such a result shows the linear regression model is too simple to predict the physical relationship in this study. Although more complicated non-linear parametric model could be utilized to improve the performance, we noticed here that the choice of parametric model has strong dependence to the physical essences of the problem. For a complicated problem, choosing a proper parametric model is not easy. Therefore, we next try to use a non-parametric model to predict this problem with the same setting.

2.2 Non-parametric regression model

Now, the Gaussian process is used to replace the ridge regression to be the regression model. As a non-parametric model, Gaussian process takes the input points as the prediction and extend those points to their surrounding points through a covariance matrix. The results are shown in Figure. 4. Compared to the results with ridge regression, Gaussian process gives about three times smaller error with the same number of the labeled data points and shows no sign of convergence, as shown in Figure. 4a. This comparison shows the advantages of using a non-parametric model for a complicated physical problem.

In most physical problems, variables are often not independent of each other, and some could even be combined to form a dimensionless number. When the accessible data number is large, such an relationship could be obtained through the supervised clustering method. However, the accessible data in this study is limited. Therefore, to further improve the performance of the active learning model, we changed the distance metric of the k-means clustering from the squared Euclidean to

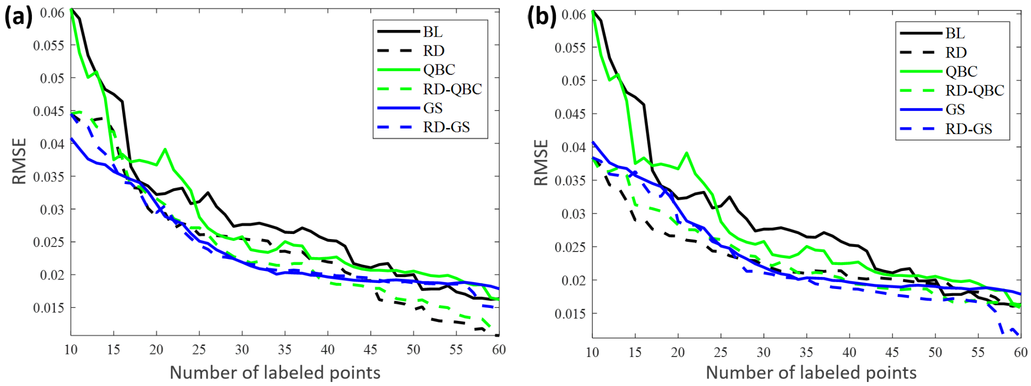


Figure 4: The testing error as function of the number of labeled points with (a) the squared Euclidean distance metric (b) the cosine distance metric.

cosine that could show the idea of the dimensionless number to a certain extent—comparing Figure. 4b to 4a, we indeed see improvement when the number of the labeled data points is small. By having a deeper physical understanding, we could directly combine the related variables in the input space to have better model performance ($v_i/v_{y,w}$ for this problem ideally).

Lastly, we add about 7% of random perturbation to the labels to mimic the error in experimental measurements. The results are shown in Figure. 5 manifesting only slightly higher error suggesting the model developed by the active learning method in this study is robust to measurement error.

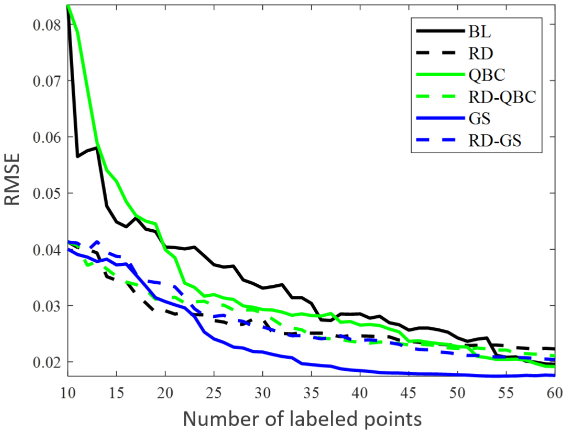


Figure 5: The testing error as function of the number of labeled points when perturbation is added.

3 Conclusion

The data labeling for CoR is expensive, making the prediction model construction with traditional methods difficult. This work used active learning for regression as the framework for labeling data selection, showing improvement compared to random sampling. By analyzing the results for different query functions, we found the query function should be chosen based on the physical features of the specific problem. Comparing the prediction accuracy from the ridge and Gaussian process regressions, we found the non-parametric model has the advantage in predicting more complicated physical problems when the number of labeled data is limited. Finally, by replacing the distance metric from the squared Euclidean to cosine, we illustrate the importance of physical understanding in limited data situations. This model is also shown to be robust to measurement error.

References

[1] Uzi, A., Levy, A. (2018). Energy absorption by the particle and the surface during impact. Wear, 404, 92-110.

[2] Wu, D. (2018). Pool-based sequential active learning for regression. IEEE transactions on neural networks and learning systems, 30(5), 1348-1359.