# Fuzzy Clustering Analysis of Power Incomplete Data based on Improved IVAEGAN Model

Yutian Hong*, Jun Lin
Guangdong Electric Power Information Technology Co., Ltd.
Guangzhou, Guangdong 520000 China

***Abstract*—The scale of data generated by the complex and huge power system during operation is also very large. With the data acquisition of various information systems, it is easy to form the situation of incomplete power data information, which cannot guarantee the efficiency and quality of work, and reduce the security and reliability of the entire power grid. When incomplete data and incomplete data sets are caused by data storage failure or data acquisition errors, fuzzy clustering of data will face great difficulties. The fuzzy clustering of incomplete data of the power equipment is divided into the processing of incomplete data and the clustering analysis of "recovered" complete data. This paper proposes an IVAEGAN-IFCM interval fuzzy clustering algorithm, which uses interval data sets to fill in the incomplete data, and then completes the clustering of interval data. At the same time, the whole numerical data set is transformed into a complete interval data set. The final clustering result is obtained by interval fuzzy mean clustering analysis of the whole interval data set. Finally, the algorithm proposed in this paper and other machine learning training data sets is made for experimental analysis. The experimental results show that the algorithm proposed in this paper can complete incomplete data sets with high precision clustering. Compared with other contrast methods, it shows higher clustering accuracy. Compared with the numerical clustering algorithm, the clustering accuracy is improved by more than 4.3%, and it has better robustness. It also shows better generalization on the artificial data sets and other complex data sets. It is helpful to improve the technical level of the existing power grid and has important theoretical research value and engineering practice significance.**

***Keywords*—Power system; power equipment; incomplete data; fuzzy clustering; mining algorithm**

## I. Introduction

With the development of the economy, along with the construction of the smart grid and smart distribution network informatization, the level of automation and interactive are raised. The IoT degree of mutual penetration and integration of power distribution networks as a direct link between the user and the power grid is also improved. The safe and stable operation of the distribution network is directly related to the electricity quality and reliability of power use. The reliability of the distribution network system is particularly important [1]. Due to the massive increase of all kinds of power data and the increase of complexity, the data processing speed and processing capacity are greatly improved with the help of computer clustering analysis, and it is widely used in many key parts of power [2].

Distribution network equipment will produce a large amount of data in operation, because in the whole distribution system, there are many equipment from different manufacturers. Their models are mostly different because these distribution equipment and background monitoring system will use their own transmission protocol communication, resulting in the existence of many communication methods. Due to the conversion of various interfaces and transmission protocols, it is inevitable to cause problems such as slow data transmission rate and loss of some data information. If these missing data are not effectively analyzed, the maintenance efficiency will be low, and the equipment operation status cannot be timely, accurately and finely evaluated. Finally, it will affect the operation quality and efficiency of the large power grid.

With the in-depth study of incomplete data attribute imputation algorithm, it is further found that interval data samples can better express the ambiguity of incomplete data and improve the accuracy and robustness of clustering [3]. In the fields of pattern recognition, image processing, fuzzy control, and clustering analysis, there were some relevant researches [4]. Scholars in China and abroad have proposed many improved algorithms and applications for interval fuzzy sets. Zhang and other scholars proposed an improved BP neural network interval filling incomplete data clustering algorithm, which uses a neural network to fill the interval range of missing attributes [5]. Li et al. proposed an interval kernel function fuzzy clustering algorithm, which uses the nearest neighbor rule to determine the missing attribute interval, and uses the kernel function fuzzy C-means to map and cluster the samples in high dimensions [6]. The above research content has obvious advantages for dynamic data and big data, but it is prone to complex and difficult computing problems in model construction. Trang and other scholars proposed an interval fuzzy co-clustering algorithm, which applies interval data to co-clustering to make the clustering results more accurate [7]. Reference [8] uses mutually exclusive maintenance, simultaneous maintenance, power grid security, resources and other constraints to constrain interrelated equipment, which comprehensively considers the status of power grid equipment, and power grid operation, and uses a tabu search algorithm to optimize the starting period of power transmission and transformation equipment maintenance in the whole network. The calculation methods in references [7] and [8] are relatively simple, but the accuracy of estimation needs to be further improved. The above research content does not fully tap the potential value of incomplete data, and the effective information is not fully utilized. The data processing of distribution network equipment operation big data will mainly

*Corresponding Author.

study the incomplete data objects in the distribution network equipment operation big data. The vector data repair technology based on genetic nearest neighbor clustering will study the data repair technology-oriented to the distribution network operation big data to facilitate the filling and repair of incomplete distribution network operation data objects [9].

To solve the problem of data loss in large-scale data, a deep learning generation model based on the conditional generation antagonistic network is designed according to the data characteristics. An optimization function is proposed according to the data characteristics. The replacement of training data is optimized, and the optimization constraints are proposed to make the interpolation model more effective. Based on the IVAEGAN-FCM algorithm, this paper proposes an IVAEGAN-IFCM interval fuzzy clustering algorithm, which uses interval data sets to cluster incomplete data. This study is closely related to the operation status of power equipment, which is conducive to improving the safe operation efficiency of power grid enterprises.

The main innovations of this paper are as follows:

*1)* A certain estimation error exists in the estimation of the missing attribute through the IVAEGAN model. The average value of all errors is taken as the interval size range of the interval type data to construct an interval type data set.

*2)* Calculate the extreme value of the attribute of the nearest neighbor sample of the incomplete data to restrict the interval size of each data.

*3)* Calculate the local density of the sample in the adjacent area of each sample, and further constrain the size of the interval through the local density of the sample.

*4)* Convert the whole numerical data set into a complete interval data set. To improve the accuracy of incomplete data clustering, the interval fuzzy C-means (IFCM) clustering analysis is carried out on the whole interval data set.

The main contents of this paper are as follows:

*1)* This paper introduces the importance of equipment data integrity in the power system operation.

*2)* In the related work, the power system equipment data and the IVAEGAN model are introduced.

*3)* The construction of the IVAEGAN interval model and the calculation of data integrity by the fuzzy clustering algorithm are done.

*4)* The IFCM algorithm is improved by using the IVAEGAN model.

*5)* The accuracy and effectiveness of the proposed algorithm are verified by experimental simulation analysis.

*6)* Finally, the research contents and results of this paper are summarized.

## II. Related work

### A. Analysis of Data Sources for Typical Power Equipment

Power grid automation helps power grid dispatchers to grasp the operating conditions of the power grid under their jurisdiction in real-time so that dispatchers can make correct dispatching decisions. At the same time, it can also provide data support such as load forecast for short-term and medium-term power grid production and development plans [10]. The centralized control mode of the substation is established to realize more and more automation systems with different functions. To improve the reliability, efficiency, and power supply quality of distribution network operation, the distribution automation system has been developed. At present, the dispatching automation system, substation integrated automation system, and distribution automation system have become the main components of the power grid automation system [11]. The power data studied in this paper takes the distribution equipment data as an example, and carries out data resources according to the equipment material information, as shown in Table I.

### B. IVAEGAN Value Estimation Principle

*1) An IVAEGAN* model performs estimated value filling on missing attributes $x_{ik}$ in an incomplete data set. Since the IVAEGAN model also performs estimated value calculation on data with complete attributes in a training process, an absolute error exists between an expected estimated value and an actual value, and the absolute value of an error average value is used as the interval size of the estimated value of the missing attributes. The left and right endpoints of the valuation interval are: $x_{ik}^{-}$, $x_{ik}^{+}$ respectively. Then the incomplete data can be expressed in the form of interval, namely $\left[x_{ik}^{-}, x_{ik}^{+}\right]$;

For the complete data $x$ in the incomplete data set, its complete attribute value $x_{ij}$ also needs to be converted into the form of interval data, that is $\left[x_{ik}^{-}, x_{ik}^{+}\right]$, where $x_{ij}^{-}=x_{ij}=x_{ij}^{+}$. Therefore, the complete attribute of the complete data is also expressed in the interval type, and the left and right interval endpoints of the interval attribute are equal to the attribute value [12].

TABLE I. Power Cable Index System Table

| part | First class indicator | Secondary indicator (state) |
|---|---|---|
| power cable | Cable body | Line load, Insulation resistance, Exterior, Fire fire flame retardant, Depth, filthy |
| | Cable terminal | |
| | Cable middle head | |
| | Cable channel | |
| | Auxiliary facilities | |

Through the training process of the IVAEGAN and the evaluation process of incomplete data, the average evaluation error of the training process is determined as the range of interval size. The evaluation of incomplete data attributes is determined as the median of the interval to determine the evaluation interval of the incomplete data [13]. And then the complete numerical data set after the valuation "recovery" is transformed into a complete interval data set. Finally, the complete interval data set is subjected to fuzzy clustering analysis through the IFCM clustering algorithm, and the clustering result is obtained [14].

## III. IVAEGAN Fuzzy Clustering Algorithm for Interval Estimation

### A. IVAEGAN Model Construction

VAE (Variational autoencoder) is fused as the generator of GAN (Generative Adversarial Network), and the fused model is improved to obtain the IVAEGAN model. The IVAEGAN network is proposed to better estimate, predict and fill the missing data attribute values contained in the incomplete data set [15]. The core idea of the IVAEGAN network proposed in this paper is to feedback the difference between the predicted value of the incomplete data through the network generator and the discriminator to the input layer so that the network can obtain more information, and thus can better realize the estimation filling of missing attributes in incomplete data. This can make the filling value of missing attributes more reasonable, thus improving the effectiveness of clustering analysis. The topology of the improved model is shown in Fig. 1:

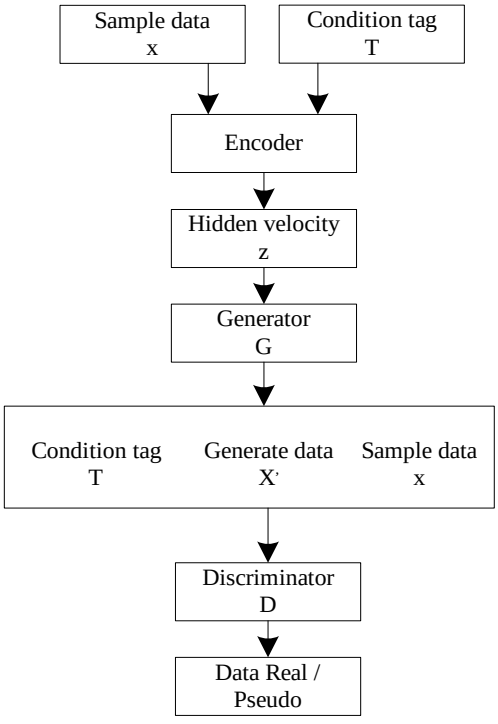


Fig. 1. IVAEGAN Model Structure.

The integration of variational learning and adversarial learning provides strong support for learning and reasoning in generative models, and these methods are used to establish new hybrid reasoning methods.

### B. Interval reconstruction of numerical incomplete data set

Interval data $x=[x^-,x^+]$ and interval data $y=[y^-,y^+]$: the addition operation of interval type is expressed as: $x+y=\left[x^-+y^-,x^++y^+\right]$. Similarly, the definition of subtraction operation of interval data is expressed as: $x-y=\left[x^--y^-,x^+-y^+\right]$ [16]. Euclidean distance is commonly used in calculating relative distance, and the Euclidean distance formula for interval data is:

$$D(x,y)=\sqrt{(x^--y^-)^2+(x^+-y^+)^2} \tag{1}$$

The density of samples in a local area reflects the degree of clustering of samples, and also reflects the similarity between samples. The greater the density of samples in a local area, the closer the attribute values are. In this paper, the regional density of sample points is used to calculate sample density a, and the obtained attribute estimation interval $[\min,\max]$ is constrained to obtain a new interval: $\left[\min+\frac{a(\max-\min)}{2},\max-\frac{a(\max-\min)}{2}\right]$.

The calculation formula for the distance between samples $x_p$ and other samples is shown in (2):

$$d_{pq}=\left\|x_p-x_q\right\|_2^2 \tag{2}$$

Let the attribute dimension be the s-interval dataset $\overline{X}=\{\overline{x}_1,\overline{x}_2,\cdots,\overline{x}_n\}$. The data $\overline{x}_i=\left[\overline{x}_{1i},\overline{x}_{2i},\cdots,\overline{x}_{si}\right]^T$, for any $\overline{x}_{ji}=\left[x_{ji}^-,x_{ji}^+\right]$, the objective function formula of the interval fuzzy C-means algorithm is shown in (3):

$$J_m(U,\overline{V})=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^m\left\|\overline{x}_k-\overline{v}_i\right\|_2^2 \tag{5}$$

$$\sum_{i=1}^{c}u_{ik}=1,k=1,2,\cdots,n \tag{4}$$

$$\left\|x_k-v_i\right\|_2^2=\sqrt{(x_k^{\ -}-v_i^-)^T(x_k^{\ -}-v_i^-)+(x_k^{\ +}-v_i^+)^T(x_k^{\ +}-v_i^+)} \tag{5}$$

$\overline{v}_i$ represents the ith cluster center. $\overline{V}$ is the cluster center matrix, and is expressed as: $\overline{V}=\left[\overline{v}_{ji}\right]=\left[\overline{v}_1,\overline{v}_2,\cdots,\overline{v}_c\right]$, $\overline{v}_{ji}=\left[v_{ji}^-,v_{ji}^+\right]$, $i=1,2,\cdots,c$, $j=1,2,\cdots,s$。

The minimum condition of Equation (5) is:

$$v_i^-=\frac{\sum_{k=1}^{n}u_{ik}^m x_k^-}{\sum_{k=1}^{n}u_{ik}^m},i=1,2,\cdots,c \tag{6}$$

$$v_i^+ = \frac{\sum_{k=1}^{n} u_{ik}^m x_k^+}{\sum_{j=1}^{n} u_{ik}^m}, i = 1, 2, \cdots, c \tag{7}$$

If there is an interval type data sample $\overline{x}_k$ within the interval value of a cluster center, its membership degree is set to 1; otherwise, its membership degree is 0 and it does not belong to this category. The formula is as follows:

$$u_{ij} = \begin{cases} 0, i \neq h \\ 1, i = h \end{cases} \tag{8}$$

Otherwise：

$$u_{ij} = \left[ \sum_{t=1}^{c} \left( \frac{\left\| x_j - v_i \right\|_2^2}{\left\| x_j - v_t \right\|_2^2} \right)^{\frac{1}{m-1}} \right]^{-1}, i = 1, 2, \cdots, c; j = i = 1, 2, \cdots, n \tag{9}$$

The basic steps of IFCM algorithm are shown in Figure 2:

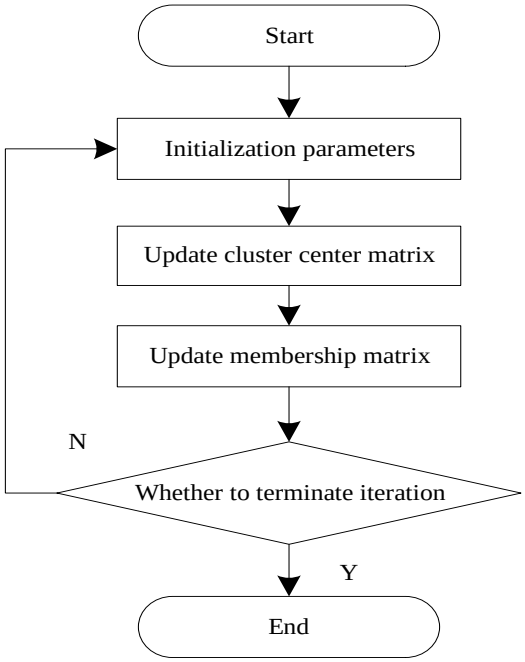


Fig. 2. IFCM Algorithm Process.

*1)* Initialize algorithm parameters, namely set an iteration stopping threshold $\varepsilon$ , a fuzzy clustering parameter m. The number of clustering is $c(2 \le c \le \sqrt{n})$ and the maximum iteration number is G. Initialize the membership degree matrix $U^{(0)}$.

*2)* Update a clustering center matrix: when that iteration time reach the $l(l = 1, 2, \cdots)$ time, update the clustering center matrix $U^{(l-1)}$ by using the clustering prototype calculation formulas (6) and (7) in combination with the membership degree partition matrix $\overline{V}^{(l)}$.

*3)* Update the membership degree matrix: update the partition membership degree matrix $U^{(l)}$ by using formulas (8) and (9).

*4)* Iteration termination condition of the algorithm: when the number of iterations reaches the maximum, or $\forall i$ , k, $\max \left| u_{ik}^{(l)} - u_{ik}^{(l-1)} \right| < \varepsilon$ , the iteration is stopped, then the IFCM algorithm stops and outputs the partition matrix U and the clustering prototype matrix $\overline{V}$ ; otherwise $l = l + 1$ , return to step (2) [18].

### C. The Autoencoder Algorithm Normalization Flow

The basic flow of the auto-encoder algorithm is as follows:

*1) Input normalization.* All the attributes of the input data is transformed into the number between [0, 1] to improve the flexibility of the model and eliminate the difference between the orders of magnitude of each attribute.

*2) Initialize the auto-encoder model parameters.* The parameters include the node number of input layer, the layer number of hidden layer, the node number of each layer of hidden layer, the encoder weight matrix, decoder weight matrix, center vector, maximum training times, and the model self-learning rate and parameters.

*a)* Input data. Sample points are drawn from the data set.

*b)* Random sampling in noisy data.

*c)* The SGVB gradient estimation.

*d)* Update parameters and weight.

*e)* Judge whether that parameter are converged.

*f)* Determination of termination. When the model is completely converged or reaches the maximum training times, the training is ended; otherwise, the step (3) is repeated.

## IV. IVAEGAN-IFCM ESTABLISHMENT OF ALGORITHM

To solve the problem of incomplete data fuzzy clustering, an interval fuzzy clustering algorithm based on the IVAEGAN estimation (IVAEGAN-IFCM) is proposed to cluster the incomplete data sets. Through the IVAEGAN model, the incomplete data set is "restored" into a complete numerical data set, and the complete numerical data set is converted into an interval data set according to the interval rule proposed in this paper, and then the interval data set is analyzed by fuzzy clustering [19].

The specific algorithm flow of the IVAEGAN-IFCM algorithm is shown in Fig. 3.

*1)* Construct a nearest neighbor sample set for an incomplete data sample. The nearest neighbor samples are selected according to the nearest neighbor rule, and the nearest neighbor sample set of incomplete data is constructed.

*2)* Construct an attribute nearest neighbor sample interval of incomplete data. According to the interval rule proposed in this paper, the maximum and minimum values of missing attributes are determined in the nearest neighbor sample set of incomplete data to construct the nearest neighbor interval of attributes.

*3)* Determine the density of data samples in the neighborhood of incomplete data samples [20].

*4)* Input sample normalization. All data are converted into numbers between intervals$^{[0,1]}$, thus eliminating the difference of orders of magnitude between dimensions.

*5)* Initialize the IVAEGAN model. Initialize the network parameters in the IVAEGAN model, weight, bias value, maximum number of iterations, and training error.

*6)* Train the IVAEGAN model. The IVAEGAN model was trained on the complete data.

*7)* Fill in the missing attributes. The IVAEGAN model proposed in this paper estimates and predicts each missing data attribute in incomplete data, and at the same time obtains the estimation error of the IVAEGAN network for the complete attributes in the data set.

*8)* Interval data set: according to the interval type transformation rule proposed in this paper, all the data in the numerical data set are transformed into interval type, and then the interval type matrix is constructed.

*9)* Initialize the IFCM algorithm parameters. Initialize the membership degree matrix, and set the number of clustering categories, the number of cycles, the termination threshold and the fuzzy index.

*10)* Update the cluster center matrix. Update the clustering center matrix $V^{(l)}$ according to the clustering center matrix $U^{(l-1)}$.

*11)* Update that membership matrix. The statement $V^{(l)}$ updates the membership degree matrix $U^{(l)}$.

*12)* Algorithm condition judgment: when the number of iterations reaches the maximum, or $\max\left|U^{(l+1)}-U^{(l)}\right|\le\varepsilon$, the algorithm stops iterating; otherwise $l=l+1$, it returns to (10).

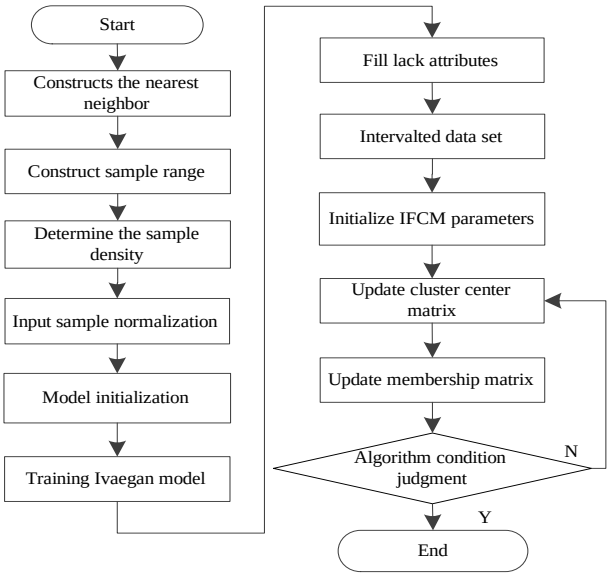


Fig. 3. IVAEGAN-IFCM Algorithm Process.

## V. Experimental Analysis

### A. Experimental Preparation

Aiming at the problem that the FCM clustering algorithm cannot directly use incomplete data for clustering analysis, this paper proposes a numerical incomplete data fuzzy clustering algorithm based on the IVAEGAN estimation. The missing attributes in the incomplete data are estimated and filled by the IVAEGAN model, and then the complete data set after recovery is analyzed by fuzzy clustering using the FCM clustering algorithm.

### B. Construct the Artificial Dataset

In this experiment, the effectiveness of the algorithm is verified by artificial data sets, which are generated by the methods in the literature. Two artificial data sets are obtained. The artificial data set Test 1 contains two types of data. Each type has 1000 data samples, and the total number of samples is 2000 [21]. The artificial data set Test 2 contains three types of data, each type of data contains 800, 1000 and 2200 samples respectively, and the total number of data sets is 4000. Two artificial data sets subject to independent two-dimensional normal distribution are generated according to the above literature, and the conditional expectation and variance are as follows:

$$u=\begin{bmatrix}u_1\\u_2\end{bmatrix}\quad \Sigma=\begin{bmatrix}\sigma_1^2 & 0\\0 & \sigma_2^2\end{bmatrix}$$

Where, the data generation parameters of manual data sets Test 1 and Test 2 are shown in Table II:

TABLE II. Artificial Datasets

| | $u_1$ | $u_2$ | $\sigma_1^2$ | $\sigma_2^2$ |
|---|---|---|---|---|
| Test 1 | 3 | 3 | 2 | 2 |
| | 7 | 9 | 2 | 2 |
| Test 2 | 28 | 24 | 6 | 6 |
| | 38 | 30 | 9 | 11 |
| | 50 | 42 | 17 | 20 |

According to the above generation method and parameter settings, generate a series of artificial data sets, in which the distribution of Test 1 is shown in Fig. 4, and the distribution of Test 2 is shown in Fig. 5.

In Fig. 4, the data in Test 1 is generated based on the first set of parameters. The two types of data are generated in equal quantities, and the data are evenly distributed, but the obvious differences of the data categories are maintained.

In Fig. 5, due to the different distribution and data amount of the three types of data, the first type of data is more concentrated, while the second and third types of data are more dispersed. The three types of data can still maintain relative independence, which is suitable for the validity verification of clustering algorithm.

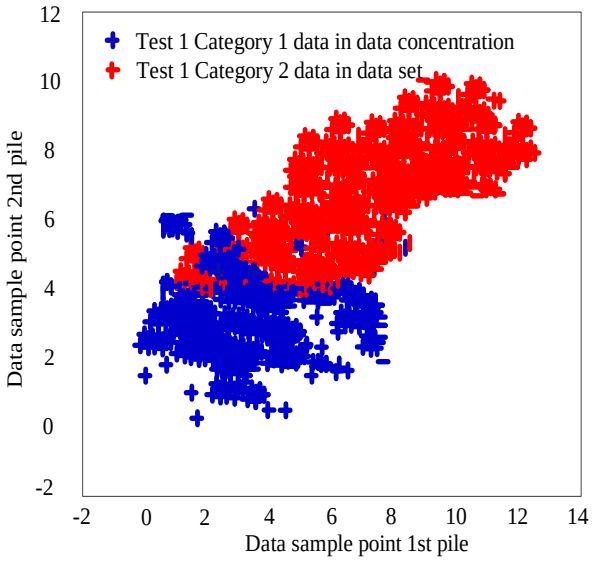


Fig. 4. Artificial Dataset Test 1.

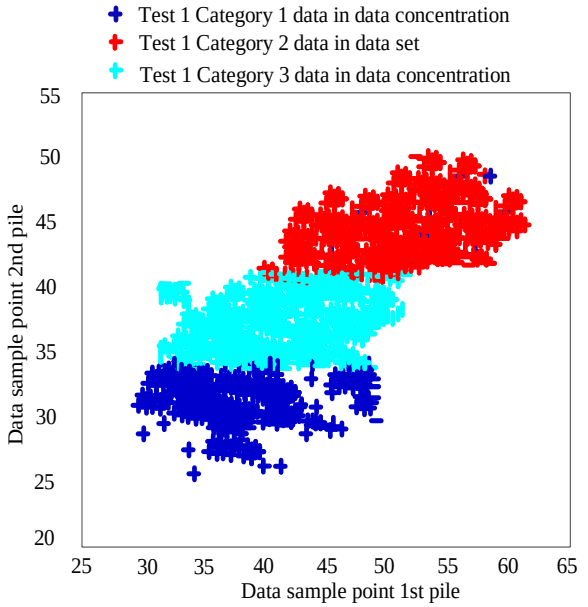


Fig. 5. Artificial Dataset Test 2.

### C. Incomplete Data Generation Rules

In the whole data processing process, in order to make the incomplete data generated in the experiment closer to reality, the incomplete data set is generated by randomly discarding the data and randomly losing the complete data in different proportions set manually. In an incomplete data set, the missing attribute of the data sample is denoted by "?" to represent. The rules for generating data missing attributes of the incomplete data sets are as follows:

*1) In an incomplete data set,* the attribute values of a sample data cannot be completely lost, that is, if the data set is n-dimensional, the incomplete data in it can lose at most attributes, and at least one attribute of an incomplete data must exist.

*2) In the incomplete data set,* any one-dimensional attribute has at least one complete attribute value, that is, the attribute column of the data set cannot be empty to ensure the reliability of the valuation.

In the algorithm comparison experiment of this paper, the complete data strategy fuzzy C-means clustering algorithm (whole data strategy fuzzy C-means clustering) WDS-FCM) and the local distance strategy fuzzy C-means clustering algorithm (partial distance strategy fuzzy C-means clustering, PDS-FCM) are rejection methods [22].

The remaining two comparison algorithms optimize the complete strategy fuzzy c-means clustering algorithm (optimal completion strategy fuzzy C-means clustering, OCS-FCM) and the nearest prototype strategy fuzzy C-means clustering algorithm (Nearest Prototype Strategy fuzzy C-means clustering, NPS-FCM) belong to the valuation method [23].

The IVAEGAN-IFCM algorithm proposed in this paper uses the IVAEGAN model to estimate the incomplete missing attributes. The IVAEGAN model performs feature extraction and data generation on the incomplete data set. Meanwhile, the model combines the median of the neighbor data as the conditional label to make the estimation range more accurate and efficient. After the IVAEGAN model performs estimation filling on the incomplete data set, the obtained complete data set completes the fuzzy clustering, which improves the clustering accuracy [24]. Iris, Bupa and Breast data sets are used in the experiment. Each data set contains 500 samples, 11 feature categories and 9 attributes.

According to the average number of iterations in Tables III to V, the algorithm in this paper can converge stably after multiple iterations under different missing rates of each data set. Compared with other clustering algorithms, the algorithm does not achieve the best results, but it can still achieve stable results after a certain number of iterations.

TABLE III. AVERAGE ITERATION TIMES OF IRIS

| Missing rate /% | WDS-FCM | PDS-FCM | OCS-FCM | NPS-FCM | Algorithm |
|---|---|---|---|---|---|
| 5 | 25.1 | 25.3 | 27.5 | 27.8 | 30.5 |
| 10 | 25.9 | 26.3 | 27.9 | 28.2 | 30.8 |
| 15 | 26.1 | 26.7 | 28.6 | 29.3 | 31.4 |
| 20 | 26.8 | 27.1 | 28.8 | 29.7 | 31.8 |

TABLE IV. AVERAGE NUMBER OF ITERATIONS OF BUPA

| Missing rate /% | WDS-FCM | PDS-FCM | OCS-FCM | NPS-FCM | Algorithm |
|---|---|---|---|---|---|
| 5 | 33.7 | 36.2 | 36.6 | 36.7 | 36.5 |
| 10 | 35.6 | 37.4 | 39.1 | 37.7 | 39.3 |
| 15 | 37.3 | 35.9 | 38.6 | 38.9 | 42.4 |
| 20 | 38.6 | 36.5 | 36.8 | 40.3 | 46.2 |

TABLE V. AVERAGE ITERATION TIMES OF BREAST

| Missing rate /% | WDS-FCM | PDS-FCM | OCS-FCM | NPS-FCM | Algorithm |
|---|---|---|---|---|---|
| 5 | 15.3 | 15.2 | 16.4 | 15.6 | 18.4 |
| 10 | 15.9 | 16.3 | 16.8 | 15.7 | 18.9 |
| 15 | 16.1 | 16.8 | 17.2 | 16.3 | 22.3 |
| 20 | 16.9 | 17.4 | 18.6 | 17.2 | 24.6 |

According to the experimental results, the algorithm in this paper is relatively better, in the case of small data loss, such as 15% loss, the clustering effect of the numerical estimation method is better, which is obviously better than other algorithms.

From the standard deviation of clustering error scores in Tables VI to VIII, the algorithm in this paper can maintain a low standard deviation of clustering error scores under different missing rates of different data sets, which reflects the stability of the algorithm. In some cases, such as the case of less missing rate, the optimal solution cannot be obtained, and in other cases, the best results can be obtained.

TABLE VI. STANDARD DEVIATION OF IRIS CLUSTERING ERROR SCORE

| Missing rate /% | WDS-FCM | PDS-FCM | OCS-FCM | NPS-FCM | Algorithm |
|---|---|---|---|---|---|
| 5 | 1.24 | 1.56 | 2.02 | 1.86 | 2.03 |
| 10 | 2.15 | 1.75 | 1.95 | 1.78 | 1.78 |
| 15 | 2.08 | 1.89 | 2.04 | 169 | 1.93 |
| 20 | 2.09 | 1.92 | 2.01 | 2.07 | 2.04 |

TABLE VII. STANDARD DEVIATION OF BUPA'S CLUSTERING ERROR SCORE

| Missing rate /% | WDS-FCM | PDS-FCM | OCS-FCM | NPS-FCM | Algorithm |
|---|---|---|---|---|---|
| 5 | 2.03 | 2.36 | 2.12 | 2.36 | 2.04 |
| 10 | 1.83 | 1.89 | 2.05 | 2.04 | 2.12 |
| 15 | 2.51 | 1.38 | 1.89 | 2.01 | 2.23 |
| 20 | 2.18 | 1.94 | 1.95 | 1.98 | 2.64 |

TABLE VIII. STANDARD DEVIATION OF BREAST CLUSTERING ERROR SCORE

| Missing rate /% | WDS-FCM | PDS-FCM | OCS-FCM | NPS-FCM | Algorithm |
|---|---|---|---|---|---|
| 5 | 3.54 | 3.12 | 3.45 | 2.89 | 4.31 |
| 10 | 3.57 | 2.92 | 3.23 | 2.96 | 4.56 |
| 15 | 3.26 | 3.02 | 3.21 | 3.25 | 4.89 |
| 20 | 3.58 | 3.08 | 3.04 | 3.12 | 4.97 |

Test results show that the algorithm in this paper is better than other algorithms on the whole. Compared with other algorithms, the performance of our algorithm is improved by 4%. 3% when the missing rate of the data set is low, such as the missing rate of 5% and 10% of the small data loss. With the increase of the missing rate, the test performance of the sample is also gradually improved. It is proved that the algorithm proposed in this paper can fully show the uncertainty of incomplete data estimation by using interval data, and has superior performance in the process of missing data filling and clustering analysis.

The WDS-FCM algorithm eliminates all the incomplete sample data in the data set and only performs fuzzy clustering on the complete sample data. The clustering accuracy of the algorithm will be greatly affected when the proportion of incomplete samples increases. The PDS-FCM algorithm replaces the Euclidean distance in FCM with the local distance formula, and only adds the iterative operation to the complete attributes in the processing of incomplete sample data, without considering the missing attribute information of incomplete samples, which does not give full play to the information value of incomplete samples. Therefore, both the WDS-FCM and the PDS-FCM do not make full use of the effective information value of the mining incomplete data. The IVAEGAN-FCM algorithm proposed in this paper does not delete and abandon incomplete data, but reconstructs the model to fill in the missing data, so that the data set can be "restored" to a complete data set. The training samples are used to train the IVAEGAN model, and the missing attributes of each incomplete attribute are estimated and filled to obtain the "recovered" complete data set, and then the complete data set is subjected to fuzzy clustering.

## VI. CONCLUSION

In this paper, according to the incomplete data formed in the power system, a numerical incomplete data fuzzy clustering algorithm based on the improved IVAEGAN estimation is proposed to solve the problem that the traditional clustering algorithm cannot directly use the incomplete data. A new fusion model is constructed by combining VAE and GAN to estimate and fill the incomplete data. The main work includes:

*1)* Fill the incomplete data attributes to realize the estimation clustering. A fuzzy clustering algorithm for numerical incomplete data based on the improved IVAEGAN estimation is proposed.

*2)* Construct a nearest neighbor sample set for that incomplete data according to the nearest neighbor rule, generating a model VAE and a model GAN, and constructing an IVAEGAN model.

*3)* Construct a missing attribute label by using that median value of the attribute of the nearest neighbor sample set so that the IVAEGAN model can obtain more effective information, and the estimation accuracy is improved.

*4)* UCI data sets and two artificial data sets are used for comparative experiments to verify the effectiveness of the algorithm, and the effectiveness of the algorithm is summarized in depth.

The model in this paper uses the gradient integral to update the parameter values, which is fast and has high computational complexity, and has obvious effect on improving the training speed. Compared with other models, this paper uses the attributes of the nearest neighbor sample set to construct condition variables, and other construction methods of condition variables are more effective, which further optimizes

the space and improves the theoretical and practical basis of data integrity research. The research results of this paper are very important to fully mine the effective information in the incomplete data of power system, and play an important role in ensuring the normal operation of power system.

In the future work, more optimization methods will be used to optimize the parameter update, such as the wolf pack optimization algorithm. The GAN model is prone to the phenomenon of gradient explosion, center collapse and training non-convergence to further enhance the space for improvement.

## REFERENCES

[1] Zhong G Q, Gao W, Liu Y B, Yang Y Z. Generative adversarial networks with decoder–encoder output noises[J]. Neural Networks,2020,127:19-28.

[2] Yoon J ,Jordon J, Mihaela S. GAIN: Missing Data Imputation using Generative Adversarial Nets[J]. arXiv preprint arXiv:1808.02920, 2018.

[3] Martin A, Soumith C, Bottou L. Wasserstein GAN[J]. arXiv preprint arXiv:1701.07875v3, 2018.

[4] Mihaela R, Balaji L, David W, Shakir M. Variational Approaches for Auto-Encoding Generative Adversarial Networks[J]. arXiv preprint arXiv:1706.04987v2,2017.

[5] Du C D, Li J P, Huang L J, He H G. Brain Encoding and Decoding in fMRI with Bidirectional Deep Generative Models[J]. Engineering,2019,5(5):948-953.

[6] Chen S M, Yu J B, Wang S J. One-dimensional convolutional auto-encoder-based feature learning for fault diagnosis of multivariate processes[J]. Journal of Process Control, 2020, 87:54-67.

[7] Ahmad M K, Mehmet S G, Mehmet R T, Hilal K. A new framework using deep auto-encoder and energy spectral density for medical waveform data classification and processing [J]. Biocybernetics and Biomedical Engineering,2019,39(1):148-159.

[8] Tang L J, Zheng S C, Zhou Z G. Estimation and inference of combining quantile and least-square regressions with missing data [J]. Journal of the Korean Statistical Society, 2018, 47(1):77-89.

[9] Liu X J, Zhang H, Niu Y G, Zeng D L. Modeling of an ultra-supercritical boiler-turbinesystem with stacked denoising auto-encoder and long short-term memory network [J].Information Sciences, 2020, 525(1):143-152.

[10] Chen G, Wang H, Jian T, Xu C, Sun S. Method for denoising and reconstructing radar HRRP using modified sparse auto-encoder[J]. Chinese Journal of Aeronautics, 2020,33(3):1026-1036.

[11] Gao Z S, Shen C, Xie C Z. Stacked convolutional auto-encoders for single space target image blind deconvolution[J]. Neurocomputing,2018,313(3):295-305.

[12] Liang Y, Ke S, Zhang J, et al. Geoman: Multi-level attention networks for geo-sensory time series prediction[A]. International Joint Conference on Artificial Intelligence[C]. 2018: 3428-3434.

[13] Han M, Zhong K, Qiu T, et al. Interval type-2 fuzzy neural networks for chaotic time series prediction: a concise overview[J]. IEEE Transactions on Cybernetics, 2018, 49(7):2720-2731.

[14] Wang H, Yang Z, Yu Q, et al. Online reliability time series prediction via convolutional neural network and long short term memory for service-oriented systems[J]. Knowledge-Based Systems, 2018, 159:132-147.

[15] Araújo R A, Nedjah N, Oliveira A L I, et al. A deep increasing–decreasing-linear neural network for financial time series prediction[J]. Neurocomputing, 2019, 347:59-81.

[16] Wang Z, Yan W, Oates T. Time series classification from scratch with deep neural networks: A strong baseline[A]. International Joint Conference on Neural Networks[C]. 2017:1578-1585.

[17] Ma Q, Zhuang W, Shen L, et al. Time series classification with Echo Memory Networks[J]. Neural networks, 2019, 117:225-239.

[18] Fawaz H I, Forestier G, Weber J, et al. Deep learning for time series classification: a review[J]. Data Mining and Knowledge Discovery, 2019, 33(4):917-963.

[19] Ma Q, Zheng J, Li S, et al. Learning Representations for Time Series Clustering[A]. Neural Information Processing Systems[C]. 2019:3776-3786.

[20] Shen L, Ma Q, Li S. End-to-end time series imputation via residual short paths[A]. Asian Conference on Machine Learning[C]. 2018:248-263.

[21] Cao W, Wang D, Li J, et al. Brits: Bidirectional recurrent imputation for time series[A]. Neural Information Processing Systems[C]. 2018:6775-6785.

[22] Luo Y, Cai X, Zhang Y, et al. Multivariate time series imputation with generative adversarial networks[A]. Neural Information Processing Systems[C]. 2018: 1596-1607.

[23] Luo Y, Zhang Y, Cai X, et al. E 2 GAN: end-to-end generative adversarial network for multivariate time series imputation[A]. International Joint Conference on Artificial Intelligence[C]. 2019:3094-3100.

[24] Zhang J, Yin P. Multivariate Time Series Missing Data Imputation Using Recurrent Denoising Autoencoder[A]. International Conference on Bioinformatics and Biomedicine[C]. 2019:760-764.