IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 29, 2020, accepted April 6, 2020, date of publication April 9, 2020, date of current version April 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986813

High-Dimensional Clustering for Incomplete Mixed Dataset Using Artificial Intelligence

MEISHAN LI[1], XIAOFENG LI[2], (Member, IEEE), AND JING LI[1]
[1]College of Information and Electronic Technology, Jiamusi University, Jiamusi 154007, China
[2]Department of Information Engineering, Heilongjiang International University, Harbin 150025, China

Corresponding author: Xiaofeng Li (mberse@126.com)

This work was supported in part by the Heilongjiang Provincial Undergraduate University Basic Research Business Expenses Research Project under Grant 2018-KYYWF-0938, in part by the Surface Scientific and Research Projects of Jiamusi University under Grant 13Z1201576, in part by the Basic Research Projects of Jiamusi University under Grant JMSUJCMS2016-009, and in part by the Ministry of Education Science and Technology Development Center Industry-University Research Innovation Fund under Grant 2018A01002.

ABSTRACT In order to address the problem that high energy consumption, high memory usage and low clustering effect in traditional data set high-dimensional clustering algorithms, we propose the high-dimensional clustering algorithm of incomplete mixed data set based on artificial intelligence. First, we construct the phase space reconstruction to ensure the invariance of features of incomplete mixed data set by analyzing the incomplete mixed data set and introduce the correlation dimension to obtain the feature correlation value. Second, we introduce the standard deviation and realize the extraction of features of incomplete mixed data set through calculating the sparsity of sample features. Third, we conduct repeated clustering for the mixed data set in the subspace according to the degree of correlation between incomplete mixed data sets in the multidimensional subspace. Last, we realize the design of high-dimensional clustering method for incomplete mixed data set in accordance with the stronger relevance in the mixed data sets. Experimental results show that the proposed algorithm has good correlation dimension processing effects, lower memory usage, time-consuming, lower and concentrated ensemble energy consumption (within 300J), good clustering effects, as high as 92%, which has some advantages and practical application value.

INDEX TERMS Artificial intelligence, subspace, phase space reconstruction, correlation dimension, mixed data set, high-dimensional clustering.

I. INTRODUCTION

Artificial intelligence covers a wide range of fields, including computer vision, machine learning, with the main objective to complete the tasks with machine [1], [2] that can only be completed with human intelligence, which has higher requirements for complexity and time [3]. The field of artificial intelligence usually involves data clustering and data classification. Classification is divides incomplete data with mixed values into existing categories according to the characteristics or attributes of data. Clustering is to divide high-similar things into meaningful and useful data group clusters according to the similarity of things [4], [5].

As one of the computing methods of Artificial intelligence, clustering analysis collects [6]–[8]. For the given data set, the data category is divided according to the similarity between the data elements, so that the similarity of the elements in the group reaches the maximum as the number of iterations increases. The similarity of elements between groups decreases to a minimum with the number of iterations, and finally they are classified by clusters. In many practical applications, we need to analyze a collection of data sets like images, text documents, etc. To model such data structures [9], [10]. At present, data cluster analysis has been successfully applied to multiple information fields such as social networks and image processing. At the same time as the rapid development of artificial intelligence technology, the capacity of computer processing and storage has also exploded, and information system data has been continuously updated with changes in time [11].

With the rapid growth of Artificial intelligence technologies, the capacity of computer processing and storage is also

The associate editor coordinating the review of this manuscript and approving it for publication was Ying Li.

on the explosive rise [12]–[14]. The data are on the rise in the process of collection, in which the incomplete mixed data set accounts for a larger proportion [15]. The efficient management for big data becomes a huge challenge for the computer information management at present.

In most practical application systems, the system is often required to be in a multi-sampling state. If the system's transmission signals include multiple forms such as fax and video, the frequency domain components of these signals are far apart. At this time, the system needs to complete the signal's automatic conversion according to artificial intelligence. The calculation of signal information needs to be completed in a high-dimensional space, which can further increase the calculation accuracy of data and information. Compared with low-dimensional space data, the data distribution in high-dimensional space is usually more sparse, and high-dimensional data sets are prone to a large number of irrelevant attributes, resulting in increased difficulty in high-dimensional clustering of the data set. Therefore, it is of great importance in researching the incomplete mixed data set, which has become the focus of many experts and scholars and received widespread concern in numerous fields [16]–[18].

At present, certain research results have been obtained for data clustering methods in different fields. Literature [19] has analyzed a method named high-dimensional data clustering (HDDC), which uses a series of Gaussian mixture models for clustering. We estimate the intrinsic dimension of each subgroup of data observed and conducted clustering analysis in the low-dimensional subspace. Such kind of model has received widespread concern due to its better classification functions. Literature [20] has insufficient data sets and noise data, therefore the clustering results can be easily disturbed. It is hard to solve the commensurability of multi-type target clustering quality in the space domain and it is complicated to use the present algorithm to solve the high-dimensional heterogeneous data of group intelligent sets. In order to overcome the above problems, Literature [21] first presents group similarity to expand multidimensional similar space region and uses the optimal value of iteration clustering function as the measure standard of clustering quality. In addition, it puts forth a fuzzy high-order mixed clustering algorithm, which not only can improve the anti-interference ability, effectively control the convergence rate of algorithm and reduce the computing time. In order to cope with different potential mechanisms, risk prognosis and therapeutic reaction, and use the information corresponding to data distribution and facilitate visualization, Literature [22] developed a parallel framework with high memory efficiency to decompose the original problem into parallel multiple regression subproblems. Sub-space clustering analysis is completed by combining sampling and parameter block partitioning, and a random optimization algorithm that minimizes the objective function. Literature [23] introduces a feature selection visual analysis concept based on star coordinates of linear discriminant analysis. This concept generates the optimal clustering and separation view in a linear sense for labeled and unlabeled data. In this way, users can explore each dimension's contribution to the cluster.

However, many problems are prone to occur during the operation of traditional clustering algorithms, such as high cluster energy consumption, high memory consumption, and low clustering effects. therefore, the paper proposed high-dimensional algorithm of incomplete mixed data set based on Artificial intelligence has extracted the features of incomplete mixed data sets and distinguished the mixed data sets in the subspace according to the correlation degree between mixed data sets in the multidimensional subspace in order to realize the high-dimensional clustering for incomplete mixed data sets and provide reference for related research in this field. The main contributions of this paper are as follows:

(1) Compared with other feature extraction methods, the proposed algorithm can extract key features under any degree of sparseness of incomplete mixed data sets distribution through introducing standard deviation calculating. It is the key of the extraction of incomplete mixed data set features. Besides, the data processing occupies a smaller memory.

(2) The algorithm carries out repeated quantitative analysis according to the dividing basis of density until it obtains the mixed data sets with a strong relevance. Then the clustering analysis algorithm is introduced. At this time, there is a strong relevance inside the mixed data sets. We can simplify the calculation process to reduce the clustering energy consumption of algorithm data set.

(3) As the algorithm cites twice the clustering analysis method in the design process, the sub-space has changed and the correlations factors changed accordingly. Hence, the test for the clustering effects in the experiment proves that the subspace clustering effect in this paper is better and can be applied in the practice.

The organizational structure of this paper firstly discusses the reconstruction of phase space, and introduces the correlation dimension to obtain the feature correlation value. Secondly, the standard deviation is introduced to calculate the sparseness of the sample features, and the key features of the incomplete mixed data set are extracted. Then, after extracting the key features, the clustering analysis of data sets in multi-dimensional subspace is described, and the clustering algorithms of data sets in different subspaces and the same subspace are given; Finally, repeated cluster analysis is performed on the mixed data sets in the subspace, and high-dimensional clustering is achieved based on the strong correlation between the data sets.

II. RELATED WORK

At present, there are many algorithms for high-dimensional clustering of incomplete mixed data sets. Literature [24] analyzes under the circumstance whether the use of emergent self-organizing feature mapping can avoid the integration of high-dimensional biomedical data clustering, that the interactive RI-based bioinformatics tools are used to recognize the subgroup structure directing different disease subtypes,

presents the data sets with different degrees of complexity to the analysis with a large number of neurons and visualises the distance structure in the high-dimensional feature space with u matrix. Literature [25] analyzes that the image data sets have many structures, therefore it is possible to disclose the binary grouping between images by projecting them into a random one-dimensional linear subspace. When the points to be clustered are complex corresponding to the image, on this basis, it presents a method of quantizing data sets clustering. Through comparing the clustering performance of image data sets and clusters generated by synthesis and verifying the clustering performance of the method projecting on the random line based on data, we get the conclusion according to probability density of measure: As the structures we found in the image data set do not meet the traditional clustering concept, we further propose a rapid method for high-dimensional data clustering with a layered approach. Literature [26] analyzes the background that the mining of uncertain data receives more and more attention, studies the problems of using the extremum learning machine to classify uncertain data and presents NU-ELM algorithm based on the uncertain data of non-uniform distribution. By calculating the probability critical value, we improved the efficiency of algorithm. Finally, through a large number of simulation experiments, we varied the effectiveness of the algorithm, which thus becomes an effective way of solving uncertain data classification, reducing execution time and improving efficiency. Literature [27] puts forth a Monte Carlo method to transform the high-dimensional potential energy surface on the discrete grid points into the form of product sum and then into more accurate Tucker form. With the variable method, it substitutes the numerical value accurate integration with Monte Carlo integration, which largely reduces the cost of numerical calculation and avoids the evaluation for potential on all grid points. Besides, it allows processing for surface with the degree of freedom up to 15 to 18, which further proves that the error of the method can be controlled and eliminated in a certain range. Literature [28] assumes that under the circumstance that the signal is in the union of sub space, the standard compression perception theory can carry out robustness recovery in low measurement quantity. However, it is limited to signal regularity with predetermined topology type and proposes a generalized model of decomposing the signals into data-driven subspace union for structured sparse representation, and thus the optimal structure and basis of subspace based on sample signals are obtained. Literature [29] uses association rule method to analyze the mixed data sets of retailers so as to guide category management, store layout and display, commodities promotion. But in the face of large amount volumes of e-commerce website, it still has the problem of low efficiency. In this end, we propose the fast clustering algorithm of sparse network of commodity related big data. First, we use the single step linked list structure to storage the joint purchase relationship matrix of retail goods. Second, we trim the low degree commodity nodes of sparse network of commodity related big data to reduce the searching space. Third, we use the fuzz K-means (FKM) clustering to conduct fast clustering for sparse network of commodity related big data and conduct clustering for remaining nodes with the thought of high connectivity value commodity node being divided by low connectivity value commodity nodes. Finally, we apply the proposed algorithm into the analysis of Amazon commodity transaction data and obtain good results. Literature [10] proposed a clustering algorithm for incomplete data in low-order subspaces, and achieved good results.

Although the above methods have obtained some research results, there are still some problems to be solved, such as clustering high energy consumption, bigger internal memory and poor clustering effects. Therefore, this paper proposes the high-dimensional clustering algorithm of incomplete mixed data set based on Artificial intelligence. First, we introduce correlation dimension and standard deviation to extract key features. On this basis, we realize data high-dimensional clustering in accordance with relevance of mixed data sets in the multidimensional subspace in order to improve the disadvantages of traditional clustering methods and we obtain better clustering results to support the research of data clustering.

III. High-DIMENSIONAL CLUSTERING ALGORITHM OF INCOMPLETE MIXED DATA SETS BASED ON ARTIFICIAL INTELLIGENCE

A. KEY FEATURE EXTRACTION OF INCOMPLETE MIXED DATA SETS

1) INTRODUCING CORRELATION DIMENSION TO OBTAIN FEATURE CORRELATION DIMENSION

In ordinary circumstances, incomplete mixed data sets are short of obvious rules and sequence and more complex. While the correlation dimension is representation of the distribution density of incomplete mixed data sets in the multidimensional space [30]. Therefore, we introduce correlation dimension to obtain the correlation value of key features of incomplete mixed data sets. In the process, we first construct the phase space reconstruction to use it to guarantee the invariance of incomplete mixed data set features. The specific process is as follows:

First, normally the sequence of incomplete mixed data sets is the non-linear time series [31]–[33]. As its focus is phase space reconstruction, we use it to guarantee the invariance of incomplete mixed data set features. The time series of incomplete mixed data set is represented by $\{q_1, q_2, \ldots, q_N\}$ and the formula of reconstructed phase space is:

$$Q = [Q_1, Q_2, \cdots, Q_N] = \begin{bmatrix} q_1 & q_2 & \cdots & q_N \\ q_{1+\tau} & q_{2+\tau} & \cdots & q_{N+\tau} \\ \vdots & \vdots & \ddots & \vdots \\ q_{1+(r-1)\tau} & q_{2+(r-1)\tau} & \cdots & q_{N+(r-1)\tau} \end{bmatrix} \quad (1)$$

where, τ represents time delay, and r represents the embedded dimension in the incomplete mixed data sets.

When $r \geq 2d'+1$, the characteristics of geometrical structure of incomplete mixed data set will be fully opened and d' represents the chaotic dimension of incomplete mixed data sets. As a mode of presentation of density degree of incomplete mixed data sets in the multidimensional space, correlation dimension mainly represents the relevance of incomplete mixed data set samples. We introduce a correlation dimension number for phase space reconstruction, so that we can get a phase space vector [34]–[36].

The max quantity difference of the Q_i sum Q_j of two random phase space vectors of incomplete mixed data sets represents the space between them. The formula is as below.

$$\left|Q_i - Q_j\right| = \max_{1\leq\delta\leq r} \left|Q_{i\delta} - Q_{j\delta}\right| \tag{2}$$

The space between them is lower than the vector of positive number l, it means there is a correlation vector. But the presence of correlation vector does not mean the incomplete mixed data sets have the relevance feature [24], [37], [38], and we go to the next step.

There are K points in the phase space of incomplete mixed data sets. We can use it to obtain the logarithm of relevant vectors of incomplete mixed data sets. The proportion of all the possible K^2 combinations, namely the correlation integral $S_k(l)$ is expressed as.

$$S_k(l) = \frac{1}{K^2} \sum_{i,j=1}^{k} H(\left|Q_i - Q_j\right|) \tag{3}$$

where, $H(\cdot)$ represents Heaviside function.

When $l \to 0$, the correlation dimension is introduced. In formula (3), there is some relevance between the correlation integrals $S_k(l)$ and l, shown as:

$$G = C \lim_{i\to\infty} S_k(l) \times 100\% \tag{4}$$

where, C represents the correlation dimension of incomplete mixed data sets. Selecting a reasonable l, and constructing the self similar structure of chaotic attrictor of incomplete mixed data sets, we get the approximate value C_l of feature correlation:

$$C_l = \frac{G \lg S_k(l)}{\lg l} \tag{5}$$

In practical application, when analyzing the double-logarithmic $\lg l - \lg S_k(l)$ of incomplete mixed data sets, we usually neglect lines with the slope of 0 or ∞ [39]–[41]. Selecting the best fitting straight line, the slope or the correlation value is expressed as C_l in the correlation formula.

2) INTRODUCE STANDARD DEVIATION TO COMPLETE KEY FEATURE EXTRACTION

Standard deviation is the representation of dispersion level of incomplete mixed data set sample points. When different data sample points of incomplete mixed data sets are not centrally distributed in space, namely the corresponding correlation dimension relatively low, the standard deviation of data samples of incomplete mixed data sets is large, specifically expressed as the sparseness of sample distribution of its correspondent incomplete mixed data sets. Therefore, when the feature correlation value is obtained, we introduce calculation of standard deviation as the key of feature extraction of incomplete mixed data sets and sparseness of processing incomplete mixed data sets.

Assume that the spatial scale of the incomplete mixed data set is in the range of $[Q_1, Q_N]$, using $p(q)$ to represent the key feature data values, and the standard deviation calculation formula is as follows:

$$\phi = \sqrt{\frac{1}{Q_1 \times Q_N} \sum_{q=Q_1}^{Q_N} \left(p(q) - \frac{1}{Q_1 \times Q_N} \sum_{q=Q_1}^{Q_N} p(q-1) \right)} \tag{6}$$

Then the distribution relationship X_i of the data set can be expressed as:

$$X_i = \alpha \frac{\phi \sigma_i}{C_p} \tag{7}$$

where, α represents the frequency doubling factor of incomplete mixed data sets. We make a new description for the sparseness of data set distribution relationship X_i according to the feature value σ_i processed by standard deviation of different incomplete mixed data set sample points.

Through the analysis above, we can regard the feature correlation value C_p as the sample points of incomplete mixed data sets, introduce standard deviation algorithm and conduct differential processing of the distribution relationship of incomplete mixed data sets, namely the sparseness to enlarge the correlation dimension and realize the extraction of key features of incomplete mixed data sets.

B. HIGH-DIMENSIONAL CLUSTERING ALGORITHM OF INCOMPLETE MIXED DATA SET BASED ON ARTIFICIAL INTELLIGENCE

On the basis of extracting key features of incomplete mixed data sets, we distinguish the mixed data sets in the subspace by using the relevance of mixed data sets in the multidimensional subspace so as to realize the high-dimensional clustering of incomplete mixed data sets [42], [43].

1) CLUSTERING ANALYSIS OF MULTIDIMENSIONAL SUBSPACE DATA SET

First, we use the clustering analysis algorithm of Artificial intelligence technology to distribute the data samples of incomplete mixed data sets in the multidimensional subspace. In the same subspace, the greater the relevance of two mixed data sets, the more suitable for clustering; when the relevance is small, we should conduct secondary division for data samples of mixed data sets in the same subspace to complete clustering [44], [45]. When the incomplete mixed data sets are in a different subspace, we need to calculate the distribution relationship of incomplete mixed data sets or the sparseness

for differential processing and identify the subspace according to the correlation property of subspace [46], [47].

Set up two mixed data sets V_i and V_k in two different subspaces M^i and M^k and use $D(i,k)$ to represent the Euclidean distance of two different subspaces and $d(i,k)$ to represent the Euclidean distance of two mixed data sets [48], [49]. The high-dimensional clustering formula of two mixed data sets of two different subspaces:

$$\begin{aligned} W(V^i, V^k) &= \left(M^i, M^k\right) \\ &= \frac{\varepsilon}{2}\begin{pmatrix} M^1 \\ \vdots \\ M^d \end{pmatrix} P(V_i)P(V_k) \\ &\quad \times \log_2 \sqrt{D(i,k)^2 + d(i,k)^2} \end{aligned} \tag{8}$$

where, ε represents the clustering factor of incomplete mixed data set subspace, $P(V_i)$ and $P(V_k)$ represent the clustering frequency of the incomplete mixed data sets V_i and V_k.

For clustering analysis of mixed data sets in the same subspace, it is necessary to distinguish between the data sets using the degree of correlation between the data sets. The correlation factor $g(i,k)$ of the mixed data sets V_i and V_k in the same space is:

$$\begin{aligned} g(i,k) = &\sqrt{\begin{pmatrix} x_{11} & \cdots & x_{i1} \\ \vdots & \cdots & \vdots \\ x_{k1} & \cdots & x_{ki} \end{pmatrix} d(i,k)} \\ &- \ln 2\frac{1}{m}\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \cdots & \vdots \\ x_{i1} & \cdots & x_{ik} \end{pmatrix} \end{aligned} \tag{9}$$

According to the correlation factor $g(i,k)$ obtained by the above formula, a high-dimensional clustering formula of two mixed data sets V_i and V_k in the same subspace is given, expressed as:

$$\begin{aligned} W(V^i, V^k) = &(P(V_i) - P(V_k))g(i,k)d(i,k)\times \\ &\frac{e^{g(i,k)}}{\begin{pmatrix} x_{11} & \cdots & x_{i1} \\ \vdots & \cdots & \vdots \\ x_{k1} & \cdots & x_{ki} \end{pmatrix}\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \cdots & \vdots \\ x_{i1} & \cdots & x_{ik} \end{pmatrix}} \end{aligned} \tag{10}$$

where, e represents the spatial correlation coefficient.

2) THE PROPOSED ALGORITHM

The specific steps of high-dimensional clustering algorithm of incomplete mixed data set based on Artificial intelligence are as follows:

Input: Incomplete mixed data sets, and input key features of data sets.

Output: Results of high-dimensional clustering of incomplete mixed data sets.

Initializing incomplete mixed data sets, specific steps of clustering as follows:

(1) Calculate the high-dimensional clustering results of the mixed data set in the subspace according to formulas (9) and (10);

(2) In actual sample data, the sample ranges of all feature distribution clusters are uneven, which results in the difference of sample density of all clusters. We need to divide the data density, namely the qualitative division based on density, divides into different types of data sets and recalculate correlation factor g. In the same space M^i, set the threshold $T(V)$ for the relevance of different incomplete mixed data sets. When the correlation factors of incomplete mixed data sets $g > T(V)$, it means the two incomplete mixed data sets have a stronger relevance and the high-dimensional clustering results of mixed data set V_i in space M^i can be expressed as follows:

$$\begin{aligned} f_{M^i}(V_i) = &\begin{pmatrix} V_1 & \cdots & V_n \\ \vdots & \cdots & \vdots \\ V_n & \cdots & V_{2n} \end{pmatrix} W(V^i, V^k) \\ &- \ln\left(\sqrt{\sum_{i=1,k=1}^{n} (P(V_i) - P(V_k))}\right) \end{aligned} \tag{11}$$

(3) When the correlation factors of mixed data set $g > T(V)$, it means the two mixed data sets have a weak relevance. We need to conduct secondary division back to the previous layer until we get the mixed data set with a stronger relevance, namely $g > T(V)$;

(4) We introduce clustering analysis algorithm again. Then there is a stronger relevance inside the mixed data sets and the high-dimensional clustering results of incomplete mixed data sets can be expressed as:

$$\begin{aligned} f_{M^i}(V_i) = &\frac{\begin{pmatrix} V_1 & \cdots & V_n \\ \vdots & \cdots & \vdots \\ V_n & \cdots & V_{2n} \end{pmatrix} W(V^i, V^k)}{(\pi e^2 - 1)} \\ &+ \frac{e}{2}\sum_{i=1,k=1}^{n} (P(V_i) - P(V_k)) \end{aligned} \tag{12}$$

(5) End.

To sum up, the flow chart of high-dimensional clustering of incomplete mixed data set based on Artificial intelligence is shown in Figure1.

Till now, we completed design of high-dimensional clustering algorithm of incomplete mixed data set based on Artificial intelligence.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL ENVIRONMENT AND DATA SET

To verify the comprehensive effectiveness of high-dimensional clustering algorithm of incomplete mixed data set based on Artificial intelligence, we need to conduct many experimental tests. The experimental procedure is written in C ++ and run on window 10, memory 4GB, CPU 2.89GHz. We compare the proposed algorithm with the experimental

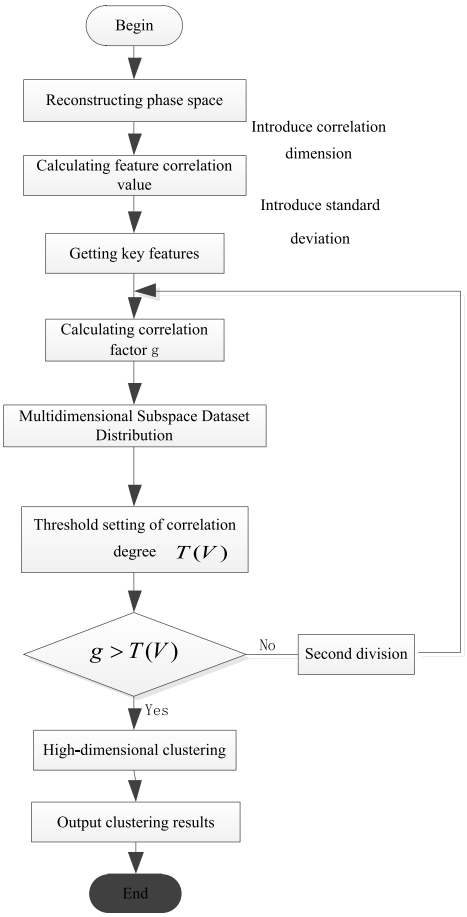


FIGURE 1. High-dimensional clustering of incomplete mixed data sets.

TABLE 1. Experimental data description.

data sets	Sample number / ten thousand	Training number	Test number
Glass	500	100	400
Ecoli	500	100	400
Segment	500	100	400
Oil	500	100	400

results in Literature [20], Literature [25], Literature [26], Literature [27] and Literature [28].

Data source is the UCI (http://archive.ics.uci.edu/ml/index.php)standard database. We select four data sets in the database Glass, Ecoli, Segment and Oil.

B. EXPERIMENTAL INDICATORS

We carry out experiment according to the given steps with the proposed algorithm. In this process, the verification of comprehensive effectiveness of high-dimensional clustering algorithm of incomplete mixed data set based on Artificial intelligence can set five experiment indicators:

1) THE PROCESSING EFFECTS OF CORRELATION DIMENSION

According to the text, in the processing effects of correlation dimension, the correlation plays an important role. We use the intensity of correlation distribution to judge the processing effects of correlation dimension. The more intensive the correlation distribution, the closer the connection degree and the better the processing effects of correlation dimension.

2) MEMORY USAGE

The formula for memory usage:

$$\rho(l) = \frac{AVE(l)}{C_l} \tag{13}$$

where, $\rho(l)$ is the memory of present data; C_l represents the data storage total load; $AVE(l)$ is the data actual load;

3) CALCULATION TIME OF STANDARD DEVIATION

When there is a weak correlation between correlation factors of incomplete mixed data sets, or the correlation factors are weak, we need to calculate the standard deviation of mixed data sets. The efficiency in the clustering phase is the main indicator to measure the entire efficiency of the algorithm. The efficiency is divided into speed and time-consuming. Under the data volume, high speed and low time-consuming, the calculation time-consuming of standard deviation is set as one of the indicators.

4) CLUSTERING ENERGY CONSUMPTION

Through the analysis of features of incomplete mixed data sets, the proposed algorithm mainly improves the problems in the relevance of nodes in the node establishment stage in order to distribute the node characteristics in the multidimensional subspace to each node evenly. As the energy consumption is different in the distribution process, the clustering energy consumption is set as one of the indicators.

5) EFFECTS OF SUBSPACE CLUSTERING

The cluster selected in the subspace clustering process is one of the evaluation standards of clustering effects. The clustering effect of each cluster in the subspace is the overall clustering effect of the subspace.

Using the data clustering success rate under different cluster conditions to verify the subspace clustering effect. The formula for calculating the clustering success rate is as follows:

$$f_{sucess} = \frac{Num_{same}}{Num_{total}} \tag{14}$$

where, Num_{same} is the number of times the same attribute data is clustered in the same category data set. Num_{total} is the total number of clusters of the data.

C. EXPERIMENTAL RESULTS

1) THE TEST FOR PROCESSING EFFECTS OF CORRELATION DIMENSION

The test for processing effects of correlation dimension is carried out for proposed algorithm and those in Literature [20],

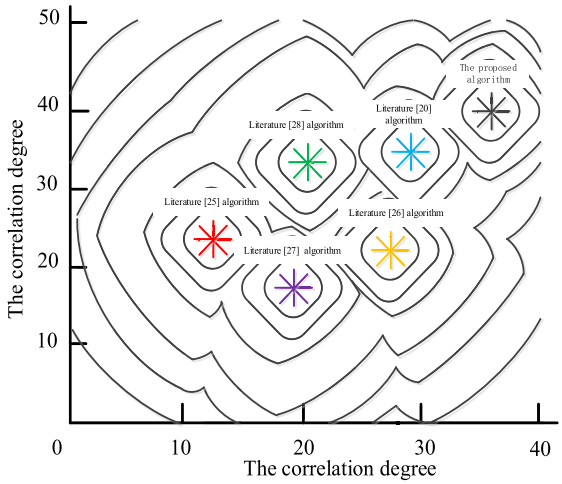


FIGURE 2. Test results of correlation dimension processing effect.

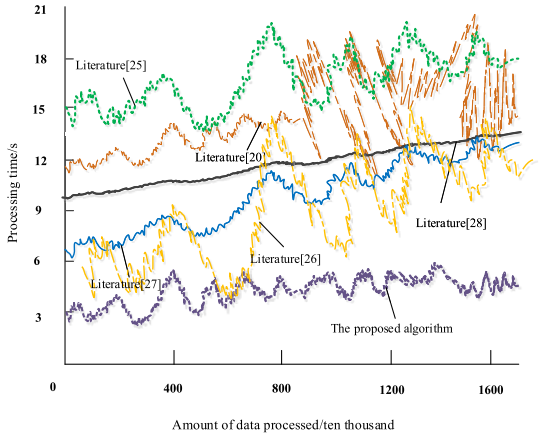


FIGURE 3. Comparison of standard deviation calculation time of different algorithms.

TABLE 2. Comparison of memory consumption of different clustering algorithms.

Data/ten thousand	CPU/MB					
	A	B	C	D	E	F
400	45	75	70	71	65	69
800	58	72	85	80	59	75
1200	65	70	92	75	67	80
1600	88	102	125	110	90	111

Literature [25], Literature [26], Literature [27] and Literature [28]. It is shown in Figure2.

From Figure 2, we can see that the connection of correlation degree of the proposed algorithm is the most intensive and the number is the largest, which means that the proposed algorithm has a great advantage in the test of processing effects of correlation dimension and good processing effects of correlation dimension. This is because before using the correlation dimension to obtain the data eigenvalue, we first construct the phase space reconstruction to ensure the invariance of data set features.

2) TEST OF MEMORY USAGE

Contrast experiment for clustering memory usage (MB) is carried out between the proposed algorithm and those in Literature [20], Literature [25], Literature [26], Literature [27] and Literature [28]. The experimental results are shown in table2. In table2, CPU represents memory usage, unit MB; A represents the proposed algorithm; B represents algorithm in Literature [20], C represents algorithm in Literature [25], D represents the algorithm in Literature [26], E represents the algorithm in Literature [27] and F represents the algorithm in Literature [28].

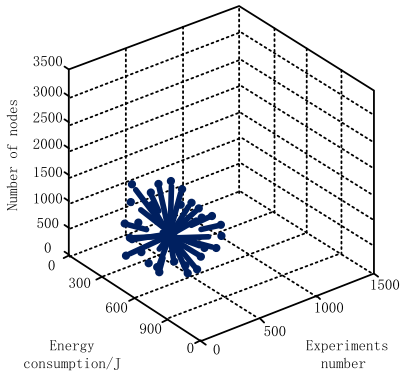
Table2 shows the clustering memory usage under different data sets. The less the clustering memory usage, the better the performance of the data set clustering of the algorithm and the more suitable for clustering for incomplete mixed data sets. From table 2, we can see the memory usage of the proposed algorithm is less, but the memory usage of algorithms in Literature [20], Literature [25], Literature [26], Literature [27] and Literature [28] used a large amount of memory for algorithm clustering, with a maximum of 111MB. Therefore, the proposed algorithm has great advantages for high-dimensional clustering of incomplete mixed data sets.

This is because the phase space reconstruction is constructed in this paper to ensure the invariance of incomplete mixed data set features, avoiding the change of data set characteristics, clustering is very difficult and takes up a lot of memory, therefore it can save the memory occupied by the algorithm, which is better than other algorithms.

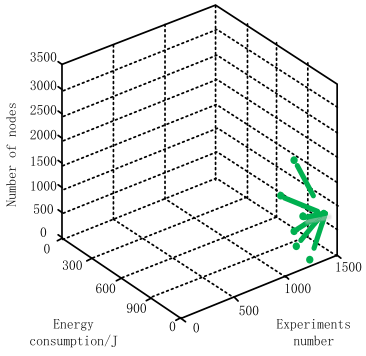
3) COMPARISON OF STANDARD DEVIATION CALCULATION TIME

The contrast experiment for calculation time of standard deviation is carried out between the proposed algorithm and those in Literature [20], Literature [25], Literature [26], Literature [27], Literature [28] and the results are shown in Figure3.

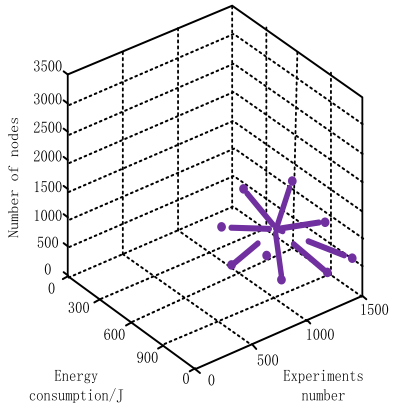
From Figure3, we can see that when the four algorithms are in the same data quantity, the calculation time of standard deviation of the proposed algorithm is lower that of the algorithms in Literature [20], Literature [25], Literature [26], Literature [27] and Literature [28], no more than 6s. The calculation time in Literature [25] is the highest and the calculation time in Literature [20] is up to 20s and not stable. The calculation time in Literature [26] and [27] is high. Although the calculation time in Literature [28] is stable but also higher than that of the proposed algorithm, averagely 12s. Through comparison, we see that the proposed algorithm has higher calculation efficiency of standard deviation and we verified that the proposed algorithm is effective. It shows that the proposed algorithm calculates the standard deviation with sparseness of data features and receives good results, which directly influences the data clustering efficiency and reduces time consuming.



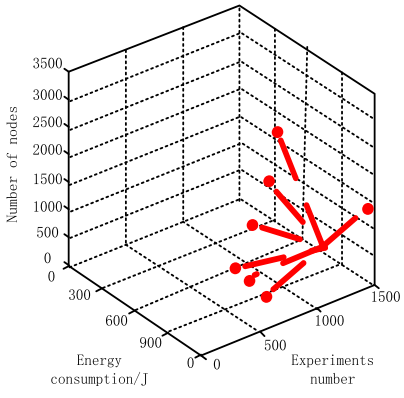
（a）The proposed algorithm



（b）Literature [20] algorithm



（c）Literature [25] algorithm



（d）Literature [26] algorithm

FIGURE 4. Comparison experiment of clustering energy consumption of different algorithms.

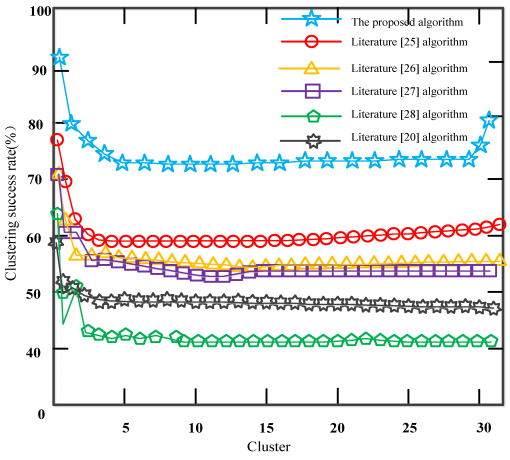


FIGURE 5. Comparison results of spatial clustering effect.

4) COMPARISON OF CLUSTERING ENERGY CONSUMPTION

The contrast experiment for clustering energy consumption is conducted between the proposed algorithm and those in Literature [20], Literature [25] and Literature [26] and the results are shown in Figure4.

From Figure4, it is shown that for nodes clustering, we can get the clustering effects of the above five algorithms. The clustering energy consumption of the proposed algorithm is obviously low and centrally distributed, with energy consumption within 3000J. But the energy consumption in Literature [20], Literature [25] and Literature [26] is more than 300J. The proposed algorithm is obviously lower than the algorithms in Literature [20], Literature [25] and Literature [26] in energy consumption. It is mainly because the proposed algorithm conducts repeated clustering for the incomplete mixed data sets in subspace based on the relevance and realizes clustering in data space with a stronger relevance. The algorithm is easy to realize and has higher efficiency for high-dimensional clustering of incomplete mixed data sets, therefore its energy consumption of clustering is low. It has the practicality..

5) COMPARISON OF SUBSPACE CLUSTERING EFFECTS

The cluster selected in the process of subspace clustering is one of the evaluation standards of clustering effects. We conduct comparison of subspace clustering effects of the algorithm with those in Literature [20], Literature [25], Literature [26], Literature [27] and Literature [28]. The results are shown in Figure5.

In Figure5, with the increase of the number of clusters in subspace, the curves of all algorithms also change variously. Compared with other algorithms, the proposed algorithm decreases slowly and has good clustering effects. The clustering success rate is up to 92%. Next is the algorithm in Literature [25], up to 78%, and the clustering success rate in Literature [25] and Literature [26] is less than 60%, and that in Literature [20] and Literature [28] is the lowest, Literature [20] is about 50%, and the clustering success rate of

Literature [28] is about 42%. It shows that the proposed algorithm has obvious advantages in subspace clustering effects. The reason is that the proposed algorithm has used twice the clustering analysis method in the research process, which changes the subspace and the correlation factors. Therefore, it obtains better clustering effects.

V. DISCUSSION

We propose a high-dimensional clustering algorithm for incomplete mixed data set based on artificial intelligence. The characteristics of incomplete mixed data sets are extracted, and the mixed data sets in the subspace are distinguished according to the degree of association between the mixed data sets in the multidimensional subspace, so as to achieve high-dimensional clustering of the incomplete mixed data sets. The proposed algorithm has better correlation dimension processing effect, with lower memory consumption and time consumption, and the cluster energy consumption is concentrated and lower.

Cluster analysis can be used as an independent tool to obtain the distribution of the data. By observing the characteristics of each cluster, focus on specific clusters for further analysis to obtain the required information. The application of cluster analysis in data mining, pattern recognition, image processing, computer vision and other fields has attracted attention. Therefore, high-dimensional clustering of incomplete mixed data sets in this paper is of great significance, and some related literature has achieved some results. To tackle this challenging problem, [50] proposes a novel intelligent weighting k-means clustering (IWKM) algorithm based on swarm intelligence. Finally verify the clustering performance of high-dimensional multi-view data. The setting the coefficients of criteria items without prior knowledge will lead to inaccurate and poor robust clustering results. To address this problem, [51] propose to optimize the multiple clustering criteria simultaneously without any predefined coefficients by a multi-objective evolutionary algorithm. These literatures have achieved certain results, but compared with the proposed algorithm, it has obvious advantages in clustering with high energy consumption, high memory consumption, and low clustering effects, and has certain advantages and practical application value.

However, the research still has some deficiencies. It has not carried out identification research for the features of incomplete mixed data. In the future, we can further combine the algorithm with classifier of higher classification accuracy for detailed analysis of features of incomplete mixed data and research of the high-dimensional clustering of incomplete mixed data in order to lay a foundation for further research.

VI. CONCLUSION

The current high-dimensional algorithms can group data on one level fairly well, but they are limited to handling linear structure or overlapping non-linear structure. Their needs for theoretical and practical incomplete mixed data sets are real. Therefore, this paper puts forth the high-dimensional clustering algorithm of incomplete mixed data set based on artificial intelligence and carries out test experiments in processing effects of correlation dimension, memory usage, calculation time of standard deviation and subspace clustering effects. The experimental results show that the algorithm is much better than the algorithms in other literature, with better clustering effects. Its theory application value provides a certain reference for the research in this field.

REFERENCES

[1] S. Bano and M. N. A. Khan, "A survey of data clustering methods," *Int. J. Adv. Sci. Technol.*, vol. 113, pp. 133–142, Apr. 2018.

[2] G. Jie, W. Jia, and Z. Yang, "Low frequency oscillation modal parameter identification based on NExT-ERA and fuzzy clustering," *Int. J. Control Autom.*, vol. 9, no. 1, pp. 309–322, Jan. 2016.

[3] X. Chen and Y. Liu, "Correlation coefficients of intuitionistic hesitant fuzzy sets and their applications to clustering analysis," *Int. J. Control Autom.*, vol. 9, no. 8, pp. 403–418, Aug. 2016.

[4] Y. Zhou, H.-F. Zuo, and J. He, "Aero-engine fault diagnosis using a feature weighting fuzzy clustering algorithm," *Int. J. Control Autom.*, vol. 10, no. 5, pp. 161–168, May 2017.

[5] Y.-G. Wang and S.-C. Xiu, "Environmental impact evaluation method of grinding process using clustering and ANFIS," *Int. J. Control Autom.*, vol. 10, no. 8, pp. 171–182, Aug. 2017.

[6] H. Gao, Y. Xu, Y. Yin, W. Zhang, R. Li, and X. Wang, "Context-aware QoS prediction with neural collaborative filtering for Internet-of-Things services," *IEEE Internet Things J.*, early access, Dec. 2, 2019, doi: 10.1109/JIOT.2019.2956827.

[7] M. Bidaki and S. R. K. Tabbakh, "Efficient fuzzy logic-based clustering algorithm for wireless sensor networks," *Int. J. Grid Distrib. Comput.*, vol. 9, no. 5, pp. 79–88, May 2016.

[8] X.-W. Zhang and J. Liang, "Multiple smooth support vector machine with FCM clustering in hidden space," *Int. J. Grid Distrib. Comput.*, vol. 9, no. 9, pp. 129–136, Sep. 2016.

[9] D. Pimentel, R. Nowak, and L. Balzano, "On the sample complexity of subspace clustering with missing data," in *Proc. IEEE Workshop Stat. Signal Process. (SSP)*, Gold Coast, VIC, Australia, Jun. 2014, pp. 280–283.

[10] M. Ashraphijuo and X. Wang, "Clustering a union of low-rank subspaces of different dimensions with missing data," *Pattern Recognit. Lett.*, vol. 120, no. 4, pp. 31–35, Apr. 2019.

[11] E. Guerrini, L. Imbert, and T. Winterhalter, "Randomized mixed-radix scalar multiplication," *IEEE Trans. Comput.*, vol. 67, no. 3, pp. 418–431, Mar. 2018.

[12] S. A. Alabady and S. Raed, "MHUCR: Mutli hop uniform clustering routing protocol for energy efficient WSN," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 6, pp. 95–106, Jun. 2018.

[13] K. S. Rao, K. V. Satyanarayana, and P. S. Rao, "Segmentation of images using two parameter logistic type distribution and K-means clustering," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 20, pp. 1–20, 2018.

[14] J. Zhang, J. Xiao, J. Wan, J. Yang, Y. Ren, H. Si, L. Zhou, and H. Tu, "A parallel strategy for convolutional neural network based on heterogeneous cluster for mobile information system," *Mobile Inf. Syst.*, vol. 2017, pp. 1–12, 2017.

[15] Y. Yin, L. Chen, Y. Xu, J. Wan, H. Zhang, and Z. Mai, "QoS prediction for service recommendation with deep feature learning in edge computing environment," *Mobile Netw. Appl.*, Apr. 2019, doi: 10.1007/s11036-019-01241-7.

[16] Y. Jin, X. Guo, Y. Li, J. Xing, and H. Tian, "Towards stabilizing facial landmark detection and tracking via hierarchical filtering: A new method," *J. Franklin Inst.*, Jan. 2020, doi: 10.1016/j.jfranklin.2019.12.043.

[17] Y. Yin, J. Xia, Y. Li, Y. Xu, W. Xu, and L. Yu, "Group-wise itinerary planning in temporary," *Mobile Social Netw.*, vol. 7, no. 1, pp. 83682–83693, 2019.

[18] J. Yu, M. Tan, H. Zhang, D. Tao, and Y. Rui, "Hierarchical deep click feature prediction for fine-grained image recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 30, 2019, doi: 10.1109/.2019.2932058.

[19] A. Pesevski, B. C. Franczak, and P. D. McNicholas, "Subspace clustering with the multivariate-t distribution," *Pattern Recognit. Lett.*, vol. 112, pp. 297–302, Sep. 2018.

[20] W. Zhong, D. Tan, X. Peng, Y. Tang, and W. He, "Fuzzy high-order hybrid clustering algorithm for swarm intelligence sets," *Neurocomputing*, vol. 314, no. 11, pp. 347–359, Nov. 2018.
[21] A. Sharma, P. J. Kamola, and T. Tsunoda, "2D–EM clustering approach for high-dimensional data through folding feature vectors," *BMC Bioinf.*, vol. 18, no. S16, pp. 547–552, Dec. 2017.
[22] B. Liu, X.-T. Yuan, Y. Yu, Q. Liu, and D. N. Metaxas, "Parallel sparse subspace clustering via joint sample and parameter blockwise partition," *ACM Trans. Embedded Comput. Syst.*, vol. 16, no. 3, pp. 1–17, Jul. 2017.
[23] A. Ultsch and J. Lötsch, "Machine-learned cluster identification in high-dimensional data," *J. Biomed. Informat.*, vol. 66, no. 2, pp. 95–104, Feb. 2017.
[24] T. Yellamraju and M. Boutin, "Clusterability and clustering of images and other 'real' high-dimensional data," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1927–1938, Jan. 2018.
[25] C. Rong-Hua, C. Yuan, and Z. Su-Xia, "Uncertain data clustering algorithm based on fast Gaussian transform," *Chin. J. Commun.*, vol. 38, no. 3, pp. 101–111, 2017.
[26] K. Cao, G. Wang, and D. Han, "An Algorithm for Classification over Uncertain Data Based on Extreme Learning Machine," in *Proceedings of ELM-2014 Volume 1* (Proceedings in Adaptation, Learning and Optimization), vol. 3, J. Cao, K. Mao, E. Cambria, Z. Man, and K. A. Toh, Eds. Cham, Switzerland: Springer, 2016, pp. 193–202.
[27] M. Schröder and H.-D. Meyer, "Transforming high-dimensional potential energy surfaces into sum-of-products form using Monte Carlo methods," *J. Chem. Phys.*, vol. 147, no. 6, Aug. 2017, Art. no. 064105.
[28] Y. Li, W. Dai, J. Zou, H. Xiong, and Y. F. Zheng, "Structured sparse representation with union of data-driven linear and multilinear subspaces model for compressive video sampling," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5062–5077, Oct. 2017.
[29] T.-Y. Li, F. Li, and Y. Chen, "Fast clustering algorithm for retail commodity associated big data sparse networks," *Control Decis.*, vol. 33, no. 6, pp. 152–157, 2018.
[30] Y. Zhang, J. Yang, and H. Hou, "The underwater acoustic target recognition algorithm based on evidence clustering," *Xibei Gongye Daxue Xuebao/J. Northwestern Polytechn. Univ.*, vol. 36, no. 1, pp. 96–102, Feb. 2018.
[31] X. Ma, H. Gao, H. Xu, and M. Bian, "An IoT-based task scheduling optimization scheme considering the deadline and cost-aware scientific workflow for cloud computing," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, Dec. 2019 Art. no. 249, doi: 10.1186/s13638-019-1557-3.
[32] A. Belesiotis, D. Skoutas, C. Efstathiades, V. Kaffes, and D. Pfoser, "Spatio-textual user matching and clustering based on set similarity joins," *VLDB J.*, vol. 27, no. 3, pp. 297–320, Jun. 2018.
[33] C.-C. Hsu and C.-W. Lin, "CNN-based joint clustering and representation learning with feature drift compensation for large-scale image data," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 421–429, Feb. 2018.
[34] T. Xin-min, L. Chen-xi, and S. Wei, "Unbalanced data classification algorithm based on density sensitive maximum soft interval SVDD," *Chin. J. Electron.*, vol. 46, no. 11, pp. 2725–2732, 2018.
[35] J. Li, X. Zhang, Z. Wang, X. Chen, J. Chen, Y. Li, and A. Zhang, "Dual-band eight-antenna array design for MIMO applications in 5G mobile terminals," *IEEE Access*, vol. 7, pp. 71636–71644, 2019.
[36] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 2, pp. 661–674, Feb. 2020, doi: 10.1109/TNNLS.2019.2908982.
[37] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Oct. 15, 2019, doi: 10.1109/TCSVT.2019.2947482.
[38] Y. Wang, J. Li, F. Nie, H. Theisel, M. Gong, and D. J. Lehmann, "Linear discriminative star coordinates for exploring class and cluster separation of high dimensional data," *Comput. Graph. Forum*, vol. 36, no. 3, pp. 401–410, Jun. 2017.
[39] K. Gong, Y. Wang, M. Xu, and Z. Xiao, "BSSReduce an $O(|U|)$ incremental feature selection approach for large-scale and high-dimensional data," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 6, pp. 3356–3367, Dec. 2018.
[40] H. Gao, W. Huang, Y. Duan, X. Yang, and Q. Zou, "Research on cost-driven services composition in an uncertain environment," *J. Internet Technol.*, vol. 20, no. 3, pp. 755–769, 2019.
[41] B. T. Lau, C. Wood-Bouwens, and H. P. Ji, "Robust multiplexed clustering and denoising of digital PCR assays by data gridding," *Anal. Chem.*, vol. 89, no. 22, pp. 11913–11917, Nov. 2017.
[42] G. Jia, G. Han, H. Wang, and F. Wang, "Cost aware cache replacement policy in shared last-level cache for hybrid memory based fog computing," *Enterprise Inf. Syst.*, vol. 12, no. 4, pp. 435–451, Apr. 2018.
[43] H. Si, Z. Chen, W. Zhang, J. Wan, J. Zhang, and N. N. Xiong, "A member recognition approach for specific organizations based on relationships among users in social networking Twitter," *Future Gener. Comput. Syst.*, vol. 92, pp. 1009–1020, Mar. 2019.
[44] T.-Y. Qian, B. Liu, L. Hong, and Z.-N. You, "Time and location aware points of interest recommendation in location-based social networks," *J. Comput. Sci. Technol.*, vol. 33, no. 6, pp. 1219–1230, Nov. 2018.
[45] P. Rathore, J. C. Bezdek, S. M. Erfani, S. Rajasegarar, and M. Palaniswami, "Ensemble fuzzy clustering using cumulative aggregation on random projections," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 3, pp. 1510–1524, Jun. 2018.
[46] X. Li, Y. Wang, and D. Li, "Medical data stream distribution pattern association rule mining algorithm based on density estimation," *IEEE Access*, vol. 7, pp. 141319–141329, 2019.
[47] Q. Zhang, L. T. Yang, Z. Chen, and F. Xia, "A high-order possibilistic C-means algorithm for clustering incomplete multimedia data," *IEEE Syst. J.*, vol. 11, no. 4, pp. 2160–2169, Dec. 2017.
[48] S. C. Slaoui, Z. Dafir, and Y. Lamari, "E-transitive: An enhanced version of the transitive heuristic for clustering categorical data," *Procedia Comput. Sci.*, vol. 127, pp. 26–34, 2018, doi: 10.1016/j.procs.2018.01.094.
[49] G. Fusco, A. Bracci, T. Caligiuri, C. Colombaroni, and N. Isaenko, "Experimental analyses and clustering of travel choice behaviours by floating car big data in a large urban area," *IET Intell. Transp. Syst.*, vol. 12, no. 4, pp. 270–278, May 2018.
[50] Q. Tao, C. Gu, Z. Wang, and D. Jiang, "An intelligent clustering algorithm for high-dimensional multiview data in big data applications," *Neurocomputing*, Jul. 2019, doi: 10.1016/j.neucom.2018.12.093.
[51] C. Liu, Y. Li, Q. Zhao, and C. Liu, "Reference vector-based multi-objective clustering for high-dimensional data," *Appl. Soft Comput.*, vol. 78, pp. 614–629, May 2019.

MEISHAN LI was born in Heilongjiang, China, in 1982. She received the B.S. and M.E. degrees from the College of Computer Science and Information Engineering, Harbin Normal University, Harbin, China, in 2006 and 2009, respectively. She currently works with the College of Information and Electronic Technology, Jiamusi University. Her research interests include artificial intelligence, and vision and image processing.

XIAOFENG LI (Member, IEEE) received the Ph.D. degree from the Beijing Institute of Technology. He is currently a Professor with Heilongjiang International University. He has published more than 50 academic articles at home and abroad, and has been indexed and collected more than 30 articles by SCI, EI. His research interests include data mining, intelligent transportation, artificial intelligence, intelligent medical, and sports engineering. He is a member of ACM and an Advanced Member of CCF.

JING LI was born in 1968. She is currently a Professor with the College of Information and Electronic Technology, Jiamusi University. Her research interests include machine learning and data privacy.