Prediction Ability Analysis of Common Machine Learning Algorithms on Flow Field

Chaoqun Li, Shuo Tang, Yi Li * and Zihai Geng

School of Astronautics, Northwestern Polytechnical University, Xi'an, China

* Corresponding author e-mail: jwshlcq@mail.nwpu.edu.cn

Abstract. In this work, the primary mission is to apply the machine learning algorithms into practical use, predicting the velocity value at different positions according to the previous velocity history. There is a large data set containing the velocity data of 12000-time steps. Using the nonlinear methods, multiple linear regression (MLR), k-nearest neighbor (KNN), support vector machine (SVM), and artificial neural networks (ANN), the predictions of u and v component of the flow in the next time step are established. The results of the prediction will be compared with the actual data which is at the next time point. Then, the mean error and standard deviation in each case are evaluated. The results illustrate that ANN method possesses the lowest error and narrow distributions, which shows good accuracy and stability, the KNN method and SVM method are weaker than the artificial neural networks, and the MLR algorithm is the most inaccurate.

Keywords: Common Machine Learning Algorithms; Flow Field; The Practical Application.

1. Introduction

In the field of fluid mechanics, with the development of measurement techniques and computational fluid mechanics, the amount of data becomes more and more huge [1]. To find the hidden correlations of the data inside the tons of data, the machine learning methods play an important role[2].

Generally, *k*-nearest neighbor, support vector machine, and artificial neural networks are common machine learning methods, and the methods can be used for regression and prediction purposes [3]. For some methods, like *k*-nearest neighbor method, the difference between the prediction and the actual value could dampen to zero [4]. In [5], Trawiński et al. investigated the performance of machine learning methods on the nonparametric analysis and compared the results of regressions. Brenner et al. revealed the potential of machine learning methods in predicting the flow field based on the training data [6]. In [7] of Rowley and Dawson, the Proper Orthogonal Decomposition (POD method) could be regarded as a linear machine learning technique. Hence, for fluid mechanics, machine learning methods can be used to simulate, predict, and analyze the flow field [8].

Based on the existing measurement of flow field, this work aims to predict the velocity using common machine learning methods. The training dataset contains flow field in 12000 time points at 11898 positions. And the multiple linear regression method, *k*-nearest neighbor, support vector machine, and artificial neural networks are employed to predict the *u*-component of wind speed, *v*-component of the flow field. Subsequently, the results are compared with the actual values. In the following sections, the methods, training data, and results are illustrated; finally, the conclusion are drawn and discussed.

2. Machine Learning Method

The machine learning algorithms are constructed based on a domain set $\mathbf{X}$ which is a subset of $\mathbf{R}^{\mathbf{m}}$, and a dependent set $\mathbf{Y}$, which is belong to $\mathbf{R}$. At the same time, the training set is also provided, $\{(x_1, y_1),\ (x_2, y_2),\ (x_3, y_3)\cdots(x_N, y_N)\}$, where $x_i \in \mathbf{X},\ y_i \in \mathbf{Y}$.

In this section, the methods of machine learning are illustrated, and the methods contain multiple linear regression and multiple linear regression of linear methods, k-nearest neighbor, support vector machine and artificial neural networks.

2.1. Multiple Linear Regression Method

The modification of an independent variable can influence the effects of another independent parameter on corresponding dependent variable. In this work, the multiple linear regression (MLR) method is extended and the interaction of the independent variables needs to be taken into account. And a simplest example of interaction among the parameters is considered, which is the term of a product of two variables. And this term is added into the formula with a weight. Thus, the function will be nonlinear. And the function relationship between one dependent variable and two independent variables can be defined as the following:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \tag{1}$$

where β_0, β_1, β_2, and β_3 are the coefficients to be determined, and they can be obtained by Cramer's rule, as the following:

$$\begin{pmatrix} N & x_1 & x_2 & x_1x_2 \\ Nx_1 & x_1^2 & x_2x_1 & x_1^2x_2 \\ Nx_2 & x_1x_2 & x_2^2 & x_1x_2^2 \\ Nx_1x_2 & x_1^2x_2 & x_2^2x_1 & x_1^2x_2^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{N}(y_i) \\ x_1\sum_{k=1}^{N}(y_i) \\ x_2\sum_{k=1}^{N}(y_i) \\ x_1x_2\sum_{k=1}^{N}(y_i) \end{pmatrix} \tag{2}$$

2.2. k-nearest Neighbor Method

The *k*-nearest neighbor (KNN) a kind of machine learning algorithm that can be simply applied to regression problems. This method believes that the nearer points will affect the target points seriously, while the effect of remote ones will be weaker, as shown in Fig.1. Hence, the KNN algorithm needs to store all the training data; then make predictions.

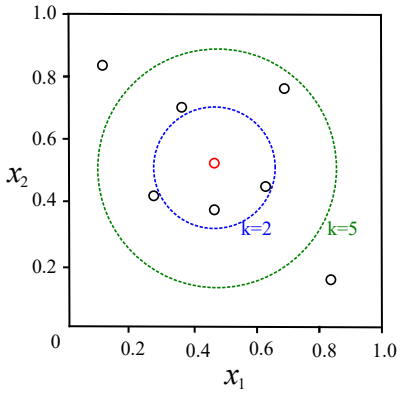


Figure 1. Schematic plot of *k*-nearest neighbor method.

In this algorithm, *k* represents the number of selected neighbors. If *k* is set to 1, the model will become a single nearest neighbor algorithm. Then, the value of the target point is totally determined by the nearest point. And this is an extreme circumstance and this will introduce inaccurate or even incorrect results. Hence for the sake of accuracy, *k* is set to large constant which is 1000 in this work. And the prediction will be based on their distances.

Subsequently, the distance is needed to be given and determined. There are some common functions to evaluate the distance, the first one is the Euclidean Distance:

$$d(X,Y)=\sqrt{\sum_{k=1}^{n}\left(x_k-y_k\right)^2} \tag{3}$$

The second one is the Manhattan Distance:

$$d(X,Y)=\sum_{k=1}^{n}\left|\left(x_k-y_k\right)\right| \tag{4}$$

The third one is the Minkowski Distance:

$$d(X,Y)=\sqrt[\frac{1}{q}]{\left|\sum_{k=1}^{n}\left(x_k-y_k\right)\right|^q} \tag{5}$$

In this work, the Equation (3), Euclidean Distance is employed. The the procedure of making the distance non-dimensional is utilized, as shown in the following:

$$x_i=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}}(i=1,2,\cdots,m) \tag{6}$$

where x_i denotes the indenpendent variable, the range of the value of which are between 0 and 1. To determine the weight of each data point, the inverse of the distance is used.

$$\hat{y}=\begin{cases}\dfrac{\sum_{i=1}^{k}\left(w_i y_i\right)}{\sum_{i=1}^{k} w_i}, d\left(x^{'},x_i^{'}\right)\neq 0\\ y_i, d\left(x^{'},x_i^{'}\right)=0\end{cases} \tag{7}$$

with

$$w_i=\frac{1}{d\left(x_i,x_i^{'}\right)} \tag{8}$$

where x_i' denotes the normalized variables in the training set, and x' denots that corresponding to parameter needed to be determined. Hence, the KNN method is established.

2.3. Support Vector Machine Method

Generally, support vector regression algorithm (SVM) is a category of nonlinear generalization of generalized portrait algorithms. By implementing the structural risk minimization inductive principle, SVM algorithm can possess accurate prediction based on a limited training data. For many methods, they try to minimizing the observed error, while SVM aims to minimizing the generalized error. In SVM algorithms, the method concentrates on solving a constrained quadratic problem with constraints. And the loss function is constructed based on which the minimum value of the convex objective function. Moreover, the simple demonstration of the SVM is illustrated in Fig.2.

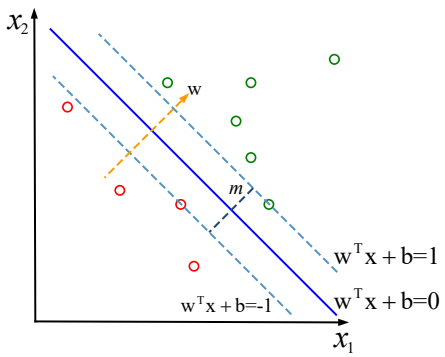


Figure 2. Schematic plot of support vector regression method.

Hence, the model of the SVM method is as the following. By solving the problem the value of the target point will be obtained.

$$\text{minimize} : \frac{1}{2}\| w\|^2 + C\sum_{k=1}\left(\xi_i + \xi_i^*\right)$$
$$\text{subject to} \begin{cases} y_i - \langle w, x\rangle - b \le \epsilon + \xi_i \\ -y_i + \langle w, x\rangle + b \le \epsilon + \xi_i^* \\ \tilde{\xi}_i, \tilde{\zeta}_i^* \ge 0 \end{cases} \tag{9}$$

2.4. Artificial Neural Networks

A typical structure of the Artificial Neural Networks (ANN) can be seen in Fig.3. The ANN contains many layers, and the node at each layer is connected with the nodes before and after the local layer. Generally, the layer consist of the input layer, hidden layers and output layer. In this work, the networks with a single hidden layer is employed, and there is only one output node. The data of the target point will be obtained by the ANN model.

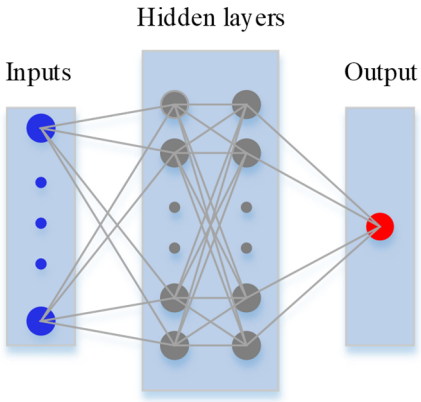


Figure 3. Schematic plot of artificial neural networks method

3. Implementation and Results

The training data in this work is a set of data where a flow field at 11848 specific points. And the information on the velocity vector at previous 12000-time steps is provided. By studying the training data, the velocity at the next time step is predicted by the algorithms at Section 2.

The training data in this work is a set of data where a flow field at a specific point. To evaluate the performance of the methods, the data obtained by the method and the real value are compared and the error is defined as the following:

$$e_{ml} = \left|\frac{x_{ml} - x_{act}}{x_{act}}\right| \times 100\% \tag{10}$$

where e_{ml} denotes the error of machine learning method, x_{ml} denotes the value obtained by the machine learning method, and the x_{act} denotes the actual value. Then, an overall mean error and standard deviation are given, and the error is defined as the following:

$$e_m = \frac{1}{N}\sum_{N} e_{ml} \times 100\% \tag{11}$$

And the standard deviation:

$$s = \sqrt{\frac{1}{N}\sum_{N}\left(e_{ml} - e_m\right)^2} \tag{12}$$

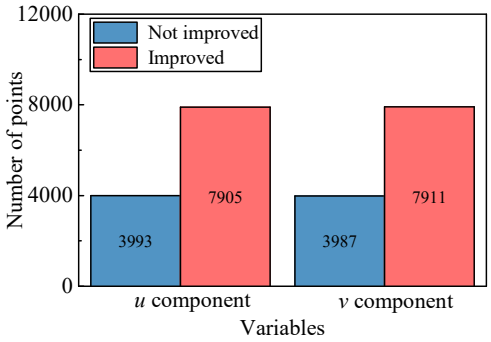
where e_{ml} denotes the overall mean error, and s denotes the standard deviation. And the mean error reflects the accuracy of the method, besides which standard deviation denotes stability of the method. The performance of the methods is shown in Table 1.

Table 1. The performance of the machine learning methods

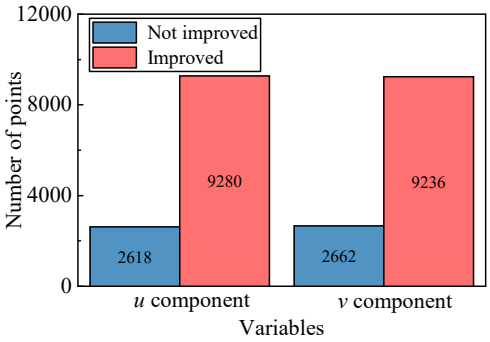
Algorithms	*u* component		*v* component	
	e_m	*s*	e_m	*s*
Multiple Linear Regression	10.16%	0.06148	20.75%	0.08877
K-Nearest Neighbor	5.344%	0.05218	7.844%	0.07505
Support Vector machine	1.308%	0.05078	1.788%	0.07268
Artificial Neural Networks	1.282%	0.05059	1.713%	0.06856

As reflected in Table 1, the ANN method possesses the best performance of the prediction with an 1.282% overall mean error, while the MLR method performs worst where the error is 10.16%. At the same time, the ANN has the finest stability, while MLR mtehod is the most unstable. Despite all this, the stability of each method is of the same magnitude.

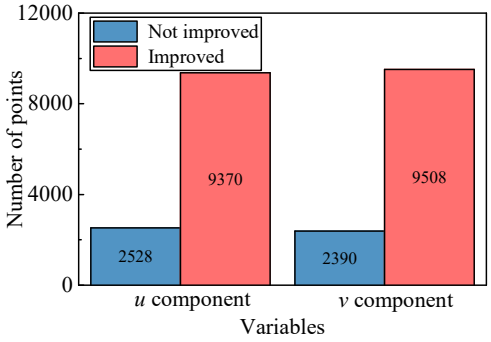
The outputs of the machine learning methods are compared with the least squares method. And the index is defined, if the error of the machine learning method is lower than the prediction of the least squares method, the result is marked with "Not improved"; on the contrary, the result is "Improved". Hence, the comparison results are as the Fig.4.



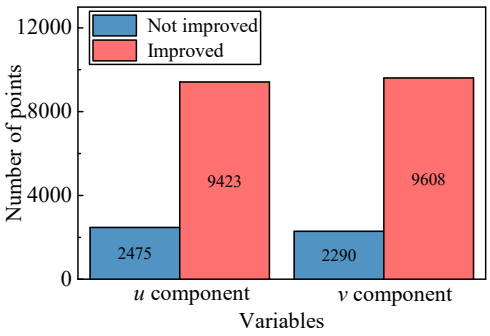
(a) Multiple linear regression



(b) *k*-nearest neighbor



(a) Support vector machine



(b) Artificial neural networks

Figure 4. Performance of different machine learning method.

From Fig.4, it can be observed that the ANN improves the predictions accuracy highly, while the MLR method shows a lowest accurancy. Besides the performance of the KNN and SVM are similar.

4. Conclusion

In this work, the prediction performance of methods of the multiple linear regression, *k*-nearest neighbor, support vector machine and artificial neural networks are examined. And the prediction results are compared with the actual value, and the main conclusion are as the following:

Among the methods, the artificial neural networks method possesses the best performance of the prediction with a lowest overall mean error, at the same time, it is the most stable. On the other hand

the multiple linear regression method performs worst with a large error and possesses the lowest stability.

Compared with the linear square's method, the machine learning methods in this work can improve the accuracy of the predictions. Artificial neural networks improve the accuracy most, the *k*-nearest neighbor and support vector machine are similar, weaker than the artificial neural networks, and the multiple linear regression is of lowest improvement.

Acknowledgments

This work was not financially supported by any funding.

References

[1] S.L. Brunton, B.R. Noack, P. Koumoutsakos, Machine learning for fluid mechanics, Annu. Rev. Fluid Mech. 52 (2020) 477–508.

[2] Y. Wei, C.D. Chiu, Machine learning with R cookbook, Packt Publishing Ltd, Birmingham, 2015.

[3] E.W. Steyerberg, T. van der Ploeg, B. van Calster, Risk prediction with machine learning and regression methods. Biom. J, 56(2014) 601-6.

[4] J. Kruppa, Y. Liu, G. Biau, M. Kohler, I. R. Konig, J. D. Malley, and A. Ziegler, Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. Biom. J, 56 (2014) 534–563.

[5] B. Trawiński, M .Smętek, Z. Telec, T. Lasota, Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. Int. J. Appl. Math. Comput. Sci, 22 (2012) 867–81.

[6] M.P. Brenner, J.D. Eldredge, J.B. Freund, Perspective on machine learning for advancing fluid mechanics, Physical Review Fluids. 4(2019) 100501.

[7] C. W. Rowley, S. T. M. Dawson, Model reduction for flow analysis and control, Annu. Rev. Fluid Mech. 49 (2017) 387–417.

[8] S. Pawar, S.M. Rahman, H. Vaddireddy, O. San, A. Rasheed, P. Vedula, A deep learning enabler for nonintrusive reduced order modeling of fluid flows, Phys. Fluids. 31 (2019) 085101.