A clustering method based on gaussian process regression and fuzzy *c*-means algorithm

Maolin Shi[*1] and Zihao Wang[2]

[*1] *School of Agricultural Engineering, Jiangsu University, Zhenjiang, 212013, China*
E-mail: shl5985336@126.com
[*2]*International School of Information Science & Engineering, Dalian University of Technology, Dalian, 116024, China*
E-mail: iswzh@outlook.com

Abstract. With the development of measurement techniques, massive data have been recorded from real-world systems, and partitioning these data can provide important information and useful references. In this paper, a new clustering method named GPR-FCM is proposed to accomplish this work. The proposed clustering method is developed based on fuzzy *c*-means algorithm. The regression relationship of data instead of the distance among the objects is utilized to evaluate the difference between the clusters, and Gaussian process regression (GPR) is used to evaluate the regression relationship of each cluster. A series of experiments on synthetic and engineering datasets are used to evaluate the performance of the GPR-FCM method. The results demonstrate higher effectiveness and advantages of the GPR-FCM method compared with conventional data clustering algorithms.

Keywords: Data clustering, fuzzy *c*-means algorithm, Gaussian process regression

1. INTRODUCTION

With the development of cyber-physical systems and measurement techniques, massive in-situ data of real-world systems are recorded. However, the patterns of these in-situ data are usually different, so it is necessary to partition them such that the characteristics of data in the same part are similar than those in the other parts. Data clustering is a mature tool to solve this issue. Data clustering is a class of data mining techniques that has been widely used in data analysis such as fault detection, image recognition, modal analysis, and risk analysis [1-3]. It is the task of finding natural partitioning within a dataset such that patterns within the same cluster are more similar than those within different clusters. Fuzzy *c*-means algorithm (FCM) is one of the most popular clustering methods. FCM clusters data according to their spatial distribution, and the spatial distribution of real-world data of different clusters is often very similar. Thus, it is necessary to utilize other patterns to improve the data clustering accuracy of FCM. For real-world systems, the regression relationship of in-situ data often varies greatly, which has considerable potentiality to improve the clustering validity. Thus, the clustering method based on the regression relationship of data is with great potential to realize the accurate data clustering results. The key step of data clustering based on the regression relationship is the evaluation method of regression relationship. In recent years, Gaussian process regression (GPR) has been widely used to describe the regression relationship of data [4]. Kumar et al. [5] used Gaussian process regression to model the regression relationship between the elastic modulus of jointed rock mass and the other six geological parameters. Pasolli et al. [6] utilized Gaussian process regression to estimate the chlorophyll concentration in subsurface water. Liu et al. [7] used Gaussian process regression to learn the degradation pattern of a lithium battery system based on the historical data and to predict its remaining useful life. Celaya et al. [8] collected the run-to-failure data of metal oxide field effect transistor (MOSFET) and used Gaussian process regression to build the failure model of MOSFET. Ocak [9] used Gaussian process regression to study the effect of operation parameters of a tunnel boring machine on the surface settlements. Hewing [10] proposed a model predictive control approach that integrates a nominal system with an additive nonlinear part of the dynamics model using Gaussian process regression. Zeng [11] used Gaussian process regression to predict the building electricity usage based on the large-scale real data collected from the building energy management system. Parveen [12] used Gaussian process regression to model the infiltration of sandy soil and compared its prediction performance between Pearson VII and radial based kernel function. Yun [13] proposed a novel methodology for predicting human gait pattern kinematics based on Gaussian process regression. The above-mentioned works indicate Gaussian process regression can effectively learn the regression relationship of data. Therefore, Gaussian process regression is utilized to evaluate the regression relationship of each cluster in this paper, and a new clustering method based on Gaussian process regression is proposed.

The rest of this paper is organized as follows. Section 2 describes the background of FCM and GPR, and Section 3 introduces the proposed clustering method. In Section 4, synthetic datasets are used to test and compare the performance of the proposed method with three conventional clustering algorithms FCM, K-means and Expectation-Maximization algorithm (EM). Two engineering datasets are used to validate the proposed clustering method in Section 5. Some conclusions are given in Section 6.

2. BACKGROUND

2.1. Fuzzy *c*-means Algorithm

Fuzzy *c*-means algorithm partitions a given set of object data $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\} \subset \mathbb{R}^{d\times n}$ into *c* fuzzy clusters by minimizing an objective function $J(\mathbf{U},\mathbf{V})$ as follows:

$$J(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}\|\boldsymbol{x}_k - \boldsymbol{v}_i\|_2^2 \quad (1)$$

where $\boldsymbol{x}_k = [x_{1,k}, x_{2,k}, \ldots, x_{j,k}, \ldots, x_{d,k}]^{\mathrm{T}}$ is an object datum, and $x_{j,k}$ is the j-th attribute value of $\boldsymbol{x}_k$; $\boldsymbol{v}_i$ is the i-th cluster prototype, and let matrix of cluster prototypes $\mathbf{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_c]^{\mathrm{T}} \in \mathbb{R}^{c\times d}$; m is a fuzzification parameter, $m \in (1, \infty)$; $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^d$; u_{ik} is the membership that represents the degree which $\boldsymbol{x}_k$ belongs to the i-th cluster and satisfies the following condition:

$$\sum_{i=1}^{c} u_{ik} = 1 \ (k = 1, 2, \ldots, n;\ \forall i, k: u_{ik} \in [0,1]) \quad (2)$$

and let partition matrix $\mathbf{U}= [u_{ik}] \in \mathbb{R}^{c\times n}$.

The necessary conditions for minimizing (1) with the constraint (2) result in the following iterative update formulas for the prototypes and the partition matrix [14]:

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^{m}\boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ik}^{m}} (i = 1, \ldots, c) \quad (3)$$

$$u_{ik} = \left[\sum_{t=1}^{c}\left(\frac{\|\boldsymbol{x}_k - \boldsymbol{v}_i\|_2^2}{\|\boldsymbol{x}_k - \boldsymbol{v}_t\|_2^2}\right)^{\frac{2}{m-1}}\right]^{-1} (i = 1, \ldots, c; k = 1, \ldots, n) \quad (4)$$

The iterations are carried out until the changes in the values of the partition matrix reported in consecutive iterations are lower than a certain predetermined threshold.

2.2. Gaussian Process Regression

GPR is briefly introduced in this section, as it is used to learn the relationship among attributes in this paper. Let the i-th input and response of interest be denoted by a dimensional vector, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$, and its response, $y_i = f(\boldsymbol{x}_i)$, respectively. The input set is denoted as $\boldsymbol{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n\}$, and the outputs are held in the *n*-dimensional vector $\boldsymbol{Y} = [y_1, \ldots, y_n]^{\mathrm{T}}$. The output of Gaussian process regression modelled as

$$f(\boldsymbol{x}_i) = \mu + z(\boldsymbol{x}_i)\ (i = 1, \ldots, n) \quad (5)$$

where μ is the overall mean, and $z(x_i)$ is a Gaussian process with $E\big(z(\boldsymbol{x}_i)\big) = 0$, $Var\big(z(\boldsymbol{x}_i)\big) = \sigma_z^2$, and $Cor\left(z(\boldsymbol{x}_i), z\big(\boldsymbol{x}_j\big)\right) = \sigma_z^2 R_{ij}$. In GPR, $f(\boldsymbol{D})$ has a multivariate normal distribution as follows:

$$f(\boldsymbol{D}) \sim N(\mathbf{1}_n\mu, \Sigma) \quad (6)$$

where $\Sigma = V\big(\boldsymbol{D}|f(\boldsymbol{D})\big) = \sigma_z^2\boldsymbol{R}$, $\boldsymbol{R}$ is the Gaussian correlation function, and $\mathbf{1}_n$ is an $n\times1$ vector of all ones. In this paper, the following correlation function is used:

$$R_{ij} = e^{-\sum_{k=1}^{d}\theta_k\|x_{i,k} - x_{j,k}\|_2^2} \quad (7)$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_d]^{\mathrm{T}}$ is the vector of hyperparameters. The model in (5) with the correlation function in (7) is used to predict responses of any new input $\boldsymbol{x}^*$. Following the maximum likelihood approach, the best linear unbiased predictor at $\boldsymbol{x}^*$ is

$$\hat{f}(\boldsymbol{x}^*) = \hat{\mu} + \boldsymbol{r}'\boldsymbol{R}^{-1}(\boldsymbol{Y} - \mathbf{1}_n\hat{\mu}) \quad (8)$$

with mean squared error

$$s^2(\boldsymbol{x}^*) = \sigma_z^2\left(1 - \boldsymbol{r}'\boldsymbol{R}^{-1}\boldsymbol{r} + \frac{(1 - \mathbf{1}_n'\boldsymbol{R}^{-1}\boldsymbol{r})^2}{\mathbf{1}_n'\boldsymbol{R}^{-1}\mathbf{1}_n}\right) \quad (9)$$

Given the hyperparameters $\boldsymbol{\theta}$, the closed form estimators of μ and σ_z^2 are as follows:

$$\hat{\mu}(\boldsymbol{\theta}) = (\mathbf{1}_n'\boldsymbol{R}^{-1}\mathbf{1}_n)^{-1}(\mathbf{1}_n'\boldsymbol{R}^{-1}\boldsymbol{Y}) \quad (10)$$

$$\hat{\sigma}_z^2(\boldsymbol{\theta}) = \frac{(\boldsymbol{Y} - \mathbf{1}_n\hat{\mu}(\boldsymbol{\theta}))'\boldsymbol{R}^{-1}(\boldsymbol{Y} - \mathbf{1}_n\hat{\mu}(\boldsymbol{\theta}))}{n} \quad (11)$$

The optimal hyperparameters $\boldsymbol{\theta}$ can be estimated using the following log-likelihood

$$log(|\boldsymbol{R}|) + nlog[((\boldsymbol{Y} - \mathbf{1}_n\hat{\mu}(\boldsymbol{\theta}))'\boldsymbol{R}^{-1}(\boldsymbol{Y} - \mathbf{1}_n\hat{\mu}(\boldsymbol{\theta})))] \quad (12)$$

where $|\boldsymbol{R}|$ is the determinant of $\boldsymbol{R}$. Recall from (7) that the correlation matrix $\boldsymbol{R}$ depends on $\boldsymbol{\theta}$ and $\boldsymbol{D}$.

3. PROPOSED CLUSTERING METHOD

In this section, a new clustering method based on Gaussian process regression is proposed. Generally, the objective function $J(\mathbf{U},\mathbf{V})$ of fuzzy clustering algorithms can rewritten as follows:

$$J(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} D_{ik}^2 \quad (13)$$

where D_{ik} denotes the distance between k-th datum and i-th cluster prototype. The necessary conditions for minimizing (13) with the constraint (4) result in the following partition matrix:

$$u_{ik} = \left[\sum_{t=1}^{c}\left(\frac{D_{ik}}{D_{tk}}\right)^{\frac{2}{m-1}}\right]^{-1} (i = 1, \ldots, c\ \&\ k = 1, \ldots, n) \quad (14)$$

To cluster the data based on their regression relationship, the D_{ik} in (14) should represent the difference among the regression relationships of different clusters. GPR is used to evaluate the regression relationship of i-th cluster as follows:

$$\widehat{y_{k,i}} = GPR(\boldsymbol{z}_k)_i \quad (15)$$

where $\widehat{y_{k,i}}$ is the estimation of output attribute of k-th datum of i-th cluster, $\boldsymbol{z}_k$ is the vector of input attributes, and $GPR(\cdot)_i$ is the GPR model of i-th cluster. If a datum belongs to i-th cluster, the estimated $\widehat{y_{k,i}}$ of i-th GPR model based on the data of i-th cluster is closer to the real y_k than those estimated by other GPR models. Thus, the distance D_{ik} in (13) is modified as follows:

$$D_{ik} = (y_k - \widehat{y_{k,i}})^2 \quad (16)$$

Thus, the objective function $J(\mathbf{U})$ can be modified as follows:

$$J(\mathbf{U}) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}\,(y_k - \widehat{y_{k,i}})^2 \quad (17)$$

where $\widehat{y_{k,i}} = GPR(\boldsymbol{z}_k)_i$, $\boldsymbol{z}_k$ represent a new vector composed of the remaining attributes after removing the output attribute of k-th datum. As the proposed objective function is still under the framework of FCM, the necessary conditions for minimizing (17) with the constraint (2) result in the following partition matrix:

$$u_{ik} = \left[\sum_{t=1}^{c}\left(\frac{(y_k - \widehat{y_{k,i}})^2}{(y_k - \widehat{y_{k,t}})^2}\right)^{\frac{2}{m-1}}\right]^{-1} \quad (18)$$

Comparing (18) with (4), the i-th cluster prototype $\boldsymbol{v}_i$ in (4) is replaced by i-th GPR model based on the data of i-th cluster. To ensure that the i-th GPR model can learn the regression relationship among attributes accurately, only the data with relatively high membership are used to construct the GPR model in each iteration. In this paper,

Table. 1 Details of synthetic datasets

Data set	Cluster	Regression relationship	D
D400A2C2F1	1	$y = sin^2(0.5z_1 - 6)\,(z_1/4)^2 + 10$	[0, 30]
	2	$y = sin^2(0.5z_1 - 6)(z_1/5)^2$	
D400A2C2F2	1	y =6-1/$((z_1 - 0.5)^2 + 0.03)$-1/$((z_1 - 0.9)^2 + 0.04)$	[0, 1]
	2	y =4-1/$((z_1 - 0.5)^2 + 0.03)$-1/$((z_1 - 0.9)^2 + 0.04)$	
D400A3C2F1	1	$y = (4 + z_1 sin(z_1))(4 + exp(-z_2^2))$	[-2, 2]
	2	$y = (6 + z_1 sin(z_1))(4 + exp(-z_2^2))$	
D400A3C2F2	1	$y = z_1 exp(-z_1^2 - z_2^2)$	[-2, 2]
	2	$y = z_1 exp(-z_1^2 - z_2^2) + 1$	
D400A4C2F1	1	$y = sin(z_2 - z_1) + (z_3 - z_2)^2$	[1, 2]
	2	$y = cos(z_2 - z_1) + (z_3 - z_2)^2$	
D400A4C2F2	1	$y = -z_1 z_2 z_3/6$	[0, 2]
	2	$y = -\frac{z_1 z_2 z_3}{6} + 0.5z_1^2 + 0.5$	
D400A5C2F1	1	$y = 0.8(10\,sin(z_1\pi) + \sum_{i=1}^{4} z_i + z_1 z_2 + z_3 z_4)$	[0, 1]
	2	$y = 0.6(10\,sin(z_1\pi) + \sum_{i=1}^{4} z_i + z_1 z_2 - z_3 z_4) + 0.75(z_1^2 + z_2^2) + 4.5$	
D400A5C2F2	1	$y = 1 + 2exp(-2((z_1 - 1)^2 + z_2^2) - 0.5(z_3^2 + z_4^2))$	[0, 1]
	2	$y = 1 + exp(-2((z_1 - 1)^2 + z_2^2) - 0.5(z_3^2 + z_4^2))$	
D400A6C2F1	1	$y = \sum_{i=1}^{5}[\frac{3}{10} + sin\left(\frac{15}{16}z_i - 1\right) + sin^2(\frac{15}{16}z_i - 1)]$	[-1, 1]
	2	$y = \sum_{i=1}^{5}[\frac{3}{10} + sin\left(\frac{15}{16}z_i - 1\right) + cos^2(\frac{15}{16}z_i - 1)]$	
D400A6C2F2	1	$y = \sum_{i=1}^{4}[(z_{i+1} - z_i)^2 + (z_i - 1)^2]$	[-1, 1]
	2	$y = \sum_{i=1}^{4}[(z_{i+1} - z_i)^2 + (z_i - 1)^2] + 5$	
D400A7C2F1	1	$y = \sum_{i=1}^{5}[(z_{i+1} - x_i)^2 + (z_i - 1)^2]$	[-1, 1]
	2	$y = \sum_{i=1}^{5}[(z_{i+1} - z_i)^2 + (z_{i+1} - 1)^2] + 6$	
D400A7C2F2	1	$y = \prod_{i=1}^{6} z_i + \sum_{j=1}^{6}[ln(z_i - 2)^2 + ln(10 - z_i)^2]$	[-1, 1]
	2	$y = \prod_{i=1}^{6} z_i + \sum_{j=1}^{6}[ln(z_i - 2)^2] + 25$	

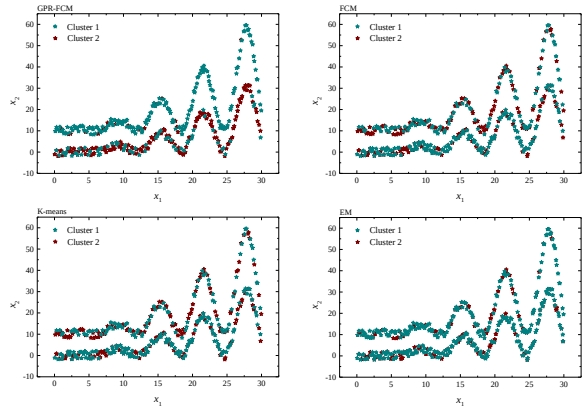

Fig. 1 Clustering results of datasets D400A2C2F1

the membership criterion is set as $\mathbf{U}[\,,i] \geq 0.8$. Considering FCM has shown benefits the data mining and knowledge discovery of real-world data [15], the clustering results of FCM is also utilized in the proposed method. The total procedure of the proposed clustering method is described as follows.

Step (1) Setting the clustering number c, the objective attribute y, and the iteration number $g = 1$;

Step (2) Using FCM to cluster data and the obtained membership matrix $\mathbf{U}_{\mathrm{FCM}}$ are used as the initial membership matrix $\mathbf{U}^{(\mathbf{0})}$;

Step (3) Selecting the data in i-th cluster with $\mathbf{U}[\,,i] \geq 0.8$ and obtaining i-th cluster dataset;

Step (4) Using i-th cluster dataset as training data to obtain i-th GPR model;

Step (5) Using the GPR models from Step (4) to get the responses of all the data and creating the responses matrix $\mathbf{Y}_{\boldsymbol{n}\times\boldsymbol{c}}$;

Step (6) Calculating the partition matrix $\mathbf{U}^{(g)}$ using (18);

Step (7) If $\forall\, i,k$: $\max\left|u_{ik}^{(g)} - u_{ik}^{(g-1)}\right| < \varepsilon$, then stop and get partition matrix $\mathbf{U}_{\mathbf{GPR}}$, otherwise set $g = g + 1$ and return to Step (3).

4. EXPERIMENTS ON SYNTHETIC DATASETS

To assess the validity of the proposed method, several synthetic datasets are used in this section. The synthetic datasets are created as follows. In each cluster, the input attributes are randomly sampled using Latin hypercube sampling method ("*randomLHS*" in **R**-package "*lhs*"). The output attribute of each object is calculated according to the setting regression relationships and then added with a random relative error from -5% to $+5\%$ as the final value. After that, the data of different clusters are combined as the obtained dataset. Each dataset is given a denomination by the number of object data, attributes, clusters and relationship among attributes. For instance, N400A2C2F1 denotes that the dataset contains 400 object data and can be divided evenly into two clusters, and A2 denotes the dataset has two attributes. The clustering performance of the proposed method is compared with three popular clustering methods, FCM, K-means and Expectation-Maximization algorithm (EM). The proposed method and FCM are coded by the authors using ***R***, K-means is using "*pam*" from ***R***-package "*cluster*" [16], and EM is using "*emcluster*" from ***R***-package "*EMCluster*" [17]. The clustering performance is evaluated in terms of misclassification (*MS*), which is calculated as follows:

$$MS = \frac{N_{error}}{N_{total}} \times 100\% \qquad (19)$$

where N_{error} is the number of misclassified data; N_{total} is the total number of data.

In the experiments, the parameters of proposed method and FCM are set as follows: the fuzzification parameter m is 2, the threshold value ε is 10^{-6}, and the maximum iteration is 100. The settings of K-means and EM are the default parameters of their ***R***-packages.

4.1. Effect of Attribute Number

12 synthetic datasets are used to test the performance of the GPR-FCM method. The datasets have different attributes, and the two clusters of each dataset have similar but different regression relationships. The relationship of each dataset is selected from the benchmark functions in [18] which are widely used to test the performance of regression approaches such as GPR. The details are listed in Table 1.

To illustrate the characteristics of the synthetic datasets used in this section, dataset D400A2C2F1 is shown in Fig. 1, where the object data belonging to two different clusters are represented by wine and cyan stars. It can be found that the two clusters have similar spatial

distribution, but their attributes have different regression relationships. The experiments on each dataset are conducted 30 times, and the average experimental results are shown in Table 2. It can be found that the GPR-FCM method produces much better clustering results than the

Table. 2 Clustering Results of Synthetic Datasets

Data set	GPR-FCM	FCM	K-means	EM
D400A2C2F1	**0.00%**	39.00%	39.25%	46.25%
D400A2C2F2	**1.75%**	47.75%	51.50%	48.00%
D400A3C2F1	**0.00%**	13.75%	13.00%	28.00%
D400A3C2F2	**21.50%**	46.50%	46.25%	48.25%
D400A4C2F1	**23.75%**	27.75%	30.50%	38.50%
D400A4C2F2	**0.00%**	13.00%	24.75%	1.50%
D400A5C2F1	**0.00%**	33.75%	33.75%	49.50%
D400A5C2F2	**1.75%**	3.75%	6.00%	49.50%
D400A6C2F1	17.25%	19.50%	26.50%	**5.50%**
D400A6C2F2	**3.75%**	26.25%	26.25%	18.00%
D400A7C2F1	**13.00%**	20.75%	24.75%	13.75%
D400A7C2F2	**21.25%**	26.00%	24.50%	47.25%

FCM, K-means and EM for most datasets. The highest misclassification rate of the GPR-FCM method is 23.75%, and the misclassification rates are mostly lower than 20.00% and even close to 0.00%. For dataset D400A6C2F1, the misclassification rate of the GPR-FCM method is 17.25% which is higher than EM. But EM achieves around 50% of misclassification rate for six datasets. It can be concluded that the GPR-FCM method exhibits much better results than EM for the 12 synthetic datasets in this section. To further show the difference between the proposed method and other clustering algorithms, the clustering results of dataset D400A2C2F1 is shown in Fig. 2. It is observed that the GPR-FCM method is clustering the data according to the regression relationship, but the other methods are based on spatial distribution. Thus, the GPR-FCM method can provide much better clustering performance than other methods.

4.2. Effect of Sample Number

24 synthetic datasets with different samples are used to test the performance of the GPR-FCM method on the number of samples. The datasets used here are based on some datasets in Section 4.1, but their samples 300, 400, 500 and 600 samples, respectively. The experiments on each dataset are conducted 30 times, and the average experimental results are shown in Table 3. From this table, it is observed that the GPR-FCM method produces better clustering results than the other clustering algorithms for most datasets. FCM and K-means provide similar clustering performance for the datasets with same regression relationship but with different sample numbers, but the clustering results of EM are different with them. The reason is that FCM and K-means cluster data according to their spatial distance but EM is based on their statistical distribution. With the sample number changing, the center of each cluster does not vary greatly but the statistical distribution might change greatly. Thus, FCM and K-means provide similar results, but the clustering results of EM are different. For datasets D3/4/5/600A2C2F1 and D3/4/5/600A3C2F1, their misclassification rates given by the GPR-FCM method are 0.00%. For datasets D3/4/5/600A4C2F1, D3/4/5/600A5C2F1 and D3/4/5/600A7C2F1, the GPR-FCM method tends to provide better clustering results with sample number increasing. When the dataset has more samples, more data can be used to build the GPR model of each cluster. The GPR models are closer to the real regression relationship, so the clustering results are improved. For dataset D300/400/500/600A6C2F1, the misclassification rates of the GPR-FCM method ranges from 6.80% to 19.67% which is smaller than EM. Considering EM achieves much higher misclassification rate for the other datasets than the GPR-FCM method, the GPR-FCM method still exhibits much better clustering performance than EM for the synthetic datasets tested in this sub-section.

Table. 3 Effect of sample number on the clustering performance

Data set	GPR-FCM	FCM	K-means	EM
D300A2C2F1	**0.00%**	39.67%	39.33%	47.33%
D400A2C2F1	**0.00%**	39.00%	39.25%	46.25%
D500A2C2F1	**0.00%**	38.80%	39.20%	38.40%
D600A2C2F1	**0.00%**	38.83%	35.67%	37.00%
D300A3C2F1	**0.00%**	14.67%	15.67%	41.00%
D400A3C2F1	**0.00%**	13.75%	13.00%	28.00%
D500A3C2F1	**0.00%**	15.20%	13.40%	35.60%
D600A3C2F1	**0.00%**	15.00%	12.33%	28.50%
D300A4C2F1	**24.33%**	30.33%	30.67%	47.00%
D400A4C2F1	**23.75%**	27.75%	30.50%	38.50%
D500A4C2F1	**23.70%**	27.00%	30.40%	40.20%
D600A4C2F1	**10.50%**	28.00%	39.83%	39.50%
D300A5C2F1	**29.00%**	30.33%	30.67%	51.33%
D400A5C2F1	**0.00%**	33.75%	33.75%	49.50%
D500A5C2F1	**0.00%**	29.80%	29.40%	33.60%
D600A5C2F1	**0.00%**	31.83%	31.83%	33.50%
D300A6C2F1	19.67%	20.33%	29.33%	**5.33%**
D400A6C2F1	17.25%	19.50%	26.50%	**5.50%**
D500A6C2F1	6.80%	22.80%	25.80%	**6.80%**
D600A6C2F1	11.33%	22.33%	30.50%	**8.33%**
D300A7C2F1	**18.00%**	23.67%	23.67%	22.00%
D400A7C2F1	**13.00%**	20.75%	24.75%	13.75%
D500A7C2F1	**2.40%**	25.20%	24.40%	19.00%
D600A7C2F1	**3.50%**	22.67%	22.83%	16.67%

4.3. Effect of Noise

24 synthetic datasets are used to test the performance of the GPR-FCM method on the noise of regression relationship. The noise level is set as ±2.5%, ±5.0%, ±7.5%, and ±10.0%, respectively. 30 times' experiments are conducted on each dataset, and the average experimental results are shown in Table 4. From this table, it is observed that the GPR-FCM method produces better clustering results than FCM, K-means and EM. For datasets D400A4C2F2, the misclassification rates for different levels of noise given by the GPR-FCM method are all 0.00%. The clustering performance of FCM and K-means is similar with each other, but is much different with EM. From the theory of FCM, K-means and EM, it is known that FCM and K-means cluster the data according to their spatial distribution but EM is based on their statistical distribution. With the level of noise increasing, the spatial distribution of each cluster changes slightly but the statistical distribution might change greatly. Thus, FCM and K-means provide similar results, but EM does not. In addition, it is also observed that misclassification rate of the GPR-FCM method tends to increase with the level of noise increasing, which is mainly because that the overlap between the two clusters

is increasing with level of noise increasing. The GPR-FCM method is still able to provide better clustering results than the other methods when clustering the datasets with different levels of noise.

Table. 4 Effect of noise on the clustering performance

Data set	Noise	GPR-FCM	FCM	K-means	EM
D400A2C2F2	±2.5%,	**0.00%**	47.75%	47.75%	48.00%
D400A2C2F1	±5.0%,	**1.75%**	47.75%	48.50%	48.00%
D500A2C2F1	±7.5%,	**8.75%**	48.00%	48.50%	48.75%
D600A2C2F1	±10.0%	**10.75%**	47.75%	47.50%	47.75%
D400A3C2F2	±2.5%,	**14.50%**	46.25%	46.00%	48.25%
D400A3C2F1	±5.0%,	**21.50%**	46.50%	46.25%	48.25%
D500A3C2F1	±7.5%,	**24.75%**	47.00%	46.50%	48.25%
D600A3C2F1	±10.0%	**27.00%**	46.75%	45.00%	49.50%
D400A4C2F2	±2.5%,	**0.00%**	20.00%	25.00%	10.25%
D400A4C2F1	±5.0%,	**0.00%**	13.00%	24.75%	1.50%
D500A4C2F1	±7.5%,	**0.00%**	19.75%	19.00%	19.50%
D600A4C2F1	±10.0%	**0.00%**	19.00%	18.00%	2.25%
D400A5C2F2	±2.5%,	**14.50%**	43.75%	36.25%	24.75%
D400A5C2F1	±5.0%,	**10.00%**	44.25%	44.25%	31.75%
D500A5C2F1	±7.5%,	**21.50%**	42.75%	47.00%	30.50%
D600A5C2F1	±10.0%	**23.00%**	40.75%	47.00%	31.00%
D400A6C2F2	±2.5%,	**3.50%**	24.25%	26.75%	19.25%
D400A6C2F1	±5.0%,	**3.75%**	26.25%	26.25%	18.00%
D500A6C2F1	±7.5%,	**4.75%**	26.50%	27.00%	20.25%
D600A6C2F1	±10.0%	**15.50%**	27.00%	28.00%	13.00%
D400A7C2F2	±2.5%,	**14.75%**	21.00%	23.75%	33.75%
D400A7C2F1	±5.0%,	**21.25%**	26.00%	24.50%	47.25%
D500A7C2F1	±7.5%,	**26.75%**	27.50%	28.50%	48.25%
D600A7C2F1	±10.0%	**29.75%**	31.50%	32.00%	35.00%

4.4. Effect of Cluster Number

Three synthetic datasets are designed to test the performance of the GPR-FCM method on cluster numbers in this section. The first dataset D600A4C3 has three clusters, and their regression relationship are similar. The second dataset D800A3C4 has four clusters, and each pair of clusters has similar regression relationship. The last dataset D1000A2C5 has five clusters, but each cluster has its special regression relationship. The details are shown in Table 5. The experiments are conducted 30 times on each dataset, and the average experimental results are shown in Table 6. From this table, it is observed that the GPR-FCM method produces better clustering results than the other clustering methods. The GPR-FCM method can provide competitive clustering results when clustering the datasets with different cluster numbers.

Table. 5 Details of the synthetic datasets with different cluster numbers

Data set	Cluster	Relationship	D
D400A4C3	1	$y = -z_1 z_2 z_3/6$	[0, 2]
	2	$y = -\frac{z_1 z_2 z_3}{6} + 0.5z_1^2 + 0.5$	
	3	$y = -\frac{z_1 z_2 z_3}{6} + 0.5z_2^2 + 0.5z_1^2 + 0.5$	
D400A3C4	1	$y = (4 + z_1 sin(z_1))(4 + exp(-z_2^2))$	[-2, 2]
	2	$y = (6 + z_1 sin(z_1))(4 + exp(-z_2^2))$	
	3	$y = z_1 exp(-z_1^2 - z_2^2)$	
	4	$y = z_1 exp(-z_1^2 - z_2^2) + 1$	
D400A3C45	1	$y = -5/((z_1 - 0.5)^3 + 0.03) - 1/((z_1 - 9)^3 + 0.04)$	[0, 30]
	2	$y = 5sin(0.3 * z_1)$	
	3	$y = 12/exp(-0.1z_1 + 1) * sin(0.5z_1)$	
	4	$y = \frac{36 \sin^2(0.1z_1)}{z_1} - (z_1/30)^2 - 14$	
	5	$y = sin^2(0.5z_1 - 6)(z_1/10)^2 + 20$	

Table. 6 Clustering results of the synthetic datasets with different cluster numbers

Data set	GPR-FCM	FCM	K-means	EM
D400A4C3	**10.17%**	47.83%	37.00%	37.50%
D400A3C4	**9.38%**	49.75%	42.13%	46.76%
D400A2C5	**36.40%**	47.10%	49.10%	47.00%

4.5. Comparison of Overall Performance

In this paper, Wilcoxson Test [19] is used to statistically compare the GPR-FCM method with FCM, K-means, and EM. MS is used to validate the clustering performance for the GPR-FCM method, FCM, K-Means and EM. The clustering results in the Sections 4.1~4.4 are used, and the results of Wilcoxson test are shown in Table 7. It is observed that all P-values of MS are much less than 0.05. The null hypothesizes, the GPR-FCM method is better than FCM, K-means and EM, are accepted. Thus, the performance of the GPR-FCM method is significantly better than FCM.

Table. 7 Results of wilcoxson test

Index	Null hypothesis	Alternative hypothesis	P-value
MS	GPR-FCM is better than FCM	FCM is better than GPR-FCM	2.63E-12
MS	GPR-FCM is better than K-means	K-means is better than GPR-FCM	2.64E-12
MS	GPR-FCM is better than EM	EM is better than GPR-FCM	8.06E-11

5. EXPERIMENTS ON ENGINEERING DATASETS

5.1. Mill Dataset

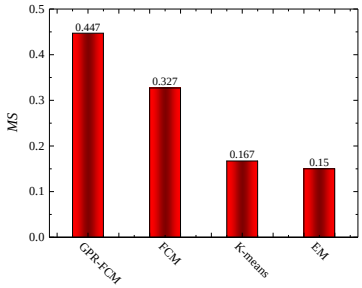

Figure. 2 Clustering results of the Mill dataset

The Mill dataset is from three experiments running on a milling machine under different operating conditions [20]. The dataset is to study the tool wear in a regular cut. Data sampled by three different types of sensors (acoustic emission sensor, vibration sensor, current sensor) are acquired at several positions, in which the DC spindle motor current, AC spindle motor current, table vibration, spindle vibration, acoustic emission at table, and acoustic emission at spindle of each experiment are recorded. The initial data have 9000 samples. In this paper, we choose the datum every 60 data (1, 61, 121, ..., 8941), and only the data of the milling process is retained. Finally, the used dataset has 300 samples and includes three classes. The parameters of the clustering algorithms are set as follows: the fuzzification parameter m is 2, the threshold value is 10^{-6}, the maximum iteration is 50, and the membership criterion θ is 0.33. The DC spindle motor current is set as the output attribute, and the AC spindle

motor current, table vibration, spindle vibration, acoustic emission at table, and acoustic emission at spindle are set as the input attributes. The obtained clustering results are shown in Figure 2. It can be found that the GPR-FCM method produces the smallest misclassification rate. The proposed method is able to provide competitive clustering results compared with the other conventional clustering methods for the Mill dataset.

5.2. Borehole Dataset

The borehole dataset comes from the water flow rate problem [21]. The dataset has 400 samples, and can be evenly divided into two clusters according to the radius of borehole. In the experiment, the water flow rate is set as the output attribute, and the other the parameters are set as follows: the fuzzification parameter m is 2, the threshold value is 10^{-6}, the maximum iteration is 50, the membership criterion θ is 0.6. The experimental results are shown in Figure 3. It can be seen the misclassification rate of the GPR-FCM method is 0.125, which is smaller than the other clustering methods.

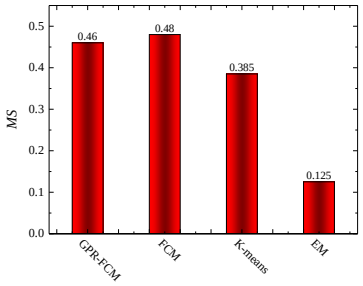

Figure. 3 Clustering results of the Borehole dataset

6. CONCLUSIONS

In this paper, a clustering method is proposed based on Gaussian process regression and fuzzy *c*-means algorithm. In the proposed method, the regression relationship instead of the distance among the objects is utilized to evaluate the difference between the clusters, and is described through Gaussian process regression. The clustering objective function of the proposed method and its optimization method are designed under the framework of fuzzy *c*-means algorithm. The effect of attribute number, sample number, level of noise and cluster number on the performance of the proposed method is investigated through a series of synthetic and engineering datasets. The results indicate that the GPR-FCM method can provide much better clustering performance than the conventional clustering methods based on spatial distribution, which indicates the applicability and potential of the proposed method in data clustering.

7. ACKNOWLEDGE

This work was supported by the Jiangsu University Senior Talent Fund Project (20JDG068).

REFERENCES:

[1] S. E. Schaeffer, "Graph clustering," *Computer science review*, vol. 1, no. 1, pp. 27-64, Aug. 2007.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264-323, Sep. 1999.

[3] C. Fraley, and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American statistical Association*, vol. 97, no. 458, pp. 611-631, Dec. 2002.

[4] M. Seeger, "Gaussian processes for machine learning," *International journal of neural systems*, vol. 14, no. 2, pp. 69-106, 2004.

[5] M. Kumar, M. Bhatt, and P. Samui, "Modeling of elastic modulus of jointed rock mass: Gaussian process regression approach," *International Journal of Geomechanics*, vol. 14, no. 3, pp. 06014001, Jun. 2014.

[6] L. Pasolli, F. Melgani, and E. Blanzieri, "Gaussian process regression for estimating chlorophyll concentration in subsurface waters from remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 3, pp. 464-468, Feb. 2010.

[7] J. Liu, W. Wang, F. Ma, Y. Yang, and C. Yang, "A data-model-fusion prognostic framework for dynamic system state forecasting," Engineering *Applications of Artificial Intelligence*, vol. 25, no. 4, pp. 814-823, Jun. 2012.

[8] J. Celaya, A. Saxena, S. Saha, and K. Goebel, "Prognostics of power MOSFETs under thermal stress accelerated aging using data-driven and model-based methodologies," *Annual Conference of the Prognostics and Health Management Society*, 2011.

[9] I. Ocak, and S. E. Seker, "Calculation of surface settlements caused by EPBM tunneling using artificial neural network, SVM, and Gaussian processes," *Environmental earth sciences*, vol. 70, no. 3, pp. 1263-1276, Jan. 2013.

[10] L. Hewing, J. Kabzan, and M. N. Zeilinger. "Cautious model predictive control using gaussian process regression." *IEEE Transactions on Control Systems Technology*, vol. 28, no. 6, pp. 2736-2743, 2019.

[11] A. Zeng, H. Hodde, and Y. Yu. "Prediction of building electricity usage using Gaussian Process Regression." *Journal of Building Engineering*, vol. 28, pp. 101054, 2020.

[12] S. Parveen, N. K. Tiwari, and Subodh Ranjan. "Modelling of infiltration of sandy soil using gaussian process regression." *Modeling Earth Systems and Environment*, vol. 3, pp. 1091-1100, 2017.

[13] Y. Yun, H. C. Kim, S. Y. Shin, J. Lee, A. D. Deshpande, and C. Kim. "Statistical method for prediction of gait kinematics with Gaussian process regression." *Journal of biomechanics*, vol. 47, no. 1, pp. 186-192, 2014.

[14] J. C. Bezdek, "Objective Function," *Clustering, Pattern recognition with fuzzy objective function algorithms*. Boston, MA, USA: Springer, 1981, pp. 43-93.

[15] A. M. Ali, G. C. Karmakar, and L. S. Dooley, "Review on fuzzy clustering algorithms," *Journal of Advanced Computations*, vol. 2, no. 3, pp. 169-181, 2008.

[16] A. P. Reynolds, G. Richards, B. de la Iglesia, and V. J. Rayward-Smith, "Clustering rules: a comparison of partitioning and hierarchical clustering algorithms," *Journal of Mathematical Modelling and Algorithms*, vol. 5, no. 4, pp. 475-504, Mar. 2006.

[17] R. Maitra, "On the Expectation-Maximization algorithm for Rice-Rayleigh mixtures with application to noise parameter estimation in magnitude MR datasets," *Sankhya B*, vol. 75, no. 2, pp. 293-318, Jan. 2013.

[18] X. Song, L. Lv, J. Li, W. Sun, and J. Zhang, "An advanced and robust ensemble surrogate model: extended adaptive hybrid functions," *Journal of Mechanical Design*, vol. 140, no. 4, Feb. 2018.

[19] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL, USA: Chapman and Hall/CRC, 2003.

[20] A. Agogino and K. Goebel. BEST lab, UC Berkeley. "Milling Data Set ", NASA Ames Prognostics Data Repository (http://ti.arc.nasa.gov/project/prognostic-data-repository), NASA Ames Research Center, Moffett Field, CA, 2007.

[21] R.B. Gramacy, H. Lian. "Gaussian process single-index models as emulators for computer experiments," *Technometrics*, vol. 54, no. 1, pp. 30-41, 2012.