A k-interpolation Model Clustering Algorithm based on Kriging Method

Guoyan Chen*, Yaping Qian
School of Mechanical Engineering
Jiangsu University of Technology, Changzhou, China

***Abstract*—In this work, a k-interpolation model clustering algorithm is proposed based on Kriging method, aim to partition data according to the relationship between the response of interest and input variables. Kriging method is used to describe the relationship between the response of interest and input variables. For each datum, the estimation errors of the interpolation models of the clusters are used to decide its assignment. An optimization strategy is proposed to obtain the final clustering results. The key factors of the proposed algorithm on its performance are studied through the synthetic and real-world datasets. The results show that the proposed algorithm is able to cluster the data according to the response of interest and input variables, and provides competitive clustering performance compared with the other clustering algorithms.**

***Keywords*—Data clustering; Kriging method; k-means algorithm; interpolation model**

I. Introduction

In recent years, massive data have been generated and recorded from real-world systems. The information mined from these data represents the characteristics of the real-world system, which can be used to analyze and improve the performance of the system. In most data mining tasks, it is necessary to build the performance prediction model first, aiming to accurately estimate the response of interest according to the input variables. However, the relationship between the response and input often changes greatly, which is difficult to evaluate through a unified prediction model [1-3]. Obviously, this issue can be solved by partitioning the data so that the data in the same part have a more similar relationship between response and input than the data from the other parts, and this work can be accomplished by data clustering.

Data clustering is a class of algorithms and techniques aiming to partition a dataset such that the data characteristics in the same cluster are more similar than the other clusters [4]. Many clustering algorithms have been proposed in the past decades, such as k-means algorithm, fuzzy c-means algorithm, Gaussian mixture model, and so on [5-6]. Since the k-means algorithm is easy to understand and implement, it has been widely used in many data mining tasks such as image recognition, modal analysis, and outlier detection. Shubair, and Al-Nassiri used the least square method to estimate the centers of clusters in k-means algorithm and applied the clustering algorithm in the preparation process of data streams [7]. Aldino et al. used k-means algorithm to group the corn-producing regions based on the collected data of corn crops to assist in the formulation of corn planting [8]. Yu et al. proposed multi-layers framework to increase the performance of k-means algorithm on the dataset with outliers and noisy values [9]. In addition, genetic algorithm is used to obtain the optimal clustering results. Zhu et al. proposed a grid k-means algorithm to improve the clustering accuracy and stability and validated its performance on the dataset with the noise points [10]. Cuomo et al. used parallel techniques to reduce the computation cost of k-means algorithm for the large data analytic problem and provides solutions for the problems of GPU space limitation and host-device data transfer time [11]. k-means algorithm clusters data according to their spatial distance, resulting in it being difficult to ensure that the data in the same cluster have a similar or same relationship between the response of interest and input variables. Thus, it is necessary to develop a new clustering method under the framework of k-means algorithm.

In recent years, an interpolation model, Kriging method, has been widely used to model the relationship between the response and input variables of the measured data of the real-world systems. For example, Echard et al. assessed the failure probabilities of an engineering system using the importance sampling method and Kriging method, which has been successfully used in the reliability analysis of engineering systems [12]. Keshtegar et al. used Kriging method to estimate the solar radiation based on the meteorological data [13]. Wojciech proposed a digital terrain estimation method based on Kriging method, in which a neighbor points selection method is designed to accelerate the training speed Kriging method [14]. Belkhiri et al. estimate the groundwater quality for drinking purposes using Kriging method [15]. The results indicate that the Kriging model with electrical conductivity as co-variable produces the best performance compared with the other Kriging models. From the above works, it can be seen that Kriging method can effectively learn the relationship between the response of interest and input variables from the measured data. Thus, a k-interpolation model clustering algorithm is proposed based on Kriging method under the framework of k-means algorithm in this work, aims to partition data according to the relationship between the response of interest and input variables. Kriging method is used to evaluate the relationship between the response and input variables. For each datum, the estimation errors of the interpolation models of the clusters are used to decide its assignment. An optimization strategy is designed to obtain the clustering results. Finally, the performance of the proposed algorithm is validated through several synthetic and real-world datasets. The remainder of this work is organized as follows. The proposed algorithm

*Corresponding Author.

including the background of k-means algorithm and Kriging method is introduced in Section 2. The synthetic datasets and engineering datasets are used to test and compare the performance of the proposed algorithm with the conventional clustering algorithms in Sections 3 and 4. The conclusions are provided in Section 5.

II. Literature Review

In recent years, several data clustering algorithms have been proposed to partition data according to their relationship between the response of interest and input. Peng et al. [16] introduced ridge regression to evaluate the relationship of two-dimensional data in their clustering. Chen et al. [17] used the least square method to evaluate the features of data, and then applied fuzzy c-means algorithm to cluster them. However, only the linear relationship is considered in the above methods. To realize data clustering based on their nonlinear relationship between the response of interest and input, artificial neural networks and Gaussian process regression have been used to replace linear models. For example, Blažič et al. [18] used artificial neural networks to evaluate the nonlinear regression relationship to identify the state of engineering systems. Fuhg et al. [19] applied Gaussian process regression to evaluate the relationship among attributes to partition data according to their variation ranges. Fang et al. [20] used artificial neural networks to evaluate the relationship among data attributes to cluster the in-situ data of a tunnel boring machine.

III. Proposed Method

A. K-means Algorithm

k-means algorithm is developed in the area of signal processing, which aims to partition n data into k clusters in which each datum belongs to the cluster with the nearest mean (the prototype of the cluster). Generally, the clustering process of k-means algorithm can be subdivided into two stages: assignment step and update step as follows.

Assignment step: each datum is assigned to the cluster with the nearest prototype as follows

$$S_i^t = \{datum_p : \|d_{ip}\|^2 \le \|d_{jp}\|^2 \forall j, 1 \le j \le k\} \quad (1)$$

where d represents the distance between the datum and the mean (Euclidean distance is usually used), and $datum_p$ is assigned to exactly one S_i^t.

Update step: the mean (prototype) of each cluster is recalculated as follows.

$$m_i^{t+1} = \frac{1}{|S_i^t|}\sum_{datum_j \in S_i^t} datum_j \quad (2)$$

The iterations are carried out until the assignments no longer change.

B. Kriging Method

In Kriging method (KRG), the following model is used to model the outputs at the samples:

$$Y(x) = f^{\mathrm{T}}(x)\beta + Z(x) \quad (3)$$

where $Y(x)$ is the function of interest, $f = [f_1(x), f_2(x), \dots, f_m(x)]^{\mathrm{T}}$ is the basis functions, and $\beta = [\beta_1, \beta_2, \dots, \beta_m]^{\mathrm{T}}$ is the corresponding coefficient vector. $Z(x)$ is a Gaussian stationary process with zero mean and covariance.

$$\mathrm{Cov}(x_i, x_j) = \sigma^2 R(\theta, x_i, x_j)\ i, j = 1,2, \dots, n \quad (4)$$

where σ^2 is the process variance, $R(\theta, x_i, x_j)$ is the correlation function of the stochastic process, θ is the hyper-parameters of $R(\theta, x_i, x_j)$, and n is the sample number, The maximum likelihood method is used to optimize θ, where the likelihood function is expressed as follows:

$$L = (2\pi\sigma^2)^{-\frac{n}{2}}|R|^{-\frac{1}{2}}\exp\left[-\frac{1}{2\sigma^2}(Y - F\beta)^T R^{-1}(Y - F\beta)\right] \quad (5)$$

where R is the correlation matrix and F is a vector including the value of $f(x)$. β and σ^2 are estimated through the least-square method as follows.

$$\hat{\beta} = (F^T R^{-1} F)^{-1} F^T R^{-1} Y \quad (6)$$

$$\widehat{\sigma^2} = \frac{1}{n}(Y - F\beta)^{\mathrm{T}} R^{-1}(Y - F\beta) \quad (7)$$

By taking the logarithm of Eq. (5) with the imposed σ^2 value and multiplying by -1, the maximum problem to obtain the optimal θ is revised as

$$\text{minimize } \frac{1}{2}\ln(|R|) + \frac{n}{2}\ln(|\sigma^2|) \quad (8)$$

The prediction of Kriging method for a new sample x^* is

$$y(x^*) = f^{\mathrm{T}}(x^*)\beta + r^{\mathrm{T}} R^{-1}(Y - F\beta) \quad (9)$$

where $r = [R(\theta, x_1, x^*), R(\theta, x_2, x^*), \dots, R(\theta, x_n, x^*)]$.

C. The Proposed Algorithm

A k-interpolation model clustering algorithm is proposed based on Kriging method in this section. From Eq. (1), it can be known that the distance d should involve the relationship between the response of interest and input variables, if we want to cluster the data based on the relationship. In this work, Kriging method is used to evaluate the relationship, and the estimated response of each datum can be obtained as flows.

$$\widehat{y_{p,i}} = KRG(x_p)_i \quad (10)$$

where $\widehat{y_{p,i}}$ is the estimated response of k-th datum of i-th cluster, x_p is the vector of input variables. The distance d is defined as follows.

$$d_{ip} = |y_p - \widehat{y_{p,i}}| \quad (11)$$

Similar to k-means algorithm, the clustering process of the proposed algorithm (named k-IM) is summarized as follows.

Step 1. Set the clustering number c;

Step 2. Generate the assignment of the data randomly;

Step 3. Construct KRG model of i-th cluster based on the data contained in the cluster;

Step 4. Using the obtained KRG models to get the responses of all the data and creating the responses matrix $Y_{n\times C}$;

Step 5. Assigning each datum to the cluster using Eq. (1) and Eq. (11).

Step 6. If any stop conditions are satisfied, the procedure is stopped, and the current assignment results are considered as the final clustering results, otherwise, return to Step 3.

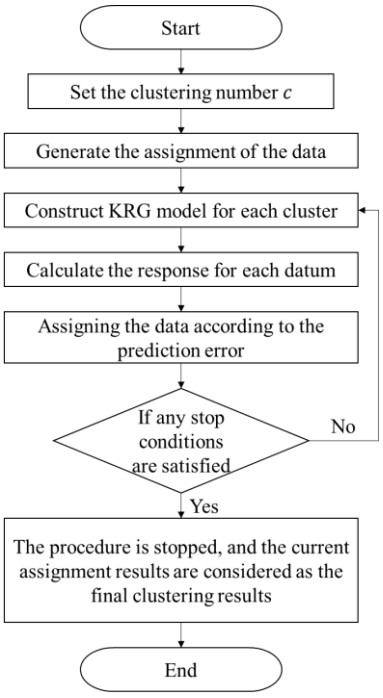

Fig. 1. Flow-chart of the Propsoed Algorithm.

IV. Experiments on Synthetic Datasets

In this section, the synthetic datasets are used to validate the proposed algorithm. For each dataset, the data of each cluster is generated first and combined as the final dataset. The Latin hypercube sampling method is used to generate the input variables, and then the corresponding responses are calculated through the setting relationship between the response and input variables. The naming of the dataset is based on its sample number and cluster number. For example, N400C2 means that the dataset has 400 samples and two clusters. The proposed algorithm is compared with three popular clustering methods, k-means algorithm (KM), fuzzy c-means algorithm (FCM), and Gaussian mixture model (GMM). The clustering performance is evaluated through the following indexes.

1) Misclassification rate (MS):

$$MS = \frac{N_{error}}{N_{total}} \tag{12}$$

where N_{error} is the number of misclassified data; N_{total} is the total number of data. The lower MS, the higher cluster validity.

2) Adjusted rand index (ARI) [21]: Given a set S of n elements, and two partitions of these elements, namely $X = \{X_1, X_2, \dots, X_s\}$ and $Y = \{Y_1, Y_2, \dots, Y_s\}$, the overlap between X and Y can be summarized in a contingency table $[n_{ij}]$ where each entry n_{ij} denotes the number of objects in common between X_i and Y_j : $n_{ij} = |X_i \cap Y_j|$ as shown in Table I. Adjusted rand index is defined as follows:

$$ARI = \frac{\Sigma_{i,j}\binom{n_{ij}}{2} - \left[\Sigma_i \binom{a_i}{2} \Sigma_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\Sigma_i \binom{a_i}{2} + \Sigma_j \binom{b_j}{2}\right] - \left[\Sigma_i \binom{a_i}{2} \Sigma_j \binom{b_j}{2}\right] / \binom{n}{2}} \tag{13}$$

The closer ARI to 1, the higher cluster validity.

TABLE I. Contingency Table

$X \backslash Y$	Y_1	Y_2	…	Y_s	Sums
X_1	n_{11}	n_{12}	…	n_{1s}	a_1
X_2	n_{21}	n_{22}	…	n_{2s}	a_2
…	…	…	…	…	…
X_s	n_{r1}	n_{r2}	…	n_{rs}	a_r
Sums	b_1	b_2		b_s	

3) Normalized mutual information (NMI) [22]:

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{14}$$

where $I(\cdot)$ is the mutual information metric and $H(\cdot)$ is the entropy metric. The closer NMI to 1, the higher cluster validity.

A. Effect of Sample Number

In this section, four synthetic datasets are used to study the effect of sample number on the performance of the k-IM algorithm. In each dataset, there are two clusters, and each cluster has the following relationship between the response and input.

Cluster 1: $y = (6x-2)^2 * \sin(12x-4)$

Cluster 2: $y = 0.6(6x-2)^2 * \sin(12x-4) + 12(x-0.5) + 6$

where $x \in [0,1]$. For each synthetic dataset, one cluster has 150, 200, 250, 300 samples, respectively. Thus, the four synthetic datasets are denoted as N100A2C2, N200A2C2, N300A2C2, N400A2C2, respectively. The obtained N400A2C2 dataset is shown in Fig. 1. From Fig. 2, it can be seen that seen that the samples of the two clusters are distributed similarly, but the relationship between the response of interest and input is different. The 30 times experiments are conducted for each dataset. The average experimental results are shown in Tables II to IV.

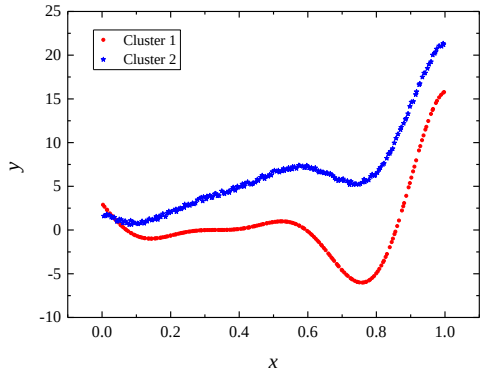

Fig. 2. N400C2 Dataset.

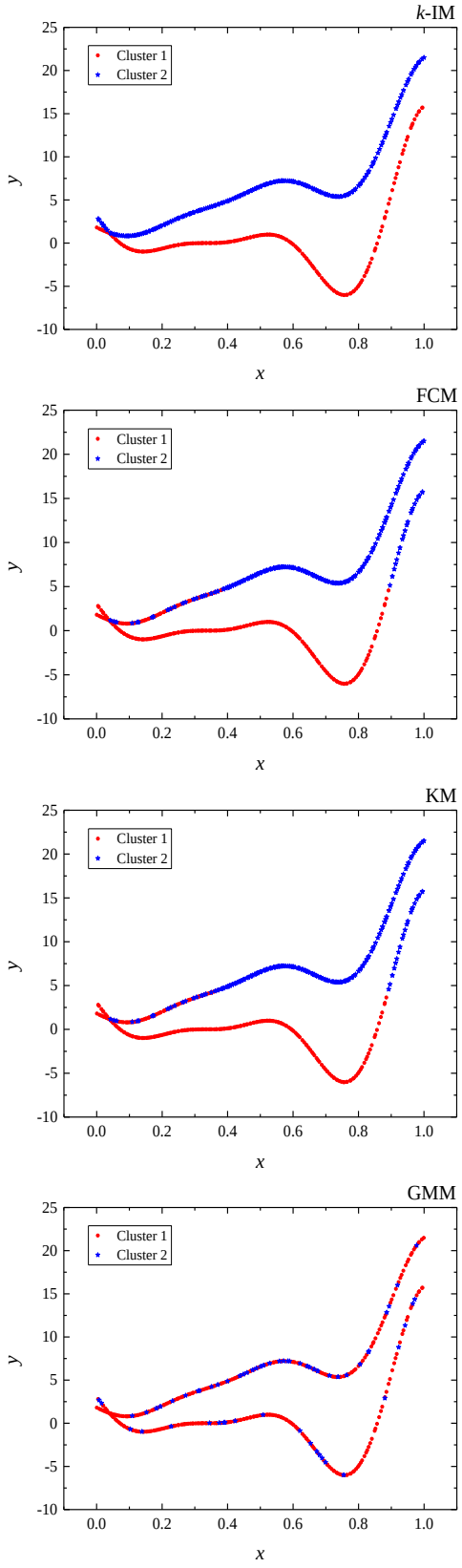

Fig. 3. Clustering Results Comparison of N400C2 Dataset.

From these tables, it can be seen that the k-IM algorithm produces much better results than the FCM, KM, and GMM algorithms. The mean misclassification rate of the proposed algorithm is less than 0.03, which is much smaller than those of FCM, KM, and GMM, indicating the k-IM algorithm is able to accurately cluster the synthetic datasets. To further compare the performance of the clustering algorithms, the clustering results of N400C2 dataset of one experiment are shown in Fig. 3. From this figure, it can be seen that the proposed algorithm clusters the data based on the relationship between the output and input. The FCM and KM algorithms cluster the data according to their spatial distribution. It is noted that the GMM algorithm assigns most data to one cluster. The reason is mainly that it clusters data with the assumption that the data obey a Gaussian mixed distribution. The assumption cannot be stratified for N400A2C2 dataset. Thus, the clustering results of the GMM algorithm are much worse than the other algorithms. From the experimental results shown in Tables II to IV, it is observed that the sample number has an effect on the proposed algorithm. With the sample number increasing from 300 to 600, the MS of the proposed algorithm first decreases to 0.013 and then increases to 0.019. Similar results can be found in the indexes ARI and NMI. The reason is explained as follows. As the sample number increases, more samples can be utilized to construct the KRG models, which mean that the relationship between the output and input can be evaluated more accurately. The performance of the k-IM algorithm increases with the increase in the sample number. However, the KRG model tends to be overfitting when the sample number is too large. Thus, the performance of the k-IM algorithm decreases with the sample number increasing from 500 to 600. The proposed algorithm produces competitive clustering results for the datasets with different sample numbers tested in this section.

TABLE II. THE EXPERIMENTAL RESULTS OF *MS*

Dataset	***k*-IM**	**FCM**	**KM**	**GMM**
N300C2	0.028	0.247	0.370	0.439
N400C2	0.027	0.248	0.391	0.467
N500C2	0.013	0.246	0.381	0.464
N600C2	0.019	0.246	0.339	0.485

TABLE III. THE EXPERIMENTAL RESULTS OF *ARI*

Dataset	***k*-IM**	**FCM**	**KM**	**GMM**
N300C2	0.894	0.253	0.107	0.099
N400C2	0.894	0.251	0.084	0.080
N500C2	0.951	0.255	0.095	0.076
N600C2	0.926	0.256	0.146	0.069

TABLE IV. THE EXPERIMENTAL RESULTS OF *NMI*

Dataset	***k*-IM**	**FCM**	**KM**	**GMM**
N300C2	0.833	0.215	0.096	0.094
N400C2	0.839	0.214	0.077	0.067
N500C2	0.923	0.216	0.085	0.063
N600C2	0.883	0.216	0.126	0.068

B. Effect of Cluster Number

Three datasets with different cluster numbers are used to test the effect of clustering number on the proposed algorithm, as shown in Table V. It is observed that the clusters of each dataset have similar but different relationships between the response and input variables. For each dataset, 30 times experiments are conducted. The average clustering results are shown in Fig. 4.

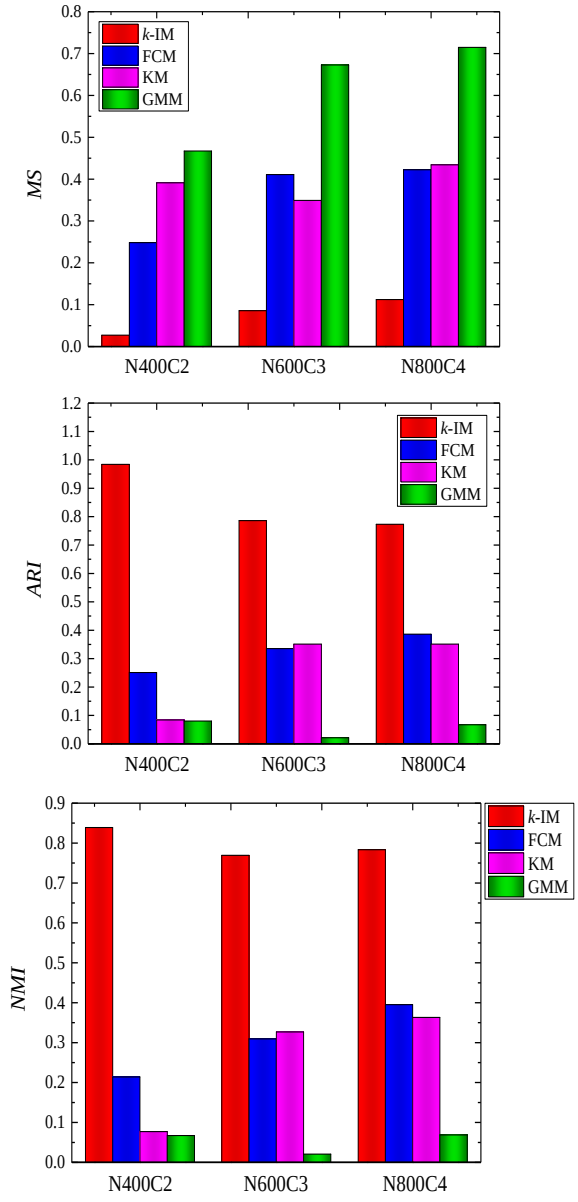

Fig. 4. Clustering Results of N400C2, N600C3 and N800C4 Datasets.

From Fig. 3, it is observed that the k-IM algorithm is still able to produce the best results among the tested four algorithms. The highest misclassification rate of the k-IM algorithm is around 0.10, which is much smaller than the conventional FCM, KM, and GMM algorithms. The index ARI of the k-IM algorithm is higher than 0.80 for all three datasets. The index NMI is around 0.80, which is higher than the other clustering algorithms as well. It is noted that the misclassification rate of the GMM algorithm is higher than 0.50 for N600C3 and N800C4 datasets. The reason is that the GMM algorithm clusters almost all the data into one class, which means that most data are misclassified. Thus, the MS is higher than 0.50. With the cluster number increasing, the performance of the k-IM algorithm decreases, but it is still much better than the other popular clustering algorithms. The proposed algorithm can produce competitive clustering results when clustering the dataset with different cluster numbers tested in this section.

C. Effect of Noise

The measured data of real-world systems usually have noise. N400A2C2 dataset is used to test the performance of the *k*-IM algorithm on the noise. The synthetic datasets are generated as follows. For each cluster, the input variables are generated. The response is calculated according to the set function. For each datum, a random value is generated according to the set interval as shown in Table VI and added to the response. The average clustering results of 30 times experiments are shown in Tables VI to VIII.

TABLE V. EFFECT OF SAMPLE ON THE CLUSTERING PERFORMANCE (MS)

Dataset	Relationship
N400C2	$y = (6x-2)^2 * \sin(12x-4)$
	$y = 0.6(6x-2)^2 * \sin(12x-4) + 12(x-0.5) + 6$
N600C3	$y = (6x-2)^2 * \sin(12x-4)$
	$y = (6x-2)^2 * \sin(12x-4) + 6(x-0.5)$
	$y = 0.6(6x-2)^2 * \sin(12x-4) + 12(x-0.5) + 6$
N800C4	$y = (6x-2)^2 * \sin(12x-4)$
	$y = (6x-2)^2 * \sin(12x-4) + 6(x-0.5)$
	$y = (6x-2)^2 + 10$
	$y = 0.6(6x-2)^2 * \sin(12x-4) + 12(x-0.5) + 6$

TABLE VI. EFFECT OF NOISE ON THE CLUSTERING PERFORMANCE (MS)

Noise	*k*-IM	FCM	KM	GMM
-	0.027	0.248	0.391	0.467
[-0.25,0.25]	0.029	0.246	0.404	0.432
[-0.50,0.50]	0.038	0.247	0.334	0.393
[-0.75,0.75]	0.040	0.248	0.390	0.413
[-1.00,1.00]	0.044	0.252	0.360	0.397

TABLE VII. EFFECT OF NOISE ON THE CLUSTERING PERFORMANCE (ARI)

Noise	*k*-IM	FCM	KM	GMM
-	0.894	0.251	0.084	0.080
[-0.25,0.25]	0.887	0.256	0.111	0.072
[-0.50,0.50]	0.854	0.255	0.173	0.116
[-0.75,0.75]	0.846	0.252	0.122	0.094
[-1.00,1.00]	0.830	0.245	0.139	0.107

TABLE VIII. EFFECT OF NOISE ON THE CLUSTERING PERFORMANCE (NMI)

Noise	k-IM	FCM	KM	GMM
-	0.839	0.214	0.077	0.067
[-0.25,0.25]	0.819	0.216	0.096	0.060
[-0.50,0.50]	0.774	0.216	0.147	0.094
[-0.75,0.75]	0.764	0.214	0.105	0.081
[-1.00,1.00]	0.741	0.209	0.120	0.087

The performance of the k-IM algorithm is better than the other popular clustering algorithms even if the dataset has noise in the relationship between the response of interest and input variables. The MS of the k-IM algorithm is smaller than 0.05, which means that less than five percent of the data are misclassified. Similar results can be found in the experimental results of the performance index ARI and NMI. With the noise level increasing, the performance of the k-IM algorithm decreases. When the dataset has higher noise in the relationship between the response of interest and input variables, the Kriging method is more difficult to accurately evaluate the relationship. Thus, the performance of the proposed algorithm is worse when the noise level is higher. But, the MS of the k-IM algorithm is still smaller than 0.045. The k-IM algorithm can produce competitive clustering results for the datasets tested in this section.

V. EXPERIMENTS ON ENGINEERING DATASETS

In this section, three engineering datasets are used to further test the proposed algorithm. Since the classification information of the engineering datasets is unknown, the experiments are conducted as follows. For each engineering dataset, the dataset is first clustered into several subsets. For each subset, five cross-validation method is used to test whether the data in the same subset has a similar relationship between the response of interest and input variables. The subset is randomly divided into five parts, one part is selected as the testing data, and the remaining four parts are used as the training data. The experiments are conducted five times, and the average R-square of the five experiments is used to assess the consistency of the relationship of the subset. R-square is calculated as follows.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (15)$$

where n is the number of the testing data, y_i is the real response, $\hat{y}_i$ is the estimate response, and $\bar{y}$ is the mean of the real responses. The closer R^2 to 1, the better performance.

A. Yacht Hydrodynamics Dataset

The yacht hydrodynamics dataset is first used. The dataset comes from a series of 308 experiments on the residuary resistance of sailing yachts [23]. Severn input variables are considered, including the prismatic coefficient, longitudinal position of the center of buoyancy, length-displacement ratio, beam-draught ratio, length-beam ratio, and Froude number. The residuary resistance is evaluated through the per unit weight of displacement. The k-IM, FCM, KM, and GMM algorithms are used to cluster the yacht hydrodynamics dataset into two subsets. And, five cross-validation methods are applied to each subset to test whether the data in the same subset has a similar relationship between the residuary resistance and input variables. The experiments are conducted 30 times, and the corresponding results are shown in Fig. 5. The average R^2 of the k-IM algorithm is higher than 0.98 for the obtained two clusters, which is much better than that of the FCM, KM, and GMM algorithms, indicating that the data of the same cluster obtained by the proposed algorithm have more similar relationship between the response of interest and input variables. The k-IM algorithm is able to cluster the yacht hydrodynamics dataset according to the relationship between the residuary resistance and input variables.

B. Bolt Tensioner Dataset

Bolt tensioner is a widely used tensioning tool in the assembly of large equipment such as nuclear power generators or the construction of large buildings [24]. It is an annular jack that rises up through hydraulic pressure. The bolt tensioner dataset recorded the data from 40 simulations, including the maximum stress at the piston of the bolt tension and the corresponding structural parameters with the same hydraulic pressure. In the experiment, the cluster number is set two as well, and the k-IM, FCM, KM and GMM algorithms are used to cluster the dataset. Based on the clustering results of each clustering algorithm, five cross-validation methods are to test whether the data in the same cluster has a similar relationship between the maximum stress and structural parameters. Fig. 5 shows the experimental results. It is noted that the GMM algorithm cannot provide clustering results since the covariance matrix is ill. From Fig. 6, it can be seen that the average R^2s of the k-IM algorithm is the highest among the tested four clustering algorithms. The k-IM algorithm is able to cluster the bolt tensioner data such that the data in the same cluster have a similar relationship between the maximum stress and structural parameters.

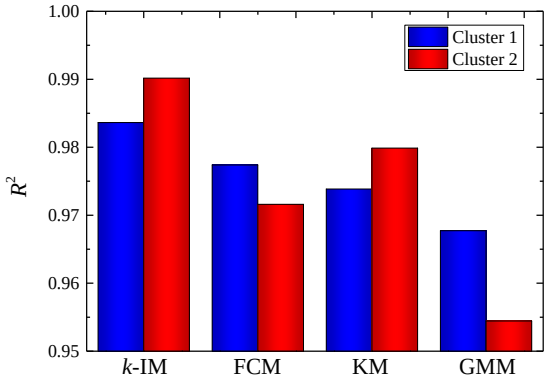

Fig. 5. Experimental Results of Yacht Hydrodynamics Dataset.

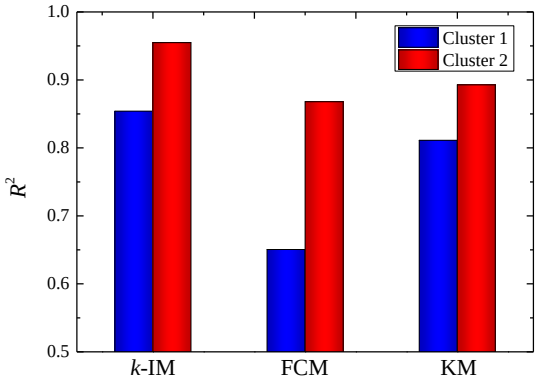

Fig. 6. Experimental Results of Bolt Tensioner Dataset.

VI. Conclusion

In this work, we proposed a *k*-interpolation model clustering algorithm (named *k*-IM) to cluster data according to the relationship between the response of interest and input variables. In the proposed algorithm, Kriging method is used to construct the interpolation models. For each datum, the estimation errors of the interpolation models of the clusters are used to decide its assignment. An optimization strategy is designed to obtain the clustering results under the framework of *k*-means algorithm. The effect of the sample number, cluster number, and noise level on the *k*-IM algorithm is studied through several synthetic datasets. The results indicate that the *k*-IM algorithm in this paper can provide competitive clustering results. Two engineering datasets are further to test the performance of the *k*-IM algorithm as well, and the experimental results show that the *k*-IM algorithm is able to cluster the data such that the data in the same part have a similar relationship between the response of interest and input variables.

Acknowledgment

The authors would like to thank the editors' and reviewers' work on this paper.

References

[1] I. Škrjanc, J. Iglesias, A. Sanchis, D. Leite, E. Lughofer, and F. Gomide, "Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification: A survey," Information Sciences, vol. 490, pp. 344-368, 2019.

[2] A. Dubey and A. Rasool, "Clustering-based hybrid approach for multivariate missing data imputation," International Journal of Advanced Computer Science and Applications, vol. 11, pp. 710-714, 2020.

[3] X. Song, M. Shi, J. Wu, and W. Sun, "A new fuzzy c-means clustering-based time series segmentation approach and its application on tunnel boring machine analysis," Mechanical Systems and Signal Processing, vol. 133, pp. 106279, 2019.

[4] R. Xu and D. Wunsch, "Survey of clustering algorithms," IEEE Transactions on neural networks, vol. 16, pp. 645-678, 2005.

[5] M. Shi, T. Zhang, L. Zhang, W. Sun, and X. Song, "A fuzzy c-means algorithm based on the relationship among attributes of data and its application in tunnel boring machinem," Knowledge-Based Systems, vol. 191, pp. 105229, 2020.

[6] J. Diaz-Rozo, C. Bielza, and P. Larrañaga, "Clustering of data streams with dynamic gaussian mixture models: an IoT application in industrial processes," IEEE Internet of Things Journal, vol. 5, pp. 3533-3547, 2018.

[7] A. Shubair, and A. Al-Nassiri, "kEFCM: kNN-based dynamic evolving fuzzy clustering method," Proc. IJACSA. vol. 6, pp. 5-13, 2015.

[8] A. Aldino, D. Darwis, A. Prastowo, and C. Sujana, "Implementation of K-means algorithm for clustering corn planting feasibility area in south lampung regency," Journal of Physics: Conference Series. vol. 1751, pp. 012038, 2021.

[9] S. Yu, S. Chu, C. Wang, Y. Chan, and T. Chang, "Two improved k-means algorithms," Applied Soft Computing, vol. 68, pp. 747-755, 2018.

[10] E. Zhu, Y. Zhang, P. Wen, and F. Liu, "Fast and stable clustering analysis based on Grid-mapping K-means algorithm and new clustering validity index," Neurocomputing, vol. 363, pp. 149-170, 2019.

[11] S. Cuomo, V. Angelis, G. Farina, L. Marcellino, and G. Toraldo, "A GPU-accelerated parallel K-means algorithm," Computers & Electrical Engineering, vol. 75, pp. 262-274, 2019.

[12] B. Echard, N. Gayton, M. Lemaire, and N. Relun, "A combined importance sampling and Kriging reliability method for small failure probabilities with time-demanding numerical models," Reliability Engineering & System Safety, vol. 111, pp. 232-240, 2013.

[13] B. Keshtegar, C. Mert, and O. Kisi, "Comparison of four heuristic regression techniques in solar radiation modeling: Kriging method vs RSM, MARS and M5 model tree," Renewable and sustainable energy reviews, vol. 81, pp. 330-341, 2018.

[14] M. Wojciech, "Kriging method optimization for the process of DTM creation based on huge data sets obtained from MBESs," Geosciences, vol. 8, pp. 433, 2018.

[15] L. Belkhiri, A. Tiri, and L. Mouni, "Spatial distribution of the groundwater quality using kriging and Co-kriging interpolations," Groundwater for Sustainable Development, vol. 11, pp. 100473, 2020.

[16] C. Peng, Q. Zhang, Z. Kang, C. Chen, and Q. Cheng, "Kernel two-dimensional ridge regression for subspace clustering", Pattern Recognition, vol.113, pp.107749, 2021.

[17] Y. Chen and Z. Yi, "Locality-constrained least squares regression for subspace clustering," Knowledge-Based Systems, vol.163, pp.51-56, 2019.

[18] S. Blažič and I. Škrjanc, "Hybrid system identification by incremental fuzzy c-regression clustering," In 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp.1-7, 2020.

[19] J. N. Fuhg and A. Fau, "A classification-pursuing adaptive approach for Gaussian process regression on unlabeled data," Mechanical Systems and Signal Processing, vol.162, pp.107976, 2022.

[20] J. Fang, X. Song, N. Yao, and M. Shi, "Application of FCM Algorithm combined with artificial neural network in TBM operation data," Computer Modeling in Engineering & Sciences, vol.126, pp.397-417, 2021.

[21] K.Y. Yeung and W. L. Ruzzo, "Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data," Bioinformatics, vol. 17, pp. 763-774, 2001.

[22] P.A. Estévez, M. Tesmer, C.A. Perez, and J.M. Zurada, "Normalized mutual information feature selection," IEEE Transactions on neural networks, vol. 20, pp. 189-201, 2009.

[23] I. Ortigosa, R. Lopez, and J. Garcia, "A neural networks approach to residuary resistance of sailing yachts prediction," In Proceedings of the international conference on marine engineering MARINE, vol. 2007, pp. 250, 2007.

[24] J. Liu, Z. Zhang, M. Wang, J. Wang, and S. He, "Main bolt tensioner for pressure vessel of 10 MW high temperature reactore," Nuclear Power Engineering, vol. 21, pp. 503-506, 2000.