UniformLIME: A Uniformly Perturbed Local Interpretable Model-Agnostic Explanations Approach for Aerodynamics

Enshuo Jiang$^{1,*}$

$^{1}$School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

$^*$E-mail: enshuojiang@uestc.edu.cn

Abstract. Machine learning and deep learning are widely used in the field of aerodynamics. But most models are often seen as black boxes due to lack of interpretability. Local Interpretable Model-agnostic Explanations (LIME) is a popular method that uses a local surrogate model to explain a single instance of machine learning. Its main disadvantages are the instability of the explanations and low local fidelity. In this paper, we propose an original modification to LIME by employing a new perturbed sample generation method for aerodynamic tabular data in regression model, which makes the differences between perturbed samples and the input instance vary in a larger range. We make several comparisons with three subtasks and show that our proposed method results in better metrics.

1. Introduction

Aerodynamic data modeling refers to the approximation of the mapping function between input and output data using appropriate models.[1] With the continuous development of aerodynamics, a variety of complex and diverse aerodynamic data continue to emerge. In the face of data from many sources, how to conduct effective analysis and modeling from it has become a difficult problem for experts in this field. Using artificial intelligence or machine learning to understand and analyze these data plays an important role in reducing experimental costs, shortening design cycles, simulation and prototyping in the field of aerodynamics. Although machine learning outperforms humans in these meaningful tasks, its performance and application have also been questioned due to its lack of interpretability. No one can know exactly the basis for making decision and whether the decision is reliable, which makes it impossible for aerodynamics experts and machine learning experts to understand each other. Ordinary users hope that the results obtained from machine learning models should be understandable by humans, rather than appearing in the form of complex matrices or equations. The result should be in a visually understandable form to facilitate verification of correctness.

Local surrogate models refer to using interpretable models (such as decision trees, linear regression, logistic regression, etc.) to approximate the prediction of the underlying black box model, and then to explain the prediction of the black box model for a single instance.[2] Local interpretable model-agnostic explanations (LIME) is a well-known, popular, and specific implementation method of local surrogate models.[3] The interpretable model usually pursues the stability and local fidelity of the interpretation result. LIME is not doing well in this regard.[4] Most of the modifications to LIME focused on solving these disadvantages.

Anchors is another paper by the original author of LIME. [5]The main improvement direction is high precision and clear converge. Anchors highlight part of the input that is sufficient for the classifier to make a prediction to make it intuitive and easy to understand. Autoencoder Based
Approach for Local Interpretability (ALIME) [6] uses a pre-trained autoencoder to convert randomly generated disturbance samples into lower-dimensional latent codes, and calculates the distance between the latent codes to measure the similarity between the disturbed samples and the input instance to be explained, instead of directly calculating the Euclidean distance between them. Deterministic Local Interpretable Model-Agnostic Explanations Approach (DLIME) [7] uses agglomerative hierarchical clustering to cluster the training data and uses K-Nearest Neighbor to select the most relevant category of the instance to be explained, as an alternative to random perturbation. Train a linear model based on the selected category to generate interpretation results. But these modifications also have some limitations and shortages. Anchors may be complex and provide low coverage. DLIME might produce incorrect interpreted results. ALIME have a poor performance on complex regression problems.

In this paper, we propose UniformLIME, which is an improvement to LIME for aerodynamic tabular data in regression model. By an original perturbed sample generation method for aerodynamic tabular data as a modification to regression LIME, UniformLIME can improve both stability and local fidelity. Besides, we make several experiments on three different subtasks on an aerodynamic dataset to study the effects and improvements.

2. Materials and Methods

2.1. LIME

Because of the nature of locality, the first step of LIME is to choose an instance which you are interested to explain. Then, LIME will generate some perturbed samples by adding random disturbances to the input instance to be explained and use the underlying black box model to make predictions on these perturbed samples. A weighted linear regression is trained by taking the prediction result of black box model corresponding to perturbed samples as the true value and the similarity between disturbance samples and the input instance to be explained as the weight. The result of linear regression can reflect the contribution of each feature to the prediction result of a single instance. Simply speaking, LIME perturbs the input data instance and observes how the corresponding prediction results change. By this way, LIME try to understand the model by this trained simple interpretable model (linear regression) and learn which feature changes will have the greatest impact on the prediction results of a single specific instance.

2.2. UniformLIME

During the process of algorithm, LIME has three different types of data representation for tabular data. The original one is continuous tabular data. The second one is discretized data. In this paper, it is assumed that the discretizer is decile, which means the discretized data is a discrete integer value from 0 to 9. The third type is binary data which only contains zero or one. The instance to be explained will be represented as an all-one vector. The perturbed samples in discretized form will be compared with the discretized sample to be explained to obtain the binary form. For each dimension, if the discretized value is the same, set the binary value to 1, otherwise set it to 0. The distance to the instance to be explained and the weight in the linear regression model are calculated based on the third binary data.

The problem is the perturbed sample generation method. LIME would generate random integers from 0 to 9 with equal probability (uniformly) for each feature individually, compare with the discretized instance to be explained to obtain a 0-1 vector, and sample from the truncated normal distribution in the corresponding interval to obtain the real disturbance sample.
Because the probability is the same from 0 to 9 for random discretized sample, the probability of one is always 10% for each dimension in the corresponding binary vector. As a result, although the discretized sample is generated randomly and uniformly, the amount of one in binary data is limited into a relatively small range, which is the key parameter to calculate distances and weights. For instance, if the discretizer of LIME is decile and the input data have 60 dimensions, the mathematical expectation of the amount of one in binary data is 6 (60 times of 10%). In other words, the random perturbed samples have similar distances to the original input instance to be explained, which means that the algorithm can hardly distinguish whether the perturbed sample is the neighbor of the original input instance to be explained and get a reasonable explanation result.
To solve this problem, we propose an original perturbed sample generation method for aerodynamic tabular data as a modification to regression LIME. Instead of generating discretized samples firstly, we generate the binary data with fixed number of one at first. By changing the number of one in binary data, we can directly control the distance between perturbed sample and input local instance and get real random samples with different distances. Next, we will shuffle the binary data and convert it into discretized data by comparing with discretized form of the input local instance. For each dimension, if the binary value is one, set the discretized value to the same value as the input, otherwise set it to a random different value (uniformly sample), as long as the discretized value is not the same as the input. The final step is not modified. According to the discretized data, sample from the truncated normal distribution in the corresponding interval to obtain the real disturbance sample.

By this way, the distance and the weight of perturbed sample would distribute uniformly in the domain. The sample with close distance would be regarded as the neighbor of the input instance to be explained, while the sample with long distance would be tagged by a low weight in linear regression. According to the experimental results, this new method improves both local fidelity and stability.

![Discretized Data distribution](image1)

![Tabular Data distribution](image2)

**Fig. 4** UniformLIME: Data distribution histogram of decile discretized data(left) and continuous tabular data(right) for one dimension (5000 random samples)

![Distance Distribution](image3)

![Weight Distribution](image4)

**Fig. 5** UniformLIME: Data distribution histogram of distance(left) and weight(right) with 5000 random samples

3. Results and Discussion

3.1. Dataset

Our experiments are based on one dataset belonging to aerodynamic domain. This dataset consisting of 600 transformations of M6 airfoil. Each transformation has 60 design variables (features) as input...
parameters, which represents the relative height of airfoil surface located at the fixed grid, and 3 aerodynamic coefficients as output parameters, including lift coefficient, drag coefficient and torque coefficient. The position of 60 features in three-dimensional space is shown in Fig. 6. The points on the lower surface are indexed from 0 to 29 while the points on the upper surface are indexed from 30 to 59. On the two surfaces, there are 6 group of points from the wing root to the winglet. Each group has 5 points, whose index increases from the trailing edge to the leading edge.

![Fig. 6 Schematic diagram of control frame and design variables of M6 airfoil(left) and one example of explanation results in the form of thermodynamic diagram (right)](image)

### 3.2. Explanation Result

As for the black box model to be explained, the neural network has two single hidden layers with 300 neurons and 30 neurons respectively with the same activation function ‘Relu’. The output layer has only 1 neuron for solving the regression problem to predict the corresponding aerodynamic coefficient (Lift coefficient, drag coefficient, torque coefficient). The training process is based on the mean squared error loss. To validate the model, we use 80 − 20 split on dataset for training and testing. The final mean squared error are 1.0619 × 10^{-7}, 1.1956 × 10^{-7}, 2.7550 × 10^{-7} on the training set, 2.0493 × 10^{-6}, 3.5445 × 10^{-6}, 5.9020 × 10^{-6} on the test set for above mentioned three aerodynamic coefficients respectively. The explanation results of LIME and UniformLIME for one random instance on predicting three coefficients from the test set is shown in Fig. 7. The vertical axis represents features of input. The horizontal axis represents the coefficient of fitted linear model. The positive coefficients (green bar) show the responding design variable has the positive influence while predicting the aerodynamic coefficient. Negative coefficients (red bar) are on the contrary. All features are ranked by the descending order of the absolute value of coefficient from top to bottom. This graph only presents top-10 design variable after explanations.
Fig. 7 Comparisons of explanation result between LIME (first row) and UniformLIME (second row).

Three columns represent drag coefficients, lift coefficients, and torque coefficients from left to right. To verify the correctness of the explanation results, we also compare the explanation result with the traditional aerodynamic method. Fig. 8 shows the gradient change of three coefficients. The horizontal axis refers to the index of design variable (feature) while the vertical axis refers to the gradient of each feature. With the larger gradient, the corresponding design variable is more important.
Fig. 8 Gradient values of drag(left) and lift(right) coefficients calculated by discrete adjoint method and finite difference method for M6 airfoil

To get the similar result from LIME and UniformLIME for direct comparison, firstly we rank all features by the absolute value of corresponding coefficients. Then we calculate the average rank of each feature after explaining all 120 instances in the test set. The result of average rank for three subtasks and two different interpretable models is shown in Fig. 9. The horizontal axis refers to the index of feature while the vertical axis refers to the average rank of the absolute value of corresponding coefficients for 120 instances in the test set. The orange line represents our modified UniformLIME while the blue line represents the original LIME.
Fig. 9 Average importance rank by LIME for three aerodynamic coefficients

By comparison between traditional method and local surrogate method, it is obvious that the tendency is the same and the feature importance is similar. For instance, in both two explanation results of lift coefficient, the feature importance decreases from the front edge to the trailing edge and varies periodically every 5 point. The most important design variable for determining the lift coefficient is the part near the wing root and the front edge.

By comparison between original LIME and our modified UniformLIME, the tendency for one group of 5 points is more obvious. For example, for the first 5 features (index from 0 to 4), the difference between the maximum rank and the minimum rank is larger for UniformLIME, which means the algorithm can distinguish the most important feature more confidently. The important feature gets the lower rank, and the unimportant feature gets the higher rank.

3.3. Metric

To evaluate LIME and UniformLIME mathematically, we calculate 6 different metrics on three subtasks to compare two models precisely. Metrics about local fidelity, including mean absolute error, mean median error, mean squared error and $R^2$ score, are shown in Fig. 10. These metrics are computed on the whole test set, regarding the prediction results from the black box model as the ground truth and the fitting result from the linear regression model as the predictive value. Besides, metrics about stability, including standard deviation and the ratio of standard deviation to mean, are shown in Fig. 11. These metrics are computed under a random selected instance from the test set with 10 iterations, recording the change of the fitting coefficients via different iterations. Additionally, we vary the number of perturbed samples used for training the local surrogate model from 500 to 10000 with an interval of 500. UniformLIME work better than LIME on both local fidelity and stability.
Fig. 10 Mean absolute error, mean median error, mean squared error and $R^2$ score.
4. Conclusion
In this paper, we proposed an original perturbed sample generation method for aerodynamic tabular data as a modification to regression LIME by uniformly generating perturbed samples to make the distance and the weight distribute uniformly in the domain. By serval experiments on three different subtasks on an aerodynamic dataset, it is showed that our method works better on stability as well as local fidelity than original LIME. For future work, we would like to explore diversified datasets and complex models to be explained in different scenarios. More theoretical verification and experimental analysis are needed.

Acknowledgments
The dataset is supported by Institute of Aerodynamic and Multidisciplinary Design Optimization, National Key Laboratory of Science and Technology on Aerodynamic Design and Research, School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, P. R. China.

References
[1] Fan H Y, Dulikravich G S, Han Z X 2005 Aerodynamic data modeling using support vector machines Inverse Problems in Science and Engineering 13(3) pp 261-278
[2] Molnar C. 2020 Interpretable Machine Learning Lulu.com
[3] Ribeiro M T, Singh S, Guestrin C 2016 Why should i trust you? Explaining the predictions of any classifier Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining pp 1135-1144
[4] Alvarez-Melis D, Jaakkola T S 2018 On the robustness of interpretability methods arXiv
preprint arXiv: 1806.08049

[5] Ribeiro M T, Singh S, Guestrin C 2018 Anchors: High-precision model-agnostic explanations

Proceedings of the AAAI Conference on Artificial Intelligence 32(1)

[6] Shankaranarayana S M, Runje D 2019 ALIME: Autoencoder based approach for local interpretability

International conference on intelligent data engineering and automated learning Springer Cham pp 454-463

[7] Zafar M R, Khan N M 2019 DLIME: a deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems arXiv preprint arXiv: 1906.10263