IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 April 2023, accepted 24 April 2023, date of publication 4 May 2023, date of current version 21 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3273010

RESEARCH ARTICLE

A Novel Hybrid SBM Clustering Method Based on Fuzzy Time Series

REN-LONG ZHANG AND XIAO-HONG LIU
School of Management, Guizhou University, Guiyang 550025, China

Corresponding author: Xiao-Hong Liu (liuxiaohonggj@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 72261005; in part by the Guizhou Science and Technology Planning Projects under Grant ZK2021G339 and Grant ZK2022G080; in part by the Guizhou Philosophy and Social Science Planning Projects under Grant 21GZYB09 and Grant 21GZYB10; in part by the Research Base and Key Special Topics of Think Tanks under Grant GDZX2021031; in part by the General Project of Liberal Arts Research of Guizhou University under Grant GDYB2021022 and Grant GDYB2021023; in part by the Talent Introduction Project of Guizhou University under Grant 2019016, Grant 2019017, Grant 20GLR001, and Grant 20GLR002; and in part by the 2022 National Social Science Fund Cultivation Project under Grant GDPY2021014 and Grant GDPY2021015.

ABSTRACT With the development of machine learning algorithm and fuzzy theory, the fuzzy clustering algorithm based on time series has received more and more attention. Based on the time series theory and considering the correlation of data attributes, it proposes a novel multivariate fuzzy time series clustering method based on Slacks Based Measure (MFTS-SBM). Compared with traditional fuzzy clustering that it has the ability to deal with fuzziness and uncertainty, the proposed hybrid SBM clustering method employs with input and output items and considers the clustering results and the influencing factors of nonparametric frontier. Thus, it is important for data decision making because decision makers are interested in understanding the changes required to combine input variables in order to classify them into the desired clusters. The simulation experiment results of different samples are given to explain the use and effectiveness of the proposed hybrid SBM clustering method. Therefore, the hybrid method has strong theoretical significance and practical value.

INDEX TERMS Fuzzy time series, SBM, nonparametric frontier, clustering algorithm.

I. INTRODUCTION

With the wide application of time series in various fields, the time series clustering can mine the internal structural characteristics of complex data by processing and analyzing, thus providing a scientific and effective basis for decision-making analysis in economic society. Compared with the traditional time series method, the fuzzy time series clustering model is mainly characterized by dealing with fuzzy sets. Generally, this feature makes it have significant clustering advantages when modeling fuzzy and uncertain data. Simultaneously, the traditional clustering algorithms have great limitations in application when modeling inaccurate and fuzzy data. Model-based time series clustering requires a parametric model with sufficient flexibility to describe the dynamics in various time series. The results showed that the method is more accurate than those obtained using the previous methods [1]. Considering the accuracy of clustering, a new time series clustering method based on non-normality and model nonlinearity is proposed [2]. The robust fuzzy clustering methods are applied to the coefficients of B-spline fitting providing feasible algorithms to implement these methods [3]. The fuzzy based classification algorithms show better accuracy than others which reveals the importance of this novel time series primitives in time series feature learning [4]. Therefore, we will compare the proposed method to other clustering methods and the results indicate that the proposed method can improve clustering accuracy for multivariate time series datasets. How to choose the optimal time series clustering method based on practical problems is the problem that needs further exploration.

At present, frontier based clustering methods are gradually attracting attention and now it has formed a mature set of theories. Although the modeling is simple, the calculation

The associate editor coordinating the review of this manuscript and approving it for publication was Easter Selvan Suviseshamuthu.

is convenient and the results have good mathematical interpretation, there are still some technical bottlenecks to be further solved. Fuzzy time series clustering algorithm can be constructed according to different complex data, as well as the different data fuzzification and fuzzy relations. Building clustering algorithms for fuzzy time series data considering the impact of multiple-feature factors is worth in-depth research. The density peak clustering algorithm is proposed to deal with the time series data and the experimental results demonstrate that the clustering performance of the proposed clustering algorithm [5]. Aiming at single dimension large sample data, multi-objective optimization algorithm and kernel fuzzy C-means clustering are adopted to realize data fuzzification. For single dimension small sample data, it is obviously unscientific to adopt multi-objective clustering algorithm due to the limitation of sample size. In addition, a time series clustering algorithm based on the combination of univariate and multivariate wavelet features is proposed and the effectiveness of the algorithm is verified [6]. Therefore, for single dimension small sample data, we should adopt the domain division method based on information optimization technology and information granularity. At the same time, it is necessary to fully tap the potential effective information of samples and establish the accuracy of fuzzy relations based on data with different distribution characteristics. Consequently, the ability and effective performance of the hybrid method for data clustering can be improved.

II. SBM MODEL

Data envelopment analysis (DEA) based on nonparametric frontier is a nonparametric linear programming method to evaluate the relative efficiency of a group of homogeneous production units with multiple inputs and outputs. In the expansion and application of DEA of non-radial distance function, it is of the great theoretical and practical value to solve such problems as how to select the appropriate direction to evaluate the efficiency of decision making units and how to evaluate and rank the efficiency of decision making units when the input/output is negative. Different DEA model with relaxation variables may reflect different effects. DEA model based on Slacks Based Measure (SBM) is considered to comprehensively evaluate the efficiency of input and output.At the same time, the model can be used to calculate the efficiency of DMU and obtain the relevant information of each input and output relaxation variable.Let's assume that the optimal variable ρ^* of the SBM model is not greater than the optimal variable θ^* and the θ^* of the DEA-CCR model.Therefore, the linear constraints of the DEA-CCR model with variable inputs can be expressed as follows.

$$\begin{cases} \theta x_0 = X\mu + t_1^- \\ \theta z_0 = Z\mu + t_2^- \\ y_0 = Y\mu - t_3^+ \\ b_0 = B\mu - t_4^+ \end{cases} \tag{1}$$

Similarly, the SBM-CCR model achieves the optimal solution $(\theta^*, \mu^*, t^{-*}, t^{+*})$.In particular, we examine the fundamental CCR model to set up the DEA-CCR clustering approach that it can be easily extended to other SBM. Through analysis, we add both sides of the equation at the same time and the following results can be obtained by item shifting and the specific situation can be expressed as follows.

$$\begin{cases} x_0 = X\mu^* + t_1^{-*} + (1-\theta^*)x_0 \\ z_0 = Z\mu^* + t_2^{-*} + (1-\theta^*)z_0 \\ y_0 = Y\mu^* - t_3^{+*} \\ b_0 = B\mu^* - t_4^{+*} \end{cases} \tag{2}$$

As a result, we can use the relevant definitions of the following variables. $\lambda = \mu^*, s_1^- = t_1^{-*} + (1-\theta^*)x_0, s_2^- = t_2^{-*} + (1-\theta^*)x_0, s_3^+ = t_3^{+*}$ and $s_4^+ = t_4^{+*}$.The variables such as $(\lambda, s_1^-, s_2^-, s_3^+, s_4^+)$ are the variables of SBM model respectively. Therefore, the efficiency can be measured quantitatively with the following formula.

$$\begin{aligned} \rho &= \frac{1 - \frac{1}{m+t}\left(\sum_{i=1}^{m}\frac{s_i^-}{x_{io}} + \sum_{i=1}^{t}\frac{s_i^-}{z_{io}} + (m+t)(1-\theta^*)\right)}{1 + \frac{1}{s+p}\left(\sum_{r=1}^{s}\frac{s_r^+}{y_{ro}} + \sum_{r=1}^{p}\frac{s_r^+}{b_{ro}}\right)} \\ &= \frac{\theta^* - \frac{1}{m+t}\left(\sum_{i=1}^{m}\frac{s_i^-}{x_{io}} + \sum_{i=1}^{t}\frac{s_i^-}{z_{io}}\right)}{1 + \frac{1}{s+p}\left(\sum_{r=1}^{s}\frac{s_r^+}{y_{ro}} + \sum_{r=1}^{p}\frac{s_r^+}{b_{ro}}\right)} \end{aligned} \tag{3}$$

Therefore, we can deduce that the following interval representation occurs.

$$\begin{aligned} \rho^* \leq \rho &= \frac{\theta^* - \frac{1}{m+t}\left(\sum_{i=1}^{m}\frac{s_i^-}{x_{io}} + \sum_{i=1}^{t}\frac{s_i^-}{z_{io}}\right)^*}{1 + \frac{1}{s+p}\left(\sum_{r=1}^{s}\frac{s_r^+}{y_{ro}} + \sum_{r=1}^{p}\frac{s_r^+}{b_{ro}}\right)} \\ &\leq \theta^* - \frac{1}{m+t}\left(\sum_{i=1}^{m}\frac{s_i^-}{x_{io}} + \sum_{i=1}^{t}\frac{s_i^-}{z_{io}}\right) \leq \theta^* \end{aligned} \tag{4}$$

We all know that the inputs and outputs of DMUs are measurable. Considering multiple DMUs, each DMU has m inputs and s outputs even though it may be in varying amounts. The vector has the following characteristics as follows. $x_j = (x_{1j}, \cdots, x_{mj})$ and $y_j = (y_{1j}, \cdots, y_{sj})$ and $z_j = (z_{1j}, \cdots, z_{pj})$. At the same time, they represent input, output and environmental variables of decision making units respectively $(j = 1, \cdots, n)$. It is important to point out the traditional data envelopment analysis model requires that both input and output are accurate. Therefore, we sometimes cannot obtain accurate input and output data for the number of DMUs is usually much larger than the number of inputs in practical applications. Combining the limitations of current clustering methods, it makes a deep exploration on the clustering methods based on nonparametric frontier from the aspects of efficiency measurement and frontier construction and the clustering rules. Moreover, the model should also consider the unexpected output, so we should build a new efficiency model that considers the unexpected output. Consequently, through the above analysis, the SBM model can be represented as follows.

$$
\begin{aligned}
&min\ \rho = \frac{1-\frac{1}{m}\sum\limits_{i=1}^{m}\frac{s_i^-}{x_{i0}}}{1+\frac{1}{s}\sum\limits_{r=1}^{s}\frac{s_r^+}{y_{r0}}}\\
&s.t.\quad \sum_{j=1}^{n}\lambda_j y_{r0} - s_r^+ = y_{r0},\ \forall r\\
&\qquad\ \sum_{j=1}^{n}\lambda_j x_{i0} + s_i^- = x_{i0},\ \forall i\\
&\qquad\ \forall j = 1,\cdots,n,\quad s_i^-, s_r^+ \geq 0
\end{aligned}
\tag{5}
$$

If the following conditions exist $1+\frac{1}{s}\sum\limits_{r=1}^{s}\frac{s_r^+}{y_{r0}} = \frac{1}{t}$, we can convert the objective function of the SBM model into the following model considering unexpected output such as $min\ \rho = t * (1-\frac{1}{m}\sum\limits_{i=1}^{m}\frac{s_i^-}{x_{i0}}) = t - \frac{1}{m}\sum\limits_{i=1}^{m}\frac{S_i^-}{x_{i0}}$. Through the sequential reasoning, the model can be transformed into the following linear model as shown as follows.

$$
\begin{aligned}
&min\ q = t - \frac{1}{m}\sum_{i=1}^{m}\frac{S_i^-}{x_{i0}}\\
&s.t.\quad 1 = t + \frac{1}{s}\sum_{r=1}^{s}\frac{S_r^+}{y_{r0}}\\
&\qquad\ t x_{i0} = \sum_{j=1}^{n}\Lambda_j x_{ij} + S_i^-\\
&\qquad\ t y_{r0} = \sum_{j=1}^{n}\Lambda_j y_{rj} - S_r^+
\end{aligned}
\tag{6}
$$

where $\rho = q, \lambda_j = \Lambda_j/t, s_i^- = S_i^-/t, s_r^+ = S_r^+/t$. Any DMU in the SBM is CCR valid if and only if it is SBM valid. And the optimal variable of the SBM model is q^* not that it is greater than the optimal variable θ^* of the CCR model. In our study, efficiency scores and relaxation variables are based on variables in the SBM model. At the same time, these relaxation variables in the mixed model can well explain the degree of ineffectiveness of DMUs. We can use the cluster center of each category as the initial cluster center of fuzzy clustering of time series data. At the same time, each category of all test time series data can be determined according to the maximum value principle determined by SBM.

On the other hand, the SBM clustering Method Based on Fuzzy Time Series can be seen as a feature selection analysis technique. A modified slacks-based super-efficiency measure in the presence of negative data for each $DMU\ j(j = 1,\cdots,n)$, let x_{ij} denote its ith $(i = 1,\cdots,m)$ input and y_{rj} denote its rth $(r = 1,\cdots,s)$ output. Consequently, the VRS-SBM model can be used to evaluate the efficiency of certain $DMU\ k(k \in \{1,\cdots,n\})$.

$$
\begin{aligned}
&min\ \frac{1-\frac{1}{m}\sum\limits_{i=1}^{m}\frac{z_i^-}{x_{ik}}}{1+\frac{1}{s}\sum\limits_{r=1}^{s}\frac{z_r^+}{y_{rk}}}\\
&s.t.\quad x_{ik} = \sum_{j=1}^{n}x_{ij}\lambda_j + z_i^-,\quad i = 1,\cdots,m\\
&\qquad\ y_{rk} = \sum_{j=1}^{n}y_{rj}\lambda_j - z_r^+,\quad r = 1,\cdots,s\\
&\qquad\ \sum_{j=1}^{n}\lambda_j = 1,\ \lambda_j \geq 0,\ j = 1,\cdots,n\\
&\qquad\ z_r^+, z_i^- \geq 0,\ r = 1,\cdots,s,\ i = 1,\cdots,m
\end{aligned}
\tag{7}
$$

Fang et al. presented an equivalent slacks-based representation of Super-SBM, the VRS version of Super-SBM can be expressed as follows.

$$
\begin{aligned}
&min\ \frac{1+\frac{1}{m}\sum\limits_{i=1}^{m}\frac{w_i^-}{x_{ik}}}{1-\frac{1}{s}\sum\limits_{r=1}^{s}\frac{w_r^+}{y_{rk}}}\\
&s.t.\quad x_{ik} \geq \sum_{j=1,j\neq k}^{n}x_{ij}\lambda_j - w_i^-,\quad i = 1,\cdots,m\\
&\qquad\ y_{rk} \leq \sum_{j=1,j\neq k}^{n}y_{rj}\lambda_j + w_r^+,\quad r = 1,\cdots,s\\
&\qquad\ w_r^+ \leq y_{rk},\quad r = 1,\cdots,s\\
&\qquad\ \sum_{j=1,j\neq k}^{n}\lambda_j = 1,\ \lambda_j \geq 0,\ j = 1,\cdots,n,\ j \neq k\\
&\qquad\ w_r^+, w_i^- \geq 0,\ r = 1,\cdots,s,\ i = 1,\cdots,m
\end{aligned}
\tag{8}
$$

Through the specific analysis, the model (2) identifies the super-efficiency of DMU k by minimizing the input savings (w_i^-) and output surpluses (w_r^+).

Hence, the VRS version of the SBM-VRS model can be determined as follows.

$$
\begin{aligned}
&min\ \frac{1-\frac{1}{m}\sum\limits_{i=1}^{m}\frac{s_i^-}{x_{ik}}}{1+\frac{1}{s}\sum\limits_{r=1}^{s}\frac{s_r^+}{y_{rk}}}\\
&s.t.\quad x_{ik} = \sum_{j=1,j\neq k}^{n}x_{ij}\lambda_j - \bar{w}_i^{-*} + s_i^-,\quad i = 1,\cdots,m\\
&\qquad\ y_{rk} = \sum_{j=1,j\neq k}^{n}y_{rj}\lambda_j + \bar{w}_r^{+*} - s_r^+,\quad r = 1,\cdots,s\\
&\qquad\ \sum_{j=1,j\neq k}^{n}\lambda_j = 1,\ \lambda_j \geq 0,\ j = 1,\cdots,n,\ j \neq k\\
&\qquad\ s_r^+, s_i^- \geq 0,\ r = 1,\cdots,s,\ i = 1,\cdots,m
\end{aligned}
\tag{9}
$$

where $\bar{w}_i^{-*}$ and $\bar{w}_r^{+*}$ are optimal variables of problem in the equation and it is demonstrated as follows.

$$\begin{aligned}
& min \ \frac{1+\sum_{i=1}^{m}\frac{\mu_i w_i^-}{p_i^-}}{1-\sum_{r=1}^{s}\frac{v_r w_r^+}{p_r^+}} \\
& s.t. \quad x_{ik} \geq \sum_{j=1,j\neq k}^{n} x_{ij}\lambda_j - w_i^-, \quad i=1,\cdots,m \\
& \qquad y_{rk} \leq \sum_{j=1,j\neq k}^{n} y_{rj}\lambda_j + w_r^+ - s_r^+, \quad r=1,\cdots,s \\
& \qquad \sum_{j=1,j\neq k}^{n} \lambda_j = 1, \ \lambda_j \geq 0, \ j=1,\cdots,n, \ j\neq k \\
& \qquad w_r^+ \leq p_r^+, \ r=1,\cdots,s \\
& \qquad w_r^+, w_i^- \geq 0, \ r=1,\cdots,s, \ i=1,\cdots,m
\end{aligned} \tag{10}$$

where $p_i^- = \max_j\{x_{ij}\} - \min_j\{x_{ij}\}, p_r^+ = \min_j\{y_{rj}\} - \min_j\{y_{rj}\}$, μ_i and v_r are known positive weights satisfying $\sum_{i=1}^{m}\mu_i = 1$ and $\sum_{r=1}^{s} v_r = 1$. The variable μ_i and v_r make model (4) more generalized than the average weights $1/m$ and $1/s$ do respectively. For unexpected decision making units, variable Z is introduced into SBM model as unexpected inputs. Simultaneously, we conduct efficiency evaluation $Y2$ as expected output and B as unexpected output. Based on the above analysis, the SBM model considering unexpected input and output can be expressed as follows.

$$\begin{aligned}
& max \ \theta \\
& s.t. \begin{cases}
\theta + \omega X_0 + \xi Z_0 - \mu Y_0 - \delta B_0 = 1 \\
\omega x_j + \xi z_j - \mu y_j - \delta b_j \geq 0, \ j=1,2,\cdots,n \\
\omega \geq 1/m(1/X_0) \\
\mu \geq \theta/s(1/Y_0) \\
\xi \geq 1/t(1/Z_0) \\
\delta \geq \theta/p(1/B_0)
\end{cases}
\end{aligned} \tag{11}$$

If the attribute values of the observation sample point change to the benchmark point, it is possible to leave the original category. Research has found that frontier clustering algorithms can well fit and approximate classified hyper surfaces. At the same time, it is found that the clustering method using two relative frontier surfaces outperforms the linear mathematical programming clustering model in classification performance. When the relationship between attribute values and category values is non-monotonic, direct application of SBM clustering frontiers performs poorly. For the case where the sample data is linearly non-separable and it does not satisfy monotonicity, it is necessary to convert the original data into high-dimensional linearly data that can be clustered. Concurrently, the dual programming form of SBM model considering unexpected input and unexpected output can be determined as.

$$\begin{aligned}
& min \ \rho = \frac{1-\frac{1}{m+t}\left(\sum_{i=1}^{m}\frac{s_i^-}{x_{io}}+\sum_{i=1}^{t}\frac{s_i^-}{z_{io}}\right)}{1+\frac{1}{s+p}\left(\sum_{r=1}^{s}\frac{s_r^+}{y_{r0}}+\sum_{r=1}^{p}\frac{s_r^+}{b_{r0}}\right)} \\
& s.t. \quad \sum_{j=1}^{n}\lambda_j x_{mj} + s_i^- = X_0, \ m=1,2,\cdots,m \\
& \qquad \sum_{j=1}^{n}\lambda_j y_{rj} - s_r^+ = Y_0, \ r=1,2,\cdots,s \\
& \qquad \sum_{j=1}^{n}\lambda_j z_{sj} + s_t^- = Z_0, \ t=1,2,\cdots,t \\
& \qquad \sum_{j=1}^{n}\lambda_j b_{rj} - s_r^+ = B_0, \ r=1,2,\cdots,p \\
& \qquad \lambda \geq 0, s_i^-, s_r^+, s_t^-, s_p^+ \geq 0
\end{aligned} \tag{12}$$

When using the SBM model to measure efficiency, ineffective decision making units reach the frontier through the same proportion of indicators that change input or output. Therefore, using SBM model can greatly improve the effectiveness and efficiency of decision-making. Because the sensitivity analysis of DMUs is important worthy of discussion in SBM, the sensitivity is considered as the essential variable index in SBM considering unexpected output. At the same time, we propose to design clustering rules according to the measurement results of direction distance function. Therefore, the hybrid SBM model can be converted to the linear CCR model form and the linear model form can be expressed as follows.

$$\begin{aligned}
& min \ \tau = \varphi - \frac{1}{m+t}\left(\sum_{i=1}^{m}\frac{s_i^-}{x_{io}}+\sum_{i=1}^{t}\frac{s_i^-}{z_{io}}\right) \\
& s.t. \begin{cases}
\sum_{j=1}^{n}\lambda_j x_{mj} + s_i^- = X_0, \ m=1,2,\cdots,m \\
\sum_{j=1}^{n}\lambda_j y_{rj} - s_r^+ = Y_0, \ r=1,2,\cdots,s \\
\sum_{j=1}^{n}\lambda_j z_{sj} + s_t^- = Z_0, \ t=1,2,\cdots,t \\
\sum_{j=1}^{n}\lambda_j b_{rj} - s_p^+ = B_0, \ r=1,2,\cdots,p \\
1 = \varphi + \frac{1}{s+p}\left(\sum_{i=1}^{m}\frac{s_r^+}{y_{ro}}+\sum_{r=1}^{p}\frac{s_r^+}{b_{ro}}\right) \\
\lambda, s_i^-, s_r^+, s_t^-, s_r^+ \geq 0
\end{cases}
\end{aligned} \tag{13}$$

III. THE MEAN CLUSTERING ALGORITHM BASED ON TIME SERIES VARIABLES

With the development of data mining technology, all kinds of data information show blowout growth. Therefore, the continuous exploration and application of data value has led to the rise of a series of data clustering research hotspots.In the meantime, the time series clustering as one of the key issues has become an important direction of data driven in recent years. By processing, analyzing and modeling data, time series clustering can mine the internal structural characteristics of data. Look for the development law and change trend of things and provide a reasonable and effective theoretical basis for decision-making analysis and policy making in economic society.However, classical time series clustering models often have strict assumptions and high requirements for data quality. At the same time, the traditional clustering models have greater limitations in application.Complex data types and constantly developing application scenarios put forward higher requirements for the performance of clustering models. Therefore, it is of great significance to build clustering models suitable for different data types. Compared with the traditional time series clustering model, the fuzzy time series clustering model is mainly characterized by dealing with fuzzy sets. In the meantime, it has significant advantages in modeling fuzzy and uncertain data.

Extensive simulation studies including multivariate linear, nonlinear and GARCH processes show that the fuzzy time algorithm is effective [7]. A new fuzzy clustering algorithm for multivariate time series is proposed according to the difference of data in different degrees [8]. Concurrently, the simulation results of multi-objective clustering algorithm show that the time complexity and model complexity are reduced [9]. In addition, the fuzzy time series clustering algorithm does not require strict premise assumptions and a large number of training samples.Further more, this feature makes up for the prominent limitations of traditional time series clustering. According to different data types, fuzzy time series clustering models can be built based on different data fuzzification and fuzzy relationship building methods. Aiming at single dimension large sample data, multi-objective optimization and kernel fuzzy mean clustering algorithm are adopted to realize data fuzzification. Subsequently, a fast and effective fuzzy time series algorithm based on clustering is introduced to deal with the classification problem [10]. Through a series of experimental analysis, it is verified that the fuzzy time series clustering model has good effectiveness in a variety of applications.By processing the fuzziness,the algorithm can effectively improve the clustering performance by learning more information from the available information [11]. At the same time, a fuzzy clustering algorithm based on multivariable time series is proposed by estimating the parameters of the membership function. Finally, the experimental results show that the of multivariate time series based on Gaussian model is a promising clustering method. [12]. Through the system based on fuzzy sets, the membership value of the input set is obtained by using clustering algorithm in the structure of the fuzzy regression function method [13].Furthermore, the performance of the clustering model can be improved effectively by introducing data preprocessing and optimization algorithms into the fuzzy time series clustering algorithm.Simultaneously, the data decomposition and integration strategy is adopted to reduce the noise impact in the data and improve the classification ability of the model in the single dimension large sample data.In small sample data, the change rate of original data is calculated to eliminate the trend in complex data and improve the generalization and universality of the model. Concurrently, it can use feature extraction to remove redundant variable attributes from multidimensional variables and selects effective variables in multidimensional sample data as model input, so as to improve the training efficiency of clustering model.Therefore, for the different time series, we need to use the variety of complex data optimization techniques to further improve the performance and parameters of clustering algorithms.

The information entropy of a random variable can reflect the amount of information and uncertainty contained in the variable.Assuming that the attribute is a cluster variable, the information entropy $H(A)$ of the clustering algorithm is calculated generally as follows.

$$H(A) = -\sum_{i}^{d} p(a^{(i)}) log(p(a^{(i)})) \tag{14}$$

Suppose there are the following sets of independent variables $V = \{A_1, \cdots, A_m\}$. Consequently, the information entropy of the variable set can be expressed as follows.

$$H(A) = -\sum_{i}^{d} p(a^{(i)}) log(p(a^{(i)})) \tag{15}$$

The lower the value of $H(V)$, the lower the uncertainty of the variable set of V. Assume there is clustering of $P = \{C_1, \cdots, C_k\}$.As a result, the conditional entropy $H(V|P)$ based on the partition is the overall information entropy of the partition, which is calculated as follows.

$$\begin{aligned} E(P) &= H(V|P) \\ &= -\sum_{j=1}^{m}\sum_{l}^{k} p(C_l) \sum_{i=1}^{d_j} p(a_j^{(i)}|C_l) log(p(a_j^{(i)}|C_l)) \\ &= -\sum_{j=1}^{m}\sum_{l=1}^{k}\sum_{i=1}^{d_j} p(a_j^{(i)}, C_l) log(p(a_j^{(i)}|C_l)) \end{aligned} \tag{16}$$

where $p(a_j^{(i)}|C_l)$ is the conditional probability of the value $a_j^{(i)}$ in the C_l. $p(C_l) = |C_l|/n$.The information entropy index $E(P)$ represents the information entropy of the whole partition with the weighted sum of all kinds of entropy, thus reflecting the quality of clustering partition. The smaller the value of $E(P)$, the more concentrated the attribute values of each category in the partition which means that objects in the same category are more similar.

The time series clustering methods are widely used in various disciplines. Similar to the clustering algorithm, the clustering algorithm calculates the difference degree of each object and class center within a class, and reflects the overall quality of clustering division by the sum of various differences. For clustering variable data, the average value of each object value has no practical significance. Therefore, the clustering algorithm replaces the mean value with the mode of the value to determine the class center. At the same time, the difference distances between an object within a class and the class center is calculated as follows.

$$d(X_{li}, Z_l) = \sum_{j=1}^{m} \delta(x_{lij}, z_{lj}) \tag{17}$$

where X_{li} is the object of attribute characteristic C_l. x_{lij} is the value of attribute variable X_{li} on A_j attribute. Z represent environmental variable and z_{lj} is the variable in Z_l.The relationship expression is as follows.

$$\delta(x_{lij}, z_{lj}) = \begin{cases} 1, x_{lij} = z_{lj} \\ 0, x_{lij} \neq z_{lj} \end{cases} \tag{18}$$

Through the above analysis, the calculation formula of clustering algorithm indicators F can be described as follows.

$$F(P) = \sum_{l=1}^{k} d_{cluster}(C_l) = \sum_{l=1}^{k} \sum_{i=1}^{|C_l|} d(X_{li}, Z_l) \tag{19}$$

where $d_{cluster}(C_l)$ is the sum of the differences between the objects C_l in the class and the class centers. Therefore, the calculation formula can be expressed as follows.

$$\begin{aligned} d_{cluster}(C_l) &= \sum_{i=1}^{|C_l|} d(X_{li}, Z_l) \\ &= |C_l| \sum_{j=1}^{m} [1 - max_{i=1}^{d_j} p(a_j^{(i)} |C_l)] \end{aligned} \tag{20}$$

The lower the value of $F(P)$ is, the more similar the objects in each category of the clustering algorithm are to the class centers and the higher the quality of clustering division can be considered. The cluster utility index CU can be used to calculate the probability of obtaining the same attribute value for each object in the same class. The higher the value of $CU(P)$, the higher the quality of clustering is. Hence, the value of the clustering utility index CU of the clustering algorithm can be calculated as follows.

$$\begin{aligned} CU(P) &= \sum_{l=1}^{k} p(C_l) \sum_{j=1}^{m} \sum_{i=1}^{d_j} [p(a_j^{(i)} |C_l)^2 - p(a_j^{(i)})^2] \\ &= \sum_{l=1}^{k} p(C_l) \sum_{j=1}^{m} \sum_{i=1}^{d_j} [p(a_j^{(i)} |C_l)^2] \\ &\quad - \sum_{l=1}^{k} \sum_{j=1}^{m} \sum_{i=1}^{d_j} p(a_j^{(i)})^2 \end{aligned} \tag{21}$$

The cluster utility index $p(a_j^{(i)} |C_l)^2$ can reflect the probability that objects in the same class have the same value. At the same time, the probability $p(a_j^{(i)})^2$ reflecting the same value of objects between different classes shall be applied. However, for a given data set, the value of $p(a_j^{(i)})^2$ is a constant, which can be determined quantitatively and scientifically according to the following methods. Consequently, the specific calculation process can be described as follows.

$$CU(P) = \sum_{l=1}^{k} p(C_l) \sum_{j=1}^{m} \sum_{i=1}^{d_j} [p(a_j^{(i)} |C_l)^2] - Constant \tag{22}$$

The above formula shows that the clustering utility index CU can reflect the similarity of objects within a class, but cannot reflect the difference of objects between classes in the process of clustering algorithm. In order to improve the index CU of clustering algorithm, we will further average the value $CU(P)/k$ of the number of classes, that is, to compare the clustering division of different number of $CU(P)$.

IV. THE FUZZY CLUSTERING ALGORITHM BASED ON VARIABLE WEIGHTING

In recent years, the clustering analysis of complex data has aroused our extensive attention. The fuzzy clustering model with time series inherit the benefits of fuzzy theory in the clustering framework is characterized in the following way. At the same time, the following conditions exist for the dataset $X = \{x_1, x_2, \cdots, x_n\}$. Assume there is an attribute weighted vector $w = [w_1, w_2, \cdots, w_s]^T$ and $\forall j : w_j > 0$. Through the above analysis, the fuzzy clustering model based on variable weight can be given as follows.

$$\begin{cases} min\ J_2(U, V, w) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \|x_k - v_i\|_w^2 \\ s.t. \sum_{i=1}^{c} u_{ik} = 1,\ k = 1, 2, \cdots, n \\ \sum_{j=1}^{s} w_j = 1 \end{cases} \tag{23}$$

We applies fuzzy time series clustering models under different data types to different aspects of social and economic life. Fuzzy time series distance models suitable for different data characteristics are constructed through different data fuzzification methods and the application range of fuzzy time series clustering models is broadened. Concurrently, it can be seen from the analysis that the attribute weighting vector w appears as a variable in the clustering objective function J_2. Therefore, we can adopt the Lagrangian multiplier method and assume that the augmented function can be expressed as follows.

$$\begin{aligned} &J_{2\lambda\eta}(U, V, w) \\ &\quad = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \|x_k - v_i\|_w^2 + \sum_{k=1}^{n} \lambda_k (\sum_{i=1}^{c} u_{ik} - 1) \\ &\quad\quad + \eta (\sum_{j=1}^{s} w_j - 1) \end{aligned} \tag{24}$$

where $\lambda = [\lambda_1, \lambda_2, \cdots, \lambda_n]^T$ and η is Lagrange multiplier.Then, under the two equality constraints of membership degree u_{ik} and attribute weight w_j, the necessary conditions for minimizing the objective function can be developed as follows.

$$\begin{cases} v_i = \sum_{k=1}^{n} u_{ik}^m x_k \Big/ \sum_{k=1}^{n} u_{ik}^m, i = 1, 2, \cdots, c \\ u_{ik} = \left[\sum_{i=1}^{c} \left(\|x_k - v_i\|_w^2 / \|x_k - v_l\|_w^2 \right)^{\frac{1}{m-1}} \right]^{-1} \\ w_j = \left[\sum_{i=1}^{c} \left(\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m (x_{jk} - v_{ji})^2 \Big/ \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m (x_{ik} - v_{ti})^2 \right) \right]^{-1} \\ i = 1, 2, \cdots, c; \ j = 1, 2, \cdots, s; \ k = 1, 2, \cdots, n \end{cases} \tag{25}$$

In the weighted fuzzy clustering algorithm, attribute weighting methods can have different clusters from different angles. For example, attribute weights are determined according to decision-makers' preferences, experiences and other factors. Simultaneously, it is still necessary to determine the attribute weighting through subjective weighting of the mathematical model according to a specific index and actual data. When clustering with attribute weighting method, we can set the attribute weight determined in advance as a constant value.

Assume that the multidimensional dataset can be represented as $X = \{x_1, x_2, \cdots, x_n\}$. At the same time, we give the vector characteristics of attribute weights as $w = [w_1, w_2, \cdots, w_s]^T$. $\forall j : w_j > 0$ and $\sum_{j=1}^{s} w_j = 1$. Then the model of clustering algorithm based on constant weight can be designed as follows.

$$\begin{cases} \min J_1(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \|x_k - v_i\|_w^2 \\ s.t. \sum_{i=1}^{c} u_{ik} = 1, \ k = 1, 2, \cdots, n \end{cases} \tag{26}$$

where variable $\|\cdot\|_w$ represents weighted Euclidean distance and it can be defined as follows.

$$\|x_k - v_i\|_w = [(x_k - v_i)^T W^T W (x_k - v_i)]^{1/2} \tag{27}$$

where it can be defined as $W = diag[w_1, w_2, \cdots, w_s]$.

By weighting samples and features, the fuzzy algorithm and sample weighted clustering algorithm are unified into a common framework. At the same time, the new fuzzy temporal clustering algorithm is quite effective for processing data sets with different distributions and characteristics.Through the above exploration, we can adopt the Lagrangian multiplier method and assume that the augmented function can be expressed as follows.

$$\begin{aligned} & J_{1\lambda}(U, V) \\ & = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \|x_k - v_i\|_w^2 + \sum_{k=1}^{n} \lambda_k (\sum_{i=1}^{c} u_{ik} - 1) \end{aligned} \tag{28}$$

where $\lambda = [\lambda_1, \lambda_2, \cdots, \lambda_n]^T$ and it is the Lagrange multiplier. Then, under the equality constraints u_{ik} on membership, the necessary conditions for minimizing the objective function J_1 of the hybrid clustering algorithm are shown as follows.

$$\begin{cases} v_i = \sum_{k=1}^{n} u_{ik}^m x_k \Big/ \sum_{k=1}^{n} u_{ik}^m, i = 1, 2, \cdots, c \\ u_{ik} = \left[\sum_{i=1}^{c} \left(\|x_k - v_i\|_w^2 / \|x_k - v_l\|_w^2 \right)^{\frac{1}{m-1}} \right]^{-1} \\ i = 1, 2, \cdots, c; \ k = 1, 2, \cdots, n \end{cases} \tag{29}$$

When using fuzzy clustering algorithm to cluster data sets, it is necessary to determine the attribute weight according to some attribute weight in the initialization process. In the meantime, the latter process is similar to the standard clustering algorithm. Through the above analysis, the hybrid algorithm flow is shown as follows.

Step 1: use some attribute weighting method to determine the attribute weight vector and take account of the contribution of different feature for cluster analysis.

Step 2: To set the fuzzy parameters, clustering number and threshold, and initialize the partition matrix randomly.

Step 3: Update the prototype like matrix of fuzzy clustering algorithm in proper sequence.

Step 4: Update the partition fuzzy matrix of fuzzy clustering algorithm in the same measure.

Step 5: If the following conditions are true $\forall i, k : max \left| u_{ik}^{(l)} - u_{ik}^{(l-1)} \right| < \varepsilon$. Then the clustering algorithm stops. Otherwise, return to step 3.

V. A NOVEL HYBRID SBM MODEL BASED ON FUZZY TIME SERIES

Time series data is a collection of observations in chronological order. It is an important data object in series data and widely exists in daily life and scientific research fields. The characteristics of time series data include large amount of data, high dimension and continuous updating. In addition, the key information in the time series usually exists in the overall change rather than a specific value. Furthermore, the increasing use of time series data has triggered a lot of research in the hot field of data clustering algorithms. At the same time, the segmentation production function derived from data envelopment analysis is also used for clustering. In the end the clustering method based on data envelopment analysis is verified to be effective by an example [14]. According to the characteristics of data, a new clustering algorithm based on the results of data envelopment analysis is also proposed [15]. An improved fuzzy clustering method based on data envelopment analysis is proposed for sparse input and output data. Consequently,the supplement sparse data in a reliable and accurate way to scientifically evaluate the relative technical efficiency of the production unit [16]. However, due to the complexity of time series data, how to accurately and efficiently find these fragments for accurateclustering. How to use the dependency between attributes and establish an effective clustering method based on nonparametric frontier for time series. Subsequently, how to implement an effective multivariate time series clustering algorithm has become a special challenge.

As the composition of time series data presents a trend of complexity and high dimension, it is very meaningful to build a nonparametric clustering algorithm based on time series under the framework of multiple time periods and dimensions. Analyze the heterogeneity of multi period and multi dimension DMUs, construct a multi period nonparametric common frontier estimator, and apply this method to study the impact of multi period and multi dimension on the inter group efficiency, intra group efficiency, intra group and inter group technical barriers of DMUs.Starting from the construction of empirical production function, this paper theoretically explores the relationship between the efficiency values given by different time series SBM models. At the same time, the methods for measuring the efficiency of time series SBM decision-making units will be different for different situations. Therefore, the foundation of DEA is to use the indicator data of the decision-making unit to simulate the empirical production function, and the production function itself depicts the relationship between various production factors and the maximum output that can be produced when the technical level is unchanged.

According to the effective definition of TS-SBM, which measures the efficiency of decision making units based on time series data, the following SBM model for measuring the efficiency of time series decision making units can be given. Therefore,the specific design of the model is shown as follows.

$$\begin{cases} min\ \theta_t - \varepsilon(\hat{e}^T s_t^- + e^T s_t^+) \\ s.t.\ \sum_{j=1}^{T} x_j^0 \lambda_j^t + s_t^- = \theta_t x_t^0 \\ \sum_{j=1}^{T} y_j^0 \lambda_j^t - s_t^+ = y_t^0 \\ \delta_1 (\sum_{j=1}^{T} \lambda_j^t - \delta_2 (-1)^{\delta_3} \lambda_{T+1}) = \delta_1 \\ s_t^-, s_t^+ \geq 0,\ \lambda_j^t \geq 0,\ j = 1, 2, \cdots, T+1. \end{cases} \tag{30}$$

When taking the different values, the (TS-SBM) model represents the time series mixed model of different returns to scale respectively.

(1) When $\delta_1 = 0$, (TS-SBM) was a time series SBM model that satisfied the constant returns to scale;

(2)When, $\delta_1 = 1$ and $\delta_2 = 0$, (TS-SBM) is a time series SBM model that satisfies variable returns to scale;

(3) When, $\delta_1 = 1$ and $\delta_2 = 1$ and $\delta_3 = 1$, (TS-SBM) is a time series SBM model that satisfies the non increasing returns to scale.

(4) When, $\delta_1 = 1$ and $\delta_2 = 1$ and $\delta_3 = 0$, (TS-SBM) is a time series SBM model that satisfies the non decreasing returns to scale.

By combining convexity hypothesis and different returns to scale hypothesis, cross period common frontier efficiency measures are constructed under single/multi period and multi-dimensional frameworks respectively, and then corresponding single/multi period and multi-dimensional decision making unit efficiency is constructed.By investigating the characteristics of the change of the common frontier efficiency of different types of DMUs, a dynamic clustering method based on fuzzy time series under unbalanced data is designed.If the input is smaller when the output is unchanged, the production of the DMU at any time is considered to be effective.

To ensure the TS-SBM clustering results for each dataset, we employ the clustering algorithms based on the different nonparametric frontier. All kinds of clustering algorithms considering for clustering time series were originally developed to be applied to static data rather than uncertain and fuzzy time series due to more and more application scenarios.Data presents characteristics such as high dimensionality, complex and diverse intra-class variability, etc. Therefore, it is necessary to study hybrid high-dimensional time series clustering methods. However, the hybrid SBM clustering algorithm can overcome the above limitations. Based on the TS-SBM, the multivariate fuzzy time series clustering method based on Slacks Based Measure (MFTS-SBM) can be expressed as follows.

$$\begin{cases} min\ \theta_t - \varepsilon(\hat{e}^T s_t^- + e^T s_t^+) \\ s.t.\ \sum_{j=1}^{T} x_j^0 \lambda_j^t + s_t^- = \theta_t x_t^0 \\ \sum_{j=1}^{T} y_j^0 \lambda_j^t - s_t^+ = y_t^0 \\ \delta_1 (\sum_{j=1}^{T} \lambda_j^t - \delta_2 (-1)^{\delta_3} \lambda_{T+1}) = \delta_1 \\ s_t^-, s_t^+ \geq 0,\ \lambda_j^t \geq 0,\ j = 1, 2, \cdots, T+1. \end{cases} \tag{31}$$

From the perspective of the relative efficiency of DMUs, the efficiency values given by the time series SBM model fully conform to the innovative ideas of the SBM method. For a group of time series decision making units, more data information should be obtained as far as possible when measuring the SBM efficiency, so as to measure the efficiency value of time series decision making units more accurately. The less incomplete the data is, the more accurate the measurement results will be. Therefore, we should give priority to the application of maximum production frontier to measure the efficiency of time series decision-making units.

Because of the imbalance of classification, the traditional classification algorithms often can not achieve the desired results. In practical applications, error clustering may sometimes lead to serious consequences. Therefore, it is of great practical significance to study the clustering problem of unbalanced data. Since the SBM method evaluates the relative efficiency, the efficiency value of the decision making unit of

SBM is a relative value no matter whether the time series data or the section data at time T are selected.When measuring their efficiency of a group of time series decision making unit, we should try to obtain more data information, so as to measure the efficiency value of time series decision making units more accurately.When measuring the efficiency and technological progress of a group of time series data, the precision of model calculation can be greatly improved if another group of cross-sectional data can be obtained. Through the scientific analysis, it is feasible, scientific and reasonable to use the hybrid SBM method to calculate the efficiency of time series decision-making units when the production system maintains irreversible technological progress.

VI. EXAMPLE ANALYSIS

Through the above research, clustering algorithm design and its effectiveness evaluation are the key issues in clustering analysis. At the same time, its effectiveness index can measure the best effectiveness of different cluster settlement methods. The experiment proves that the validity index has superior ability in determining reasonable clustering algorithm [17]. Clustering technology is also an important analysis tool and method for data processing in data mining. For example: decision tree, Bayes clustering, k-nearest neighbor(K-NN), neural network (NN), genetic algorithm (GA), Particle Swarm Optimization(PSO), support vector machine (SVM) and other classical algorithms. With the further expansion of scientific applications, the need to do clustering processing of complex data has become complex and diverse. We also know that the internal evaluation of clustering effectiveness is the key link of clustering analysis. Therefore, it is necessary to find appropriate internal evaluation indicators for analysis. However, due to the lack of inherent geometric characteristics of the clustering data, the corresponding clustering algorithm design is quite different from the data.

In this paper, some problems in clustering analysis of clustering data are studied, including the initial cluster center selection algorithm, similarity measurement and clustering algorithm for high-dimensional, massive and temporal clustering data. The clustering simulation results of different test samples from the public total data set are shown in TABLE 1 as follows.

TABLE 1. Clustering accuracy of different clusterers on test samples.

	UCI	Negative class	Positive class	Accuracy-difference
SBM	45.35	68	0	68.00
TS-SBM	35.35	0	100	-100.00
FTS-SBM	60.00	85	10	85.00
NN	62.38	90	20	86.00
GA	62.30	95	15	89.00
PSO	61.25	98	12	87.00
SBM-DA	68.65	100	0	100.00
Two stage MSD-DA	35.35	0	100	-100.00
Linear DA	68.65	100	0	100.00
Secondary DA	69.60	85	30	55.00
Logical-regression-clustering	70.00	100	10	90.00
Decision tree	58.65	70	30	45.00
Gaussian kernel-SVM	75.35	70	80	-10.00
K-NN	58.65	70	30	40.00
TS	34.50	60	40	50.00
FTS	50.50	80	30	60.00
DEA	59.65	40	90	-50.00
DEA-DA	75.35	80	70	-10.00

Through the experimental results of the public aggregate data set, we found that the clustering accuracy of non convex clustering front is higher than that of convex clustering front, no matter whether we use positive or negative class data to build clustering front. Although as far as the overall clustering accuracy is concerned, the clustering frontier based on positive class construction is slightly better than that based on negative class construction. Due to the data imbalance in this dataset, the number of positive samples is 3.20 times that of negative samples, so it is necessary to further compare the clustering accuracy under different groups. In addition, we found that the data imbalance leads to a large difference in the clustering accuracy between positive and negative classes.Simultaneously, when K nearest neighbor method (K-NN) is used, the difference between positive and negative accuracy reaches 46.00%.

At the same time, we found that in the proposed non-parametric clustering frontier method, when the clustering frontier is constructed based on negative classes, even for a few classes, it is 100.00% correct clustering through the comparative experimental results of the samples. Moreover, we found that the difference between the clustering accuracy of positive and negative classes was not large, and the difference under the nonconvex frontier would also be further reduced. For the nonparametric clustering frontier method based on positive data, the non convex frontier method also performs well in correctly clustering positive and negative classes. However, the frontier of positive convex clustering is to cluster a few classes. Finally, we found deeply that all the observation points to be clustered in the clustering front belong to the same category as the sample set training the clustering front, while all the observation points to be clustered outside the clustering front belong to a relative other category.

The monotonic relationship between attribute values of data sets and clustering results has been considered more in existing models and the characterization of attributes with non monotonic relationship with clustering results in clustering based on nonparametric frontier has not received enough attention. Therefore, we will compare the clustering performance of nonparametric frontier based methods under different sample sizes. Through the simulation experiments, the clustering performance of the method based on nonparametric frontier below when the sample size 500 is shown in TABLE 2 as follows.

When the sample size is 800, the clustering performance of the method based on nonparametric frontier is shown in TABLE 3.

Then increase the sample size. The specific performance of clustering based on nonparametric frontier method when the sample size is 1500 is shown in TABLE 4.

TABLE 2. Experimental results of nonparametric frontier method with 500 samples.

	Accuracy	precision	Recall	Specificity	$F_{0.5}$	G-Means
Single -SBM	0.7850	0.9750	0.5895	0.9820	0.8645	0.7535
Single-FDH	0.7880	0.9245	0.5895	0.9750	0.8565	0.7520
Relative-SBM	0.8580	0.8505	0.8590	0.8470	0.8445	0.8475
Relative-FDH	0.8525	0.8290	0.9250	0.8200	0.8500	0.8505
SBM	0.6925	0.6920	0.6800	0.6870	0.6015	0.6025
TS-SBM	0.7550	0.8530	0.7570	0.9050	0.8440	0.8040
FTS-SBM	0.6580	0.7864	0.4070	0.8890	0.6560	0.6105
MIP SBM -DA	0.6380	0.7850	0.4070	0.8890	0.6565	0.6205
MSD-DA	0.7950	0.9550	0.6590	0.9500	0.8520	0.7865
Two stage MSD-DA	0.8245	0.8825	0.7560	0.9035	0.8558	0.8195
Linear DA	0.8530	0.8350	0.8410	0.8450	0.8435	0.8425
Secondary DA	0.8220	0.8500	0.8050	0.8500	0.8315	0.8320
Logical-regression-clustering	0.8455	0.8510	0.8470	0.8545	0.8485	0.8460
Decision tree	0.8595	0.8590	0.8405	0.8590	0.8595	0.8580
Gaussian kernel-SVM	0.8560	0.8860	0.8860	0.8860	0.8860	0.8865
K-NN	0.8550	0.8520	0.8420	0.84354	0.8435	0.8440
TS	0.8595	0.8580	0.8580	0.8510	0.8510	0.8445
FTS	0.8860	0.8590	0.8595	0.8595	0.8590	0.8445
DEA	0.8540	0.8565	0.8560	0.8560	0.8560	0.8240
DEA-DA	0.8540	0.8520	0.8525	0.8525	0.8520	0.8540
NN	0.8661	0.8664	0.8660	0.8560	0.8565	0.8560
GA	0.8340	0.8240	0.8145	0.8345	0.8240	0.8245
PSO	0.8640	0.8642	0.8645	0.8745	0.8745	0.8700

TABLE 3. Experimental results of nonparametric frontier method with 800 samples.

	Accuracy	precision	Recall	Specificity	$F_{0.5}$	G-Means
Single -SBM	0.8050	0.9850	0.6090	0.9920	0.8840	0.8035
Single-FDH	0.7980	0.9340	0.6095	0.9850	0.8860	0.7825
Relative-SBM	0.8680	0.8600	0.8690	0.8572	0.8545	0.8585
Relative-FDH	0.8635	0.8392	0.9255	0.8305	0.8600	0.8605
SBM	0.7025	0.7025	0.7000	0.7070	0.6815	0.6825
TS-SBM	0.7650	0.8635	0.7670	0.9155	0.8540	0.8240
FTS-SBM	0.6680	0.7965	0.5270	0.8990	0.6760	0.6308
MIP SBM -DA	0.6580	0.8050	0.4270	0.8990	0.6665	0.6405
MSD-DA	0.8050	0.9650	0.6790	0.9605	0.8620	0.7965
Two stage MSD-DA	0.8345	0.8922	0.7660	0.9235	0.8858	0.8590
Linear DA	0.8630	0.8550	0.8615	0.8650	0.8535	0.8525
Secondary DA	0.8424	0.8705	0.8150	0.8600	0.8510	0.8464
Logical-regression-clustering	0.8550	0.8610	0.8580	0.8645	0.8585	0.8567
Decision tree	0.8695	0.8697	0.8505	0.8690	0.8695	0.8780
Gaussian kernel-SVM	0.8660	0.8960	0.8960	0.8963	0.8961	0.8985
K-NN	0.8850	0.8720	0.8520	0.86354	0.8735	0.8745
TS	0.8895	0.8880	0.8780	0.8618	0.8715	0.8735
FTS	0.8965	0.8790	0.8690	0.8694	0.8690	0.8645
DEA	0.8740	0.8765	0.8765	0.8760	0.8860	0.8842
DEA-DA	0.8745	0.8720	0.8820	0.8725	0.8625	0.8740
NN	0.8861	0.8865	0.8867	0.8668	0.8805	0.8860
GA	0.8540	0.8540	0.8445	0.8647	0.8542	0.8540
PSO	0.8740	0.8745	0.8745	0.8845	0.8945	0.8905

TABLE 4. Experimental results of nonparametric frontier method with 1500 samples manuscript.

	Accuracy	precision	Recall	Specificity	$F_{0.5}$	G-Means
Single -SBM	0.7950	0.8850	0.6000	0.9420	0.8440	0.7830
Single-FDH	0.7940	0.8340	0.5950	0.9350	0.8260	0.7525
Relative-SBM	0.8436	0.8400	0.8455	0.8370	0.8240	0.8085
Relative-FDH	0.8335	0.8092	0.9055	0.8005	0.8500	0.8400
SBM	0.6625	0.6825	0.6800	0.6770	0.6515	0.6520
TS-SBM	0.7550	0.8535	0.7475	0.8655	0.8440	0.8140
FTS-SBM	0.6585	0.7865	0.5070	0.8690	0.6565	0.6208
MIP SBM -DA	0.6480	0.7850	0.4070	0.8290	0.6065	0.6200
MSD-DA	0.7850	0.9050	0.6090	0.9405	0.8420	0.7865
Two stage MSD-DA	0.8045	0.8822	0.7560	0.9035	0.8458	0.8295
Linear DA	0.8530	0.8050	0.8015	0.8050	0.8435	0.8225
Secondary DA	0.8324	0.8405	0.8050	0.8505	0.8310	0.8265
Logical-regression-clustering	0.8050	0.8010	0.8080	0.8445	0.8285	0.8360
Decision tree	0.8295	0.8297	0.8205	0.8290	0.8290	0.8380
Gaussian kernel-SVM	0.8465	0.8760	0.8460	0.8363	0.8461	0.8580
K-NN	0.8650	0.8420	0.8120	0.84354	0.8535	0.8445
TS	0.8595	0.8580	0.8080	0.8018	0.8015	0.8035
FTS	0.8465	0.8490	0.8490	0.8494	0.8490	0.8340
DEA	0.8140	0.8060	0.8065	0.8060	0.8060	0.8042
DEA-DA	0.8345	0.8320	0.8320	0.8320	0.8125	0.8240
NN	0.8561	0.8465	0.8260	0.8348	0.8205	0.8165
GA	0.8040	0.8045	0.8045	0.8047	0.8042	0.8040
PSO	0.8140	0.8245	0.8345	0.8345	0.8545	0.8405

Through the above experiments, we can draw the following conclusions. First of all, we found that the precision and specificity of the clustering model based on a single SBM frontier are as high as 92.10% and 95.00%. It is 9.50% and 18.00% higher than the results of the SBM frontier respectively.With the current sample size of 1500, the MSDDA model performs best in clustering performance, with accuracy and specificity of 95.50% and 96.50%. Through the experiments, we found that the results are higher than that of single SBM frontier clustering model with the best performance among frontier clustering models. At the same time, we also found the following two important laws through data analysis.Concurrently,we found the convexity assumption can improve the clustering performance. Compared with the clustering results of two relative SBM fronts, the clustering model based on the two FDH fronts performs better on each test index after loosening the convex hypothesis.Similarly,we found interestingly that the overall performance of the clustering model based on two relative fronts is significantly better than that based on a single front.

Specifically,the corresponding relative front clustering model has improved the accuracy by 6.50% and 8.25% compared with a single SBM front and a single FDH front respectively. Further more the G-Means index increased by 10.50% and 12.80% respectively. In addition, the difference between recall and specificity of the cluster model based on a single frontier is as high as 40.50% and 40.60%. However, the difference in recall and specificity of the cluster model based on relative frontier is only 2.40% and 12.50%. Compared with the former, the degree of difference is significantly reduced. Compared with the SBM related clustering model simultaneously, the frontier clustering model shows better accuracy in the case of small samples. With the increase of the sample size, the accuracy of the SBM clustering model and the intelligent clustering algorithm is slightly higher than that of the frontier.

Whether the structured data or unstructured data, the class imbalance inevitably exists. It has brought enormous difficulties and challenges to data clustering. Currently, a large amount of work has been carried out on the clustering of class unbalanced data and achieved good results. First, the unbalanced data set is clustered, the data is initially divided and then the fuzzy time series clustering is carried out to test the validity of the model. Finally,the experiments show that the hybrid method is practical.In general, when the clustering accuracy is high, the corresponding recall rate is poor. Where, the F index is used to balance the performance of the model on clustering accuracy and recall. However, compared with the recall rate, the decision-makers prefer the model to have higher accuracy.The clustering performance of the method based on nonparametric frontier in the unbalanced data set (100,300) is shown in TABLE 5.

TABLE 5. Clustering results based on nonparametric frontier data set (100,300).

	Accuracy	precision	Recall	Specificity	$F_{0.5}$	G-Means
Single - SBM	0.7050	0.8500	0.5295	0.9020	0.8045	0.7035
Single-FDH	0.7040	0.8245	0.5095	0.9010	0.8010	0.7020
Relative-SBM	0.8080	0.8005	0.8090	0.8070	0.8045	0.8075
Relative-FDH	0.8025	0.8090	0.9050	0.8000	0.8000	0.8005
SBM	0.6225	0.6225	0.6100	0.6270	0.5815	0.5825
TS-SBM	0.7050	0.8030	0.7070	0.8050	0.8040	0.7540
FTS-SBM	0.6080	0.7064	0.3070	0.8090	0.6060	0.6005
MIP SBM -DA	0.6080	0.7050	0.4050	0.8090	0.6065	0.6005
MSD-DA	0.7450	0.9150	0.6190	0.9100	0.8120	0.7460
Two stage MSD-DA	0.8145	0.8125	0.7060	0.8035	0.8058	0.8095
Linear DA	0.8130	0.8050	0.8110	0.8150	0.8135	0.8120
Secondary DA	0.8020	0.8300	0.7850	0.8100	0.8015	0.8020
Logical-regression-clustering	0.8055	0.8110	0.8070	0.8145	0.8085	0.8060
Decision tree	0.8195	0.8190	0.8205	0.8190	0.8195	0.8185
Gaussian kernel-SVM	0.8060	0.8360	0.8362	0.8560	0.8560	0.8565
K-NN	0.8050	0.8020	0.8020	0.84154	0.8235	0.8240
TS	0.8095	0.8080	0.8080	0.8010	0.8015	0.8045
FTS	0.8260	0.8290	0.8295	0.8295	0.8290	0.8225
DEA	0.8040	0.8065	0.8060	0.8064	0.8060	0.8040
DEA-DA	0.8340	0.8320	0.8325	0.8325	0.8320	0.8340
NN	0.8461	0.8464	0.8460	0.8460	0.8465	0.8460
GA	0.8240	0.8140	0.8245	0.8245	0.8145	0.8145
PSO	0.8540	0.8542	0.8545	0.8540	0.8546	0.8500

The above validation data results show that the non convex data frontier clustering method is superior to the convex data frontier clustering method in clustering accuracy and effectiveness. However, it is the less affected by the unbalanced data volume. At the same time, compared with the existing classical clustering models, the data frontier clustering model also shows higher clustering accuracy. The clustering performances of the method based on nonparametric frontier in the unbalanced data set (300,500) are provided in TABLE 6.

TABLE 6. Clustering results based on nonparametric frontier data set (300,500).

	Accuracy	precision	Recall	Specificity	$F_{0.5}$	G-Means
Single - SBM	0.7160	0.8600	0.5495	0.9120	0.8145	0.7235
Single-FDH	0.7240	0.8445	0.5290	0.9110	0.8210	0.7220
Relative-SBM	0.8280	0.8205	0.8290	0.8270	0.8145	0.8175
Relative-FDH	0.8125	0.8190	0.9150	0.8230	0.8100	0.8205
SBM	0.6325	0.6425	0.6200	0.6450	0.6015	0.5920
TS-SBM	0.7150	0.8230	0.7270	0.8158	0.8140	0.7664
FTS-SBM	0.6180	0.7106	0.3570	0.8190	0.6260	0.6105
MIP SBM -DA	0.6280	0.7250	0.4150	0.8190	0.6265	0.6205
MSD-DA	0.7550	0.9255	0.6390	0.9200	0.8220	0.7660
Two stage MSD-DA	0.8345	0.8225	0.7260	0.8135	0.8258	0.8190
Linear DA	0.8180	0.8080	0.8210	0.8250	0.8235	0.8220
Secondary DA	0.8120	0.8400	0.8050	0.8500	0.8515	0.8420
Logical-regression-clustering	0.8255	0.8210	0.8270	0.8245	0.8285	0.8260
Decision tree	0.8295	0.8290	0.8405	0.8490	0.8305	0.8285
Gaussian kernel-SVM	0.8260	0.8560	0.8562	0.8660	0.8667	0.8765
K-NN	0.8150	0.8220	0.8220	0.8615	0.8435	0.8440
TS	0.8195	0.8180	0.8180	0.8110	0.8115	0.8140
FTS	0.8260	0.8245	0.8230	0.8225	0.8236	0.8200
DEA	0.8080	0.8015	0.8050	0.8034	0.7860	0.7940
DEA-DA	0.8240	0.8220	0.8225	0.8225	0.8020	0.8140
NN	0.8361	0.8365	0.8360	0.8360	0.8365	0.8260
GA	0.8140	0.8040	0.8045	0.8055	0.8045	0.8040
PSO	0.8440	0.8440	0.8445	0.8448	0.8355	0.8350

Under the unbalanced data set (300, 500), the model based on the FDH frontier shows smaller inter group differences in the accuracy of different groups.By relaxing the convex hypothesis, $F_{0.5}$ and G-Means are increased by 0.86% and 4.10% respectively. In addition, the convex hypothesis is relaxed to improve the $F_{0.5}$ indexfor a single frontier model, the G-Means value decreases slightly.On the other hand, the difference between the recall rate and specificity of the cluster model based on a single data envelope frontier and the cluster model based on the relative data envelope frontier is 22.50% and 12.60%.However, the difference between the recall rate and specificity of the cluster model based on a single FDH front and the relative FDH front is only 15.60% and 7.80%, which is significantly smaller than the result of the data envelope front. Specifically, the corresponding relative frontier clustering model has increased 4.50% and 4.50% on the $F_{0.5}$ index and 5.06% and 7.50% on the G-Means index respectively.

The clustering technology based on data envelopment analysis for heterogeneous data sets and standard clustering technology were proposed to find the number of clusters [18]. Through the simulation experiment, it is found that when the

CV value of the data class size distribution in the original dataset is greater than 0.90, the clustering tends to reduce the difference of the data class size distribution, while it increases the difference when the difference of the data class size distribution is small. In addition, through the analysis of the intra class compactness measure used by each index, it is proved that the intra class compactness measure without separation measure is monotonically related to the number of classes when evaluating hierarchical results.

By analyzing the evaluation ability of existing internal indicators with this method, it is found that the lack number of real classes will seriously affect the evaluation ability of indicator efficiency.Furthermore, there are also some problems such as being highly sensitive to the evaluation parameters and vulnerable to the impact of data attribute characteristics. In the meantime, the semi-supervised learning method based on small-scale data sets has the characteristics of fast training speed and wide application range. In order to prevent the loss of useful large loss samples during the experiment, we adopted supervised learning to obtain the advanced performance of the clustering algorithm. In addition, the following is a clustering simulation experiment based on the above sample data.Consequently,the simulation results of different samples according to the strong clustering and weak clustering are shown in FIGURE.1-FIGURE.6 as follows.

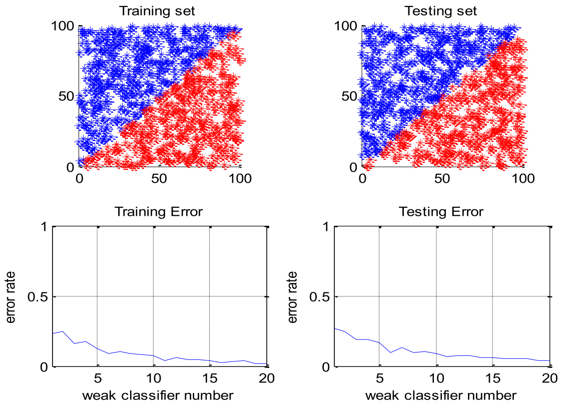

FIGURE 1. The results of samples 500 according to weak clustering.

As to the stability of resultant clusters derived from the hybrid SBM clustering method based on time series considering the nonparametric frontier, it is determined whether or not it is robust to a slight change in the input–output data and retains the existing reference set of frontiers.

Under the current unbalanced data set (100, 300), the clustering model based on the relative FDH frontier performs best overall. Although there is a serious imbalance in the amount of data in the current sample dataset, the clustering model based on a single SBM frontier has reached 97.21% and 90% in accuracy and specificity. Subsequently, the results are 2.19% and 13% higher than those of the frontier of the relative data envelope, respectively.Under the current unbalanced data set (300, 500), the clustering model based on a single data envelope frontier can perform best in the cluster group.Furthermore,the clustering model based on a single SBM front and a relative data envelope front has 13.50% and 17.10% differences in recall and specificity,which is

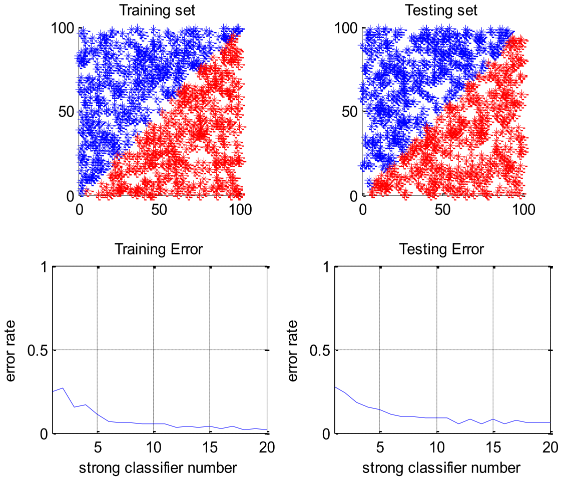

FIGURE 2. The results of samples 500 according to strong clustering.

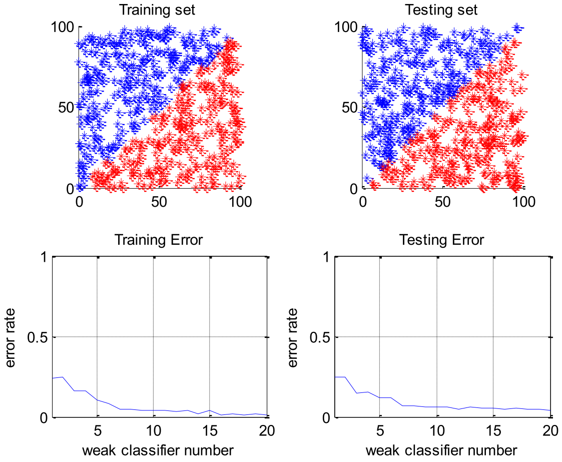

FIGURE 3. The results of samples 300 according to weak clustering.

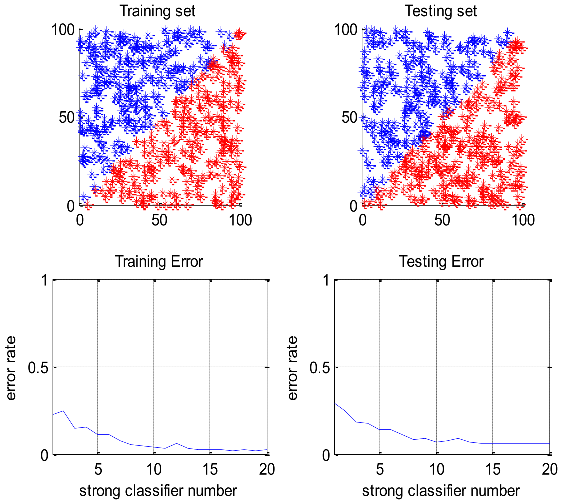

FIGURE 4. The results of samples 300according to strong clustering.

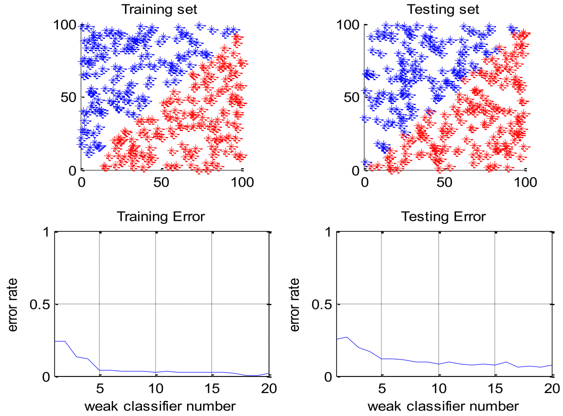

FIGURE 5. The results of samples 100 according to weak clustering.

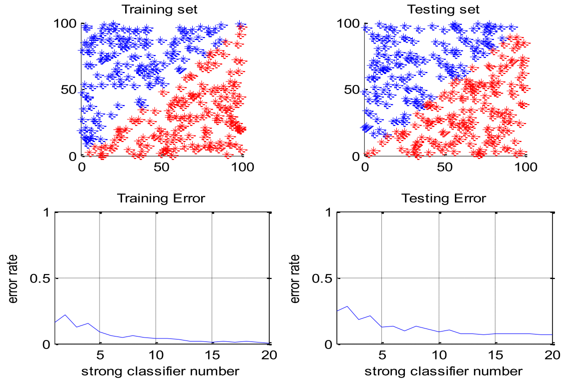

FIGURE 6. The results of samples 100 according to strong clustering.

significantly reduced compared with the data envelope front. We all known that the comparative experimental results are the judgments of the selected hybrid clustering methods on the optimal number of clusters in different data sets.

On the basis of the above simulation experiment research, we will conduct comparative experiments in groups and the experimental results of hybrid SBM clustering methodsare shown in FIGURE.7-FIGURE.12 as follows.

Through the experimental study of the above series of examples, it is found that the non convex data frontier performs better in clustering accuracy and clustering accuracy. Through the analysis of the results, we can find that the experimental error of the hybrid clustering algorithm is smaller than that of other clustering algorithms specifically. It is also found that the experimental error of weak clustering algorithm is smaller than that of the strong algorithm and the result is very sensitive to the hybrid algorithm. Moreover, we also found that the experimental error became larger and larger with the increase of sample size. Simultaneously,we have done some pioneering work on the clustering analysis of clustering variable data from the aspects of the effectiveness evaluation index principle based on nonparametric frontier clustering, the evaluation ability measurement, the proposed new effectiveness index and clustering algorithm,

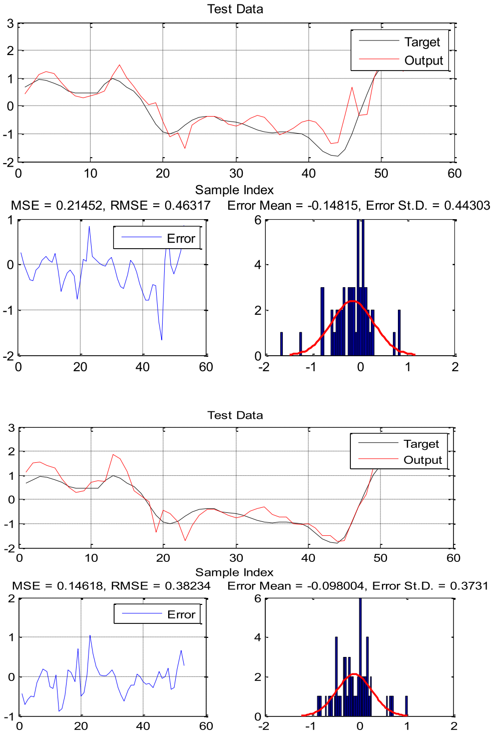

FIGURE 7. The comparison experiment results of strong clustering group 500.

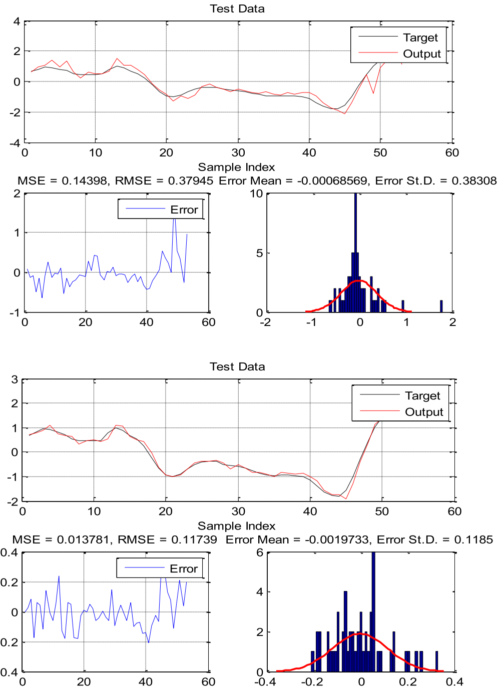

FIGURE 8. The comparison experiment results of weak clustering group 500.

but there are still problems that need further improvement and in-depth research. In the study of existing internal indicators, the number of analysis indicators can still be expanded, so that the internal evaluation of clustering variable data

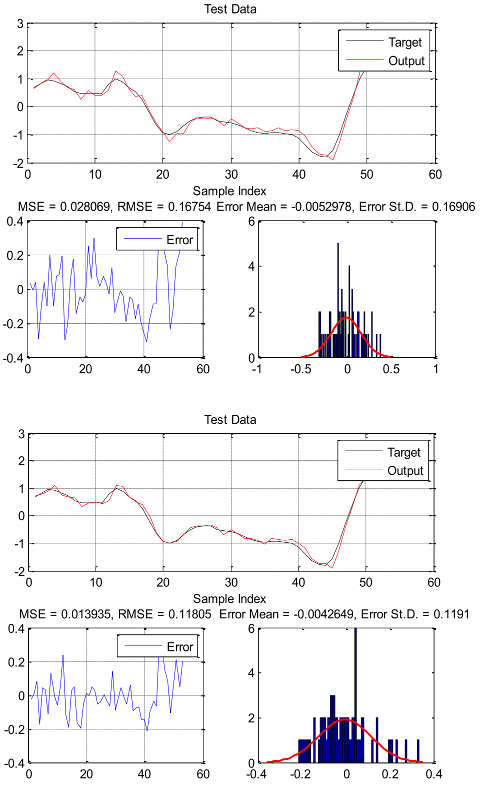

FIGURE 9. The comparison experiment results of strong clustering group 300.

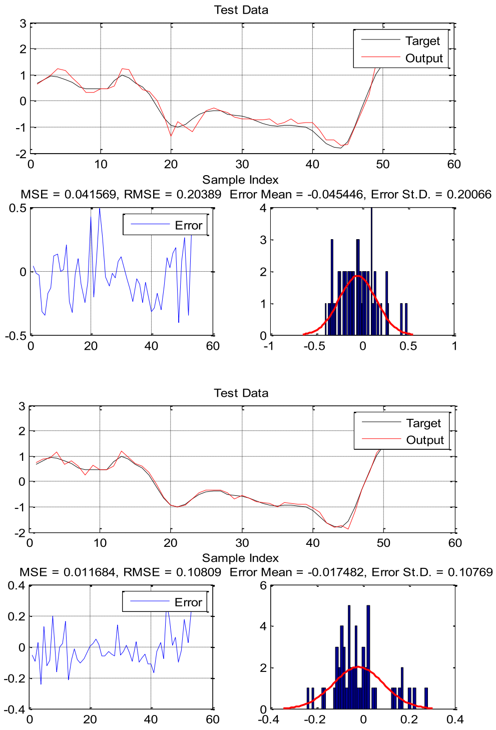

FIGURE 10. The comparison experiment results of weak clustering group 300.

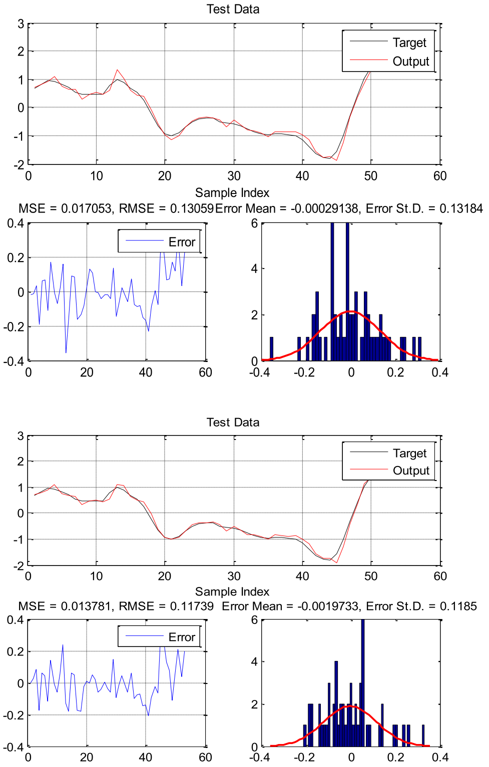

FIGURE 11. The comparison experiment results of strong clustering group 100.

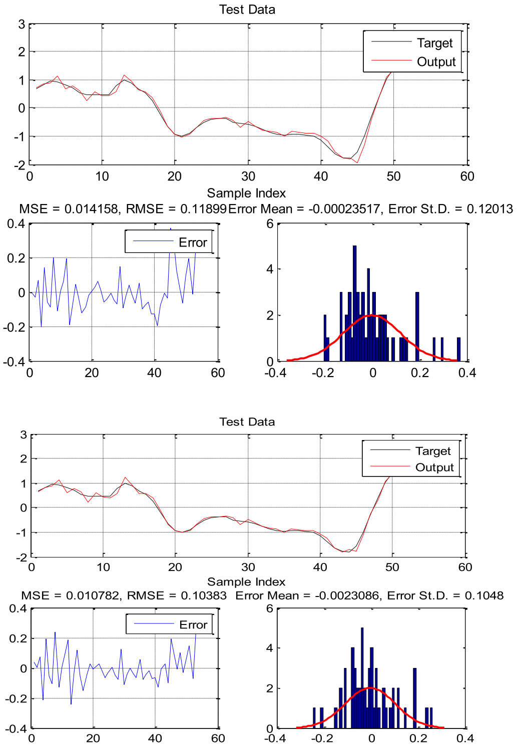

FIGURE 12. The comparison experiment results of weak clustering group 100.

can be more widely studied and making the relevant conclusions more reliable. At the same time, we can further study the compactness within attribute classes and the separation between attribute classes. By studying the monotony trend of the number of attribute classes, we explore the deep principle of the result evaluation of the hybrid clustering method. The experimental results show that by reasonably setting the weights of different indicators, the weighted indicators solve the shortcomings of traditional clustering effectiveness indicators, and can better identify the best number of clusters for a given dataset, which provides a new idea for the clustering effectiveness research of fuzzy time series.

VII. CONCLUSION

We found that the multivariate fuzzy time series clustering method based on Slacks Based Measure (MFTS-SBM) is a nonparametric linear programming model, the clustering method based on nonparametric frontier can provide a piecewise linear envelope frontier without relying on the form of clustering function, so as to distinguish the two groups of sample data. Moreover, it has found that time series data has the characteristics of large data volume and high data dimensions. The main advantage between time series clustering problems and traditional clustering problems is that each variable of time series data has an attribute ordering relationship, and the interrelationship between its attribute variables is independent of their relative positions. Meanwhile,we innovatively propose a novel hybrid clustering method based on nonparametric frontier and it can expand effectively the application of nonparametric frontier in clustering problems from different perspectives. Furthermore, the clustering mechanism of the frontier clustering model is explained through the small sample data in the example and the clustering effect of the model is verified and compared with the small sample data. Correspondingly, unlike monotone attributes, only one clustering frontier is required to depict the boundary of values. Consequently, the non-monotone attributes require multiple fronts under different preference directions to depict different combinations of upper and lower bounds and the intersection of corresponding multiple fronts constitutes the data fronts of a positive class. On the basis of data envelopment analysis, we propose a preference adaptive direction distance function to measure the relative distance between the clustering point and the frontier in a specific preference direction. In addition, we design a new clustering algorithm based on data frontier. In clustering problems with non monotone attributes, it also considers the necessity of convexity hypothesis and constructs non convex data frontier and non parametric frontier respectively.Through the simulation experiment results, we found that the hybrid SBM clustering method performs particularly well under small samples, and outperforms existing clustering models and methods in unbalanced data sets.

In a word, under the background of time series problem, it improves the existing SBM frontier clustering method by combining the clustering methods of nonparametric methods. Through the experimental results, we found that the experimental error of hybrid clustering algorithm based on nonparametric frontier for time series has great difference with the increase of sample size and the type of clustering algorithm. In the meantime, a series of clustering methods based on nonparametric frontier are constructed combined fuzzy clustering validity evaluation methods. Compared with some classical clustering methods, frontier clustering method also gives a good clustering performance. Although the existing research has proposed clustering effectiveness indicators, it lacks the comparison of algorithm effectiveness indicators. Firstly, we systematically analyzes the main clustering validity indexes of fuzzy time series and their existing problems. On this basis, the validity index of fuzzy clustering based on time series is constructed. Simultaneously, the simulation experimental analysis shows that the hybrid clustering method has high efficiency evaluation ability and low computational complexity through comparative experiments. Moreover,it is also found through multiple-experiments that the evaluation ability of hybrid clustering algorithm will be relatively reduced on the data sets with low quality of the repeated clustering results. Furthermore,we improve the performance of the hybrid clustering method in practical applications by using relevant methods. At the same time, it puts forward a new effectiveness index from the principle of cluster effectiveness evaluation index and evaluation ability measurement in this paper. At the same time,the results have made pioneering work and provided new ideas for clustering analysis of complex data. Therefore, our final results have important guiding significance and theoretical significance.

REFERENCES

[1] R. Umatani, T. Imai, K. Kawamoto, and S. Kunimasa, "Time series clustering with an EM algorithm for mixtures of linear Gaussian state space models," *Pattern Recognit.*, vol. 138, Jun. 2023, Art. no. 109375, doi: 10.1016/j.patcog.2023.109375.

[2] R. Cerqueti, M. Giacalone, and R. Mattera, "Model-based fuzzy time series clustering of conditional higher moments," *Int. J. Approx. Reasoning*, vol. 134, pp. 34–52, Jul. 2021.

[3] P. D'Urso, L. A. García-Escudero, L. De Giovanni, V. Vitale, and A. Mayo-Iscar, "Robust fuzzy clustering of time series based on B-splines," *Int. J. Approx. Reasoning*, vol. 136, pp. 223–246, Sep. 2021.

[4] M. V. N. K. Prasad, S. Nickolas, and G. R. Gangadharan, "Fuzzy representational structures for trend based analysis of time series clustering and classification," *Knowl.-Based Syst.*, vol. 222, Jun. 2021, Art. no. 106991, doi: 10.1016/j.knosys.2021.106991.

[5] F. Li, M. Zhou, S. Li, and T. Yang, "A new density peak clustering algorithm based on cluster fusion strategy," *IEEE Access*, vol. 10, pp. 98034–98047, 2022.

[6] P. D'Urso and E. A. Maharaj, "Wavelets-based clustering of multivariate time series," *Fuzzy Sets Syst.*, vol. 193, pp. 33–61, Apr. 2012.

[7] Á. López-Oriona, P. D'Urso, J. A. Vilar, and B. Lafuente-Rego, "Quantile-based fuzzy C-means clustering of multivariate time series: Robust techniques," *Int. J. Approx. Reasoning*, vol. 150, pp. 55–82, Nov. 2022.

[8] Á. López-Oriona, J. A. Vilar, and P. D'Urso, "Quantile-based fuzzy clustering of multivariate time series in the frequency domain," *Fuzzy Sets Syst.*, vol. 443, pp. 115–154, Aug. 2022.

[9] K. T. Chui, "Driver stress recognition for smart transportation: Applying multiobjective genetic algorithm for improving fuzzy c-means clustering with reduced time and model complexity," *Sustain. Comput., Informat. Syst.*, vol. 35, Sep. 2022, Art. no. 100668, doi: 10.1016/j.suscom.2022.100668.

[10] H. Saberi, A. Rahai, and F. Hatami, "A fast and efficient clustering based fuzzy time series algorithm (FEFTS) for regression and classification," *Appl. Soft Comput.*, vol. 61, pp. 1088–1097, Dec. 2017.

[11] S. Askari, N. Montazerin, and M. H. F. Zarandi, "A clustering based forecasting algorithm for multivariable fuzzy time series using linear combinations of independent variables," *Appl. Soft Comput.*, vol. 35, pp. 151–160, Oct. 2015.

[12] A. Lemos, W. Caminhas, and F. Gomide, "Multivariable Gaussian evolving fuzzy modeling system," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 1, pp. 91–104, Feb. 2011.

[13] E. Bas and E. Egrioglu, "A fuzzy regression functions approach based on Gustafson–Kessel clustering algorithm," *Inf. Sci.*, vol. 592, pp. 206–214, May 2022.

[14] R.-W. Po, Y.-Y. Guh, and M.-S. Yang, "A new clustering approach using data envelopment analysis," *Eur. J. Oper. Res.*, vol. 199, no. 1, pp. 276–284, Nov. 2009.

[15] J. J. Krüger, "Comment on 'a new clustering approach using data envelopment analysis,"' *Eur. J. Oper. Res.*, vol. 206, no. 1, pp. 269–270, Oct. 2010.

[16] D. Ben-Arieh and D. K. Gullipalli, "Data envelopment analysis of clinics with sparse data: Fuzzy clustering approach," *Comput. Ind. Eng.*, vol. 63, no. 1, pp. 13–21, Aug. 2012.

[17] B. Kim, H. Lee, and P. Kang, "Integrating cluster validity indices based on data envelopment analysis," *Appl. Soft Comput.*, vol. 64, pp. 94–108, Mar. 2018.

[18] M. G. Tsionas, "Clustering and meta-envelopment in data envelopment analysis," *Eur. J. Oper. Res.*, vol. 304, no. 2, pp. 763–778, Jan. 2023.

REN-LONG ZHANG was born in Shaoyang, in 1976. He received the Ph.D. degree from the School of Business Administration, Hunan University, Changsha, Hunan, China, in 2015. He is currently an Associate Professor and a Master Supervisor with Guizhou University. He has presided over and participated in a number of national and provincial projects. He has published more than 30 high-quality papers. His research interests include big data science and analysis, system optimization and operation management, and intelligent algorithms.

XIAO-HONG LIU was born in Yiyang, in 1980. She received the Ph.D. degree from the School of Business Administration, Hunan University, Changsha, Hunan, China, in 2019. She is currently an Associate Professor and a Master Supervisor with Guizhou University. She has presided over and participated in a number of national and provincial projects. She has published more than 20 high-quality papers. Her research interests include big data science and analysis, low-carbon supply chain management, and intelligent algorithms.