International Journal of Innovative Computing, Information and Control
Volume **18**, Number **1**, February **2022**

 ISSN 1349-4198
pp. 41–56

A SPARSE REPRESENTATION DATA COMPRESSION ALGORITHM FOR POWER PLANT

Shuanzhu Sun*, Bin Sun, Qixiang Wang and Chunlei Zhou

Jiangsu Frontier Electric Technology Co., Ltd.
No. 58, Suyuan Avenue, Nanjing 211102, P. R. China
{ 15905166668; 15905166652; 13851845492 }@163.com
*Corresponding author: 15905166613@139.com

Received July 2021; revised November 2021

Abstract. *With the advent of the big data era, the volume and rate of increase of the electric power data have reached an unparalleled level that has severely exceeded the scope of current data storage and network transmission capacity. To mitigate such circumstance, it is of particular significance to study how to reduce the data volume on the premise of retaining the useful information of power plant's original data. By analyzing the data features of the power plant's data and taking the data density distribution as the criterion, the corresponding data have been sorted out according to a new zone limitation strategy without destroying the original data structures, thus, proposing a sparse representation data compression algorithm for the power plant in the paper. The proposed algorithm filters the similar or overlapping data in the original data by constantly updating the zone positions through the iteration of time. The maximum filtration criterion can be utilized to limit the amount of data filtering in a single circular zone and prevent the data structure from being destroyed. In this way, the compressed data will be further evaluated by the twin support vector regression. The simulation experiment results indicate that, compared with the conventional method, the proposed algorithm can make the compression ratio lower than 15% and retain the features of original data after compressed. The results demonstrate the feasibility and the validity of the proposed algorithm.*

Keywords: Data compression, Density distribution, Filtration criterion, Twin support vector regression

1. **Introduction.** The power equipment, playing an important role in terms of the daily production and life, is essential to the running of the electric power system. The power equipment is indispensable to people's daily life, and any activity will depend on the power industry. As a result, government has vigorously practiced the reform of the power industry system and the upgrading of pertinent electric power technology. However, the imperfection of the process equipment and field data involving the mentioned industry leads to many problems, including instable operation state, diversified equipment parameters and higher sudden faults and parameter complexity. It is of vital forward-looking significance to monitor the safety and economy of power equipment for power production in real time to raise the real value of the power generation.

Over the past few years, rapid development has been made for the big data technology. It has been served to many fields, including public security, AI, and the state monitoring of industrial production [1]. It has become a hit about how to make the big data technology better serve the power industry in recent years. For any power enterprise, a large amount of data will be generated from the daily monitoring of the power equipment, in addition

DOI: 10.24507/ijicic.18.01.41

to that, there will be other data related to equipment management, operation, overhaul and power network maintenance. Confronting such enormous and complicated data, the challenges for the power enterprise to consider are how to efficiently storage and utilize such data to better serve the power industry. With the rapid development and constant maturity of AI and big data technology, data mining can be adopted to get the information needed from the huge amounts of data, and in this way, the conundrum of making use of power data in power enterprises can be better solved [2]. Nowadays, many researchers have begun to conduct data mining on massive power plant data, hoping to find useful information to improve the value of electricity production. Fachrizal et al. proposed on applying cluster analysis techniques to evaluate the level of power quality parameters of a virtual power plant. The conducted research concerns the application of the K-means algorithm in comparison with the agglomerative algorithm for power quality data, which have different sizes of features [3]. For Khaleel et al., an analytical model has been developed to predict the effectiveness of the closed feedwater heaters based on the measured data available from a coal-fired thermal power plant. Invoking this model, a 'predictive model' has been developed to anticipate the energy and exergy-based behavior of the coal-fired power plants. Based on this model, a numerical code has been developed to analyze an existing coal-fired power plant using Engineering Equation Solver software [4]. Liang et al. introduced the subsystems and structure of the distributed control system, and lists the commonly used analysis methods related to safety and reliability. Since traditional methods may no longer be suitable for direct analysis of current complex systems, it proposes to use principal component analysis algorithm to study power plant operation control system data. The simulation results are compared with the results of other algorithms to verify the usability and advantages of the principal component analysis algorithm [5]. Choi et al. discussed data-driven fault diagnosis of the power plant reheater tube leakage based on their operating data. From the temperature sensors, fault data and normal data are measured. Mahalanobis distance (MD) analysis was performed to quantitatively analyze whether the distribution of fault data differed from that of the normal data. Finally, they demonstrated the feasibility of the proposed approach to detect reheater tube leakage prior to the failure [6]. At the same time, Desell et al. used recurrent neural networks to predict the long term of coal fired power plant data, and the simulation results show that the method has a certain practical value [7]. The majority of research now focuses on analyzing case studies of hydropower plants in different issues. Michal et al. presented the investigation that is based on real hydropower plants. The investigation discusses different input databases of hydropower plants for cluster analysis. This enables the reduction of the size of the input database of hydropower plants with maintaining the data features for cluster analysis [8]. Although predicting the amount of power generation is crucial for efficient operations, it is not easy because of fluctuating nature of wind speed. Baek et al. applied a deep neural network method to predicting wind power generation based on weather forecast data. The prediction performance of the model was evaluated with wind power generation data, and the simulation experiment results show that the prediction effect of this method is better [9]. In addition to the above studies, there are many related studies that have proved the importance of studying power plant data.

However, the premise for the data mining of vast power plant data is to efficiently store the data, as only the stored data can be utilized for the subsequent big data analysis, mining, and prediction. Despite that a large amount of storage equipment can be used to solve the problem, the data types are various and most of the data are generated from the production process, which are dynamic with fast refreshing speed and huge amount.

It would be a major burden for the enterprises to preserve all real-time data in the form of original format in a long term.

In terms of problems mentioned above, the universal practice is to compress the power plant data, which will be stored in the storage equipment, by doing so, the storage space can be substantially reduced, and the transmission efficiency of data can be raised [10]. There are two common types of data compression in the industrial community, which are lossless compression and lossy compression. The principle of the former is to adopt the statistics redundant data for the compression, making the compression ratio lower; as for the latter, the compression ratio is larger with the sacrifice of some data. The lossy compression may not be able to completely restore the original data, the entire structure will not be compromised, though. There is no special compression algorithm to compress the data in terms of the data of power plant except for the universal compression algorithm [11-13]. As the actual data of power plant are voluminous and diversified, the data are stored in seconds, making the amplitude of fluctuation of the generated data smaller; therefore, the lossy compression is more suitable to the data compression [14,15]. The popular spinning door transformation (SDT) algorithm discards too much data when it is utilized to compress the data of power plant, bringing on the data fault and the destroyed data structure. As a result, it is necessary to consider how to reduce the data storage volume while raising the data transmission efficiency by studying the compression algorithms with higher compression ratios for specific targets via the analysis and utilization of power plant data.

In terms of above problems, starting from the actual condition of power plant data, the paper takes the compression ratio and accuracy into consideration and puts forward a sparse representation data compression algorithm for power plant. The main contributions of this paper are summarized as follows.

1) A zone limitation strategy has been adopted to filter and screen the similar or overlapping data by constantly iterating over the location of the zone. Furthermore, the zone can limit the sample numbers to be filtered, avoiding the destruction of the original data structure.

2) A filter criterion has been adopted to screen the data while automatically filtering the abnormal points. In addition, the size of the filtration volume can be limited for the data period with smaller amplitude of fluctuation, preventing the obvious fault phenomenon occurring in the compressed data.

3) The compression ratio can be controlled by constantly adjusting the pre-set parameters to achieve the higher portability and extensibility for different tasks in real life.

In addition, we conduct extensive experiments on artificial datasets, and actual power plant to demonstrate the superiorities of the proposed algorithm over some state-of-the-art algorithms in terms of compression performance and anti-interference capability.

The remainder of this paper is arranged as follows. In Section 2, we briefly review the overview of compression algorithm. In Section 3, we mainly analyze the power plant data. Section 4 introduces the proposed sparse representation data compression algorithm for power plant in detail. Section 5 presents the experimental results and analyses. The conclusions are summarized in Section 6.

2. **Overview of Compression Algorithm.**

The lossless compression can be divided into statistics-based compression and dictionary-based compression according to the compression model. The former mainly consists of the run-length coding, Huffman coding, arithmetic coding etc. In terms of the dictionary-based compression, it comprises LZ77 coding, LZW algorithm [16-20]. The lossy compression mainly includes quantization algorithm, spinning door transformation algorithm [21,22]. The Huffman coding algorithm,

the typical lossless compression and the spinning door transformation algorithm, the typical lossy compression, will be mainly introduced below.

2.1. **Huffman coding.** The major steps of Huffman coding are 1) The source symbols from large to small, Table 1 shows the sorted source symbols; 2) Add together the probability of the two symbols with the least probability while allocating the code word length, generating the new probability; 3) Re-sort the new probability set generated from the composition and repeat step 2), in this way, the sum of the last two probability will be 1; 4) Construct the code tree from bottom to top and obtain the coding of the symbol formation from the results of the tree. Figure 1 shows the process of Huffman coding.

TABLE 1. The sorted source symbols

Symbol	A	B	C	D	E
Frequency	15	7	6	6	5

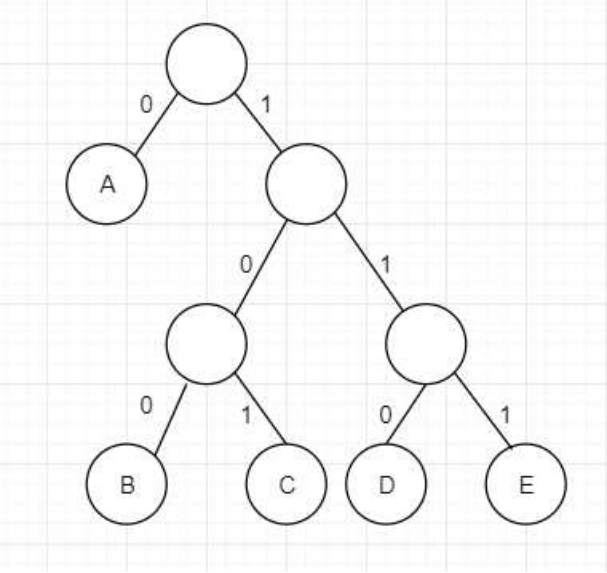

FIGURE 1. The process of Huffman coding

2.2. **Spinning door transformation algorithm.** The spinning door transformation (SDT) algorithm is a kind of lossy compression method like the linear fitting. It has many merits, including higher compression ratio, simple realization, and controllable error. Figure 2 displays its principle. The starting point t_0 is the previous storage point and the supporting points are the top and bottom points of E, which is the distance to t_0. Two doors will be established, and the doors are closed when there is only one data. With the increase of data points, the door will spin to open with an extensible. Once the door is open, it could not be closed; as long as the interior angle between two doors is smaller than 180°, it will keep spinning; it will stop spinning when the angle between two doors is no smaller than 180°, the data points of previous period will be stored, and a new round of compression will be started from this point. As shown in figure, the compression section 1 of the straight line from t_0 to t_4 replaces the data point of $t_0 \sim t_4$; in terms of compression section 2, the straight line from t_4 to t_7 substitutes the data point of $t_4 \sim t_7$.

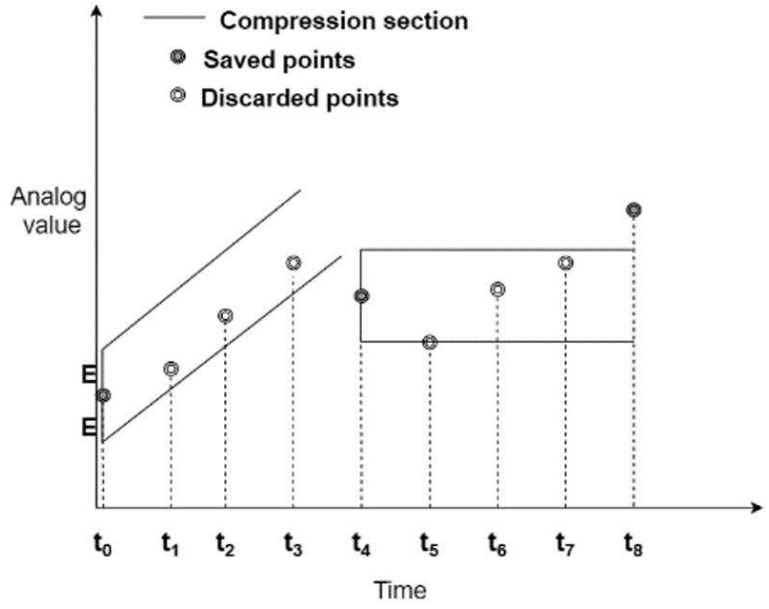

FIGURE 2. The schematic diagram of spinning door transformation algorithm

3. **Problem Analysis.** The data of power plant is a kind of time series data, featuring for the voluminous data size and weaker regularity of amplitude of fluctuation. As a result, it needs to understand the data features and configuration of power plant data to study the compression algorithm with higher compression ratio under certain goals.

3.1. **Data description.** The data format gathered from the coal-fired unit (CFU) data of power plant is .txt, among which, the integral data can be divided into formatted files of real-time and previous data. In the .txt file, each data represents the data of a measuring point and each file contains several lines, with a total number of lines within 20,000. Each line can be separated into three fields, representing the code field of measuring point, time stamp, and load value, respectively. The field will be isolated by TAB key. The time stamp is the difference formed by the time value of the measuring point minus the value of 0-hour 0-minute 0-second on January 1, 2001. The code field is the desensitization data for the name of the power plant (the sampling frequency is five seconds). Figure 3 signifies the numerical fluctuation graph obtained from the data visualization. As can be seen from it, the regularity of time series data is weak, showing significant peaks in some locations, and there is no distinct change in the amplitude of data fluctuation over a period of time.

3.2. **Problem analysis.** According to the time series data of power plant and the sampling frequency, one of the equipment can generate up to 518,400 data, while there are over 10,000 equipment in the power plant running simultaneously. Besides, the power plant is always running uninterruptedly, and the generated data volume is immeasurable. Therefore, the priority to store the original data of power plant is to compress the data with suitable method. However, both the Huffman coding and the lossless compression method based on dictionary are facing a problem of low compression ratio. Consequently, a kind of lossy compression method can be utilized with the premise that the generated loss will not exert significant impact on the subsequent data analysis and mining.

Compared with other type, the data of power plant is a time-series data, showing an irregular and random float due to the complicated system structure. The traditional lossy compression method, such as spinning door transformation algorithm, may destroy the

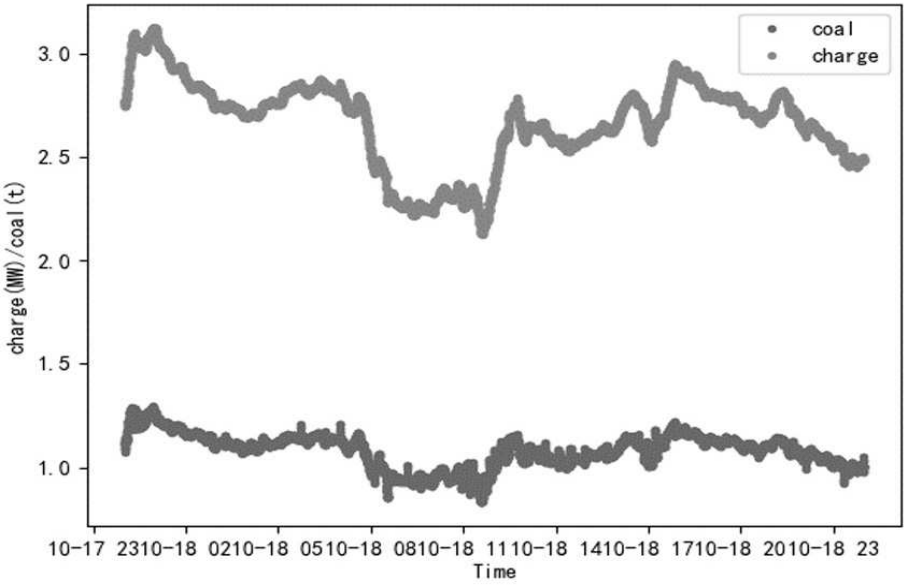

FIGURE 3. The visual figure of data

original data structure while compressing the data and have higher compression ratio. Therefore, a new sparse representation lossy compression algorithm has been put forward in this paper. The method adopts a new strategy that can not only efficiently compress the data but also effectively retain the features of original data, mitigating the destruction of data structure and high compression ratio problem caused by the conventional lossy compression.

4. A Sparse Representation Data Compression Algorithm for Power Plant.

In this section, we first discuss the sparse representation data compression algorithm for power plant.

4.1. A sparse representation data compression algorithm.

A sparse representation lossy compression algorithm has been proposed in this paper in terms of power plant data with small amplitude of fluctuation and large data volume. The specific representation for the algorithm is that the initial data point will be selected as the compressed data set. The circular zone with a radium of r will be adopted to measure the shortest distance between each point in the original data and each point in the compressed data set, determining whether this point is smaller than the self-defined δ. If it is smaller than r, then, determine whether the sample density of the point is smaller than the volume threshold δ. If it is, filter the point, otherwise, the point shall be added to the compressed data set; if it is larger than r, then the point should be directly added to the compressed data set. The final compressed data set can be obtained via repeated iteration. Figure 4 shows the schematic diagram of the sparse representation data compression algorithm of power plant.

4.2. Algorithm flow.

Given the data set of power plant $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and the compressed data set $Q = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_q\}$. The Euclidean distance between two points shall be firstly calculated:

$$dis = \arg\min D(\boldsymbol{x}_i, \boldsymbol{d}_j) \quad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, q$$

where $D(\boldsymbol{x}_i, \boldsymbol{d}_j) = \|\boldsymbol{x}_i - \boldsymbol{d}_j\|$ signifies the Euclidean distance between sampling point $\boldsymbol{x}_i$ and $\boldsymbol{d}_j$.

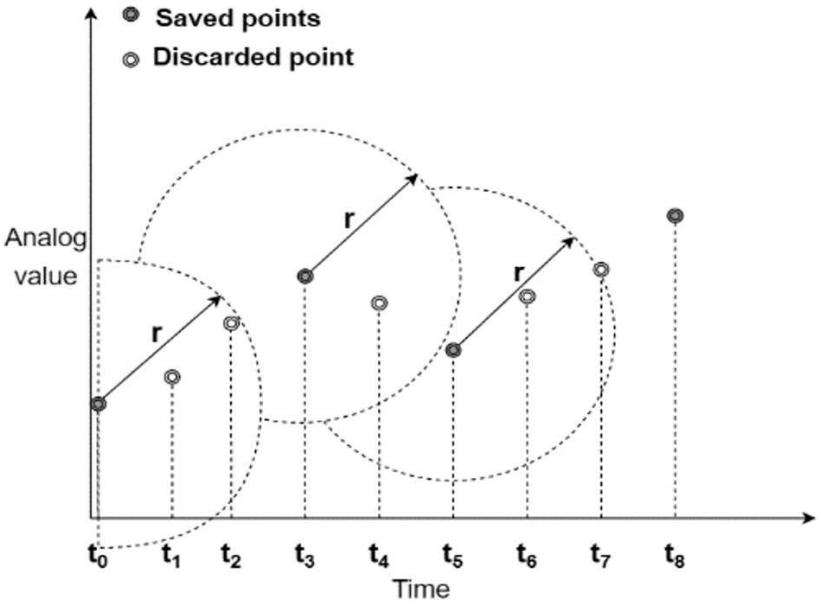

FIGURE 4. The schematic diagram of the sparse representation data compression algorithm of power plant

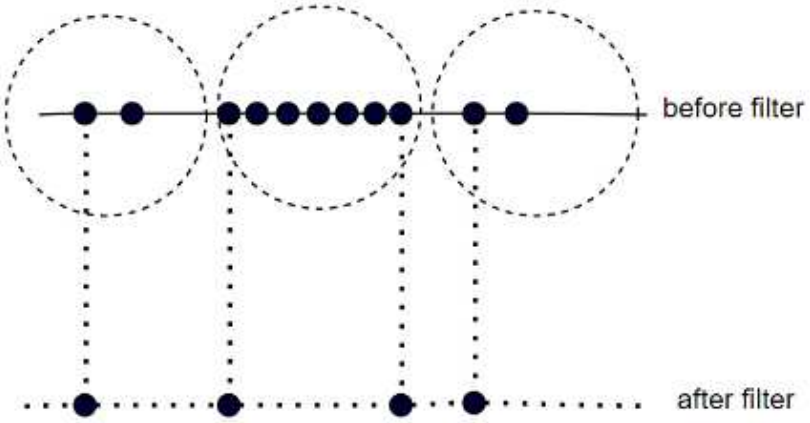

FIGURE 5. The one-dimensional visualization diagram of compression algorithm

In terms of the compressed data set $\boldsymbol{Q}$, have it initialized, $\boldsymbol{Q} = \{\boldsymbol{x}_0\}$. Firstly, iterate each data point of data set $\boldsymbol{S}$, obtaining dis and determining whether the obtained distance is smaller than the self-defined error threshold r. It is to qualify the circular zone on the compressed data set, and any point within the circular region on the compressed data set will need to be filtered. Extend Figure 4 and the one-dimensional visualization can be adopted to expound the filtering criterion of the algorithm, which is shown as Figure 5. The upper half part signifies the original data set, and the lower half part represents the compressed data set. Taking C_1 as the sample $\boldsymbol{x}_0$, the points within the circular zone of r will be eliminated. One sample point has been eliminated, and the condition of C_5 is similar to that of C_1. In terms of C_2, C_3 and C_4, instead of one point, there are 3 points reserved here. The principle will be described below. 5 sample points were retained out of 13 sample points of the original data through the compression algorithm.

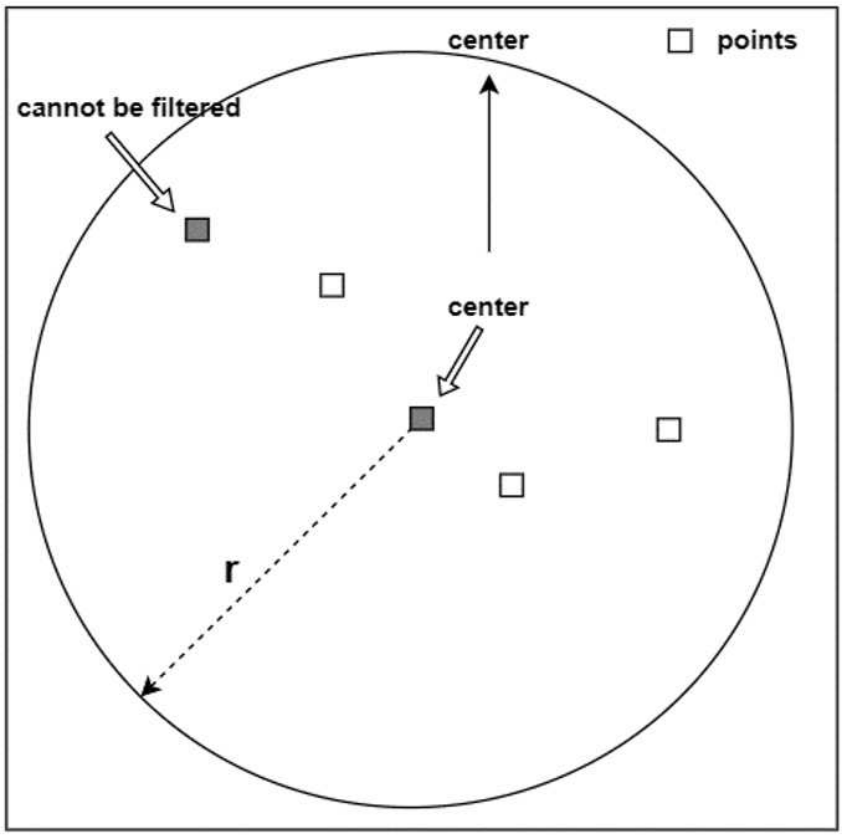

FIGURE 6. The visualization diagram of filtering criterion

As the data of power plant have small amplitude of fluctuation in certain period, the data fault phenomenon is prone to occur if all the points within the circular zone are eliminated. As a result, a new filtering mechanism is added here, which can be simply signified by Figure 6. First, there are five data points in the circular zone r, represented by '□' and the set number of threshold is δ; then, determining whether the filtered data points near the data points '□' framed by the circular zone are greater than the set threshold δ, if that is the case, the last data point greater than δ will be compulsively added to the compressed data set, which are the black data points '□'. At last, the five data points within the circular zone will be compressed, obtaining two data points, which are the black and white data points '□'. The maximum value of the filtration yield can be ensured through this strategy, which also explains the three retaining points in Figure 2, which are C_2, C_3 and C_4. Here $\delta = 5$.

In terms of circular radius r and the number of threshold δ described above, the proportion of the compressed data set and the original data, namely, the compression ratio, can be utilized to determine the optimal value of these two parameters. Or the grid search algorithm can be used to find the optimal pair of these two parameters. The pseudocode of this algorithm is shown in Algorithm 1.

4.3. **Twin support vector regression.** Given a training sample set $T = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$ (m is the size of the training sample set) such that $x_i \in \mathbb{R}^d$ (d is the dimension of the training sample set) is an input and $y_i \in \mathbb{R}$ is a target output. The training sample is represented by m row vector $\boldsymbol{A}_i$, $i = 1, 2, \ldots, m$. And the training sample $\boldsymbol{A}_i = (\boldsymbol{A}_{i1}, \boldsymbol{A}_{i2}, \ldots, \boldsymbol{A}_{id})$ is in the d-dimensional real space $\mathbb{R}$, where $\boldsymbol{A} = (\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_d)$ and $\boldsymbol{Y} = (y_1, y_2, \ldots, y_m)$. In linear case, the twin support vector regression (TSVR) determines the primal problems by the ε_1 insensitive lower bound $f_1(x) = \boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x} + b_1$ and

Algorithm 1: A sparse representation data compression algorithm

Input: $\{\boldsymbol{x}_i\}_{i=0}^{n-1}$; $\boldsymbol{x}_i \in \mathbb{R}^d$
Output: $\boldsymbol{Q} = \{\boldsymbol{d}_i\}_{i=0}^{q}$, $index$; $\boldsymbol{d}_i \in \mathbb{R}^d$
Initialization: $\boldsymbol{Q} = \{\boldsymbol{x}_0\}$, r, δ, $index = \{0\}$, $\boldsymbol{L} = \{:\}$
for $i = 1, 2, \ldots, n-1$:
//find the nearest point $\boldsymbol{x}_i$ near $\boldsymbol{d}_0$
  $\min dis = \arg\min D(\boldsymbol{x}_i, \boldsymbol{d}_0)$
  $m = len(\boldsymbol{Q})$
 for $j = 0, 1, \ldots, m-1$:
//find the nearest distance dis between the $\boldsymbol{x}_i$ and the set $\boldsymbol{Q}$
    $dis = \arg\min D(\boldsymbol{x}_i, \boldsymbol{d}_i)$
//set the dis as the neartest distance between the $\boldsymbol{x}_i$ and the set $\boldsymbol{Q}$
    if $\min dis \geq dis$:
      $\min dis = dis$
//if the dis is greater than r, put the $\boldsymbol{x}_i$ into the set $\boldsymbol{Q}$
 if $\min dis \geq r$:
    $\boldsymbol{Q} = \{\boldsymbol{x}_i, \boldsymbol{Q}\}$; $index = \{i, index\}$; $\boldsymbol{L} = \{\boldsymbol{L}; 1\}$
 else:
//if $\boldsymbol{x}_i$ do not have found point in the set $\boldsymbol{Q}$ within $\boldsymbol{L}$, put the $\boldsymbol{x}_i$ into the set $\boldsymbol{Q}$
    if $len(\boldsymbol{L}) \geq \delta$:
      $\boldsymbol{Q} = \{\boldsymbol{x}_i, \boldsymbol{Q}\}$; $index = \{i, index\}$; $\boldsymbol{L} = \{:\}$
    else:
      $\boldsymbol{L} = \{\boldsymbol{L}, 1\}$

the ε_2 insensitive upper bound $f_2(x) = \boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x} + b_2$ as follows:

$$\begin{aligned} &\min \frac{1}{2}\left\|\boldsymbol{Y} - \boldsymbol{e}\varepsilon_1 - (\boldsymbol{A}\boldsymbol{w}_1 + \boldsymbol{e}b_1)\right\|^2 + C_1\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\xi} \\ &\text{s.t.} \quad \boldsymbol{Y} - (\boldsymbol{A}\boldsymbol{w}_1 + \boldsymbol{e}b_1) \geq \boldsymbol{e}\varepsilon_1 - \boldsymbol{\xi}, \quad \boldsymbol{\xi} \geq \boldsymbol{0} \end{aligned} \tag{1}$$

$$\begin{aligned} &\min \frac{1}{2}\left\|\boldsymbol{Y} + \boldsymbol{e}\varepsilon_2 - (\boldsymbol{A}\boldsymbol{w}_2 + \boldsymbol{e}b_2)\right\|^2 + C_2\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\eta} \\ &\text{s.t.} \quad (\boldsymbol{A}\boldsymbol{w}_2 + \boldsymbol{e}b_2) - \boldsymbol{Y} \geq \boldsymbol{e}\varepsilon_2 - \boldsymbol{\eta}, \quad \boldsymbol{\eta} \geq \boldsymbol{0} \end{aligned} \tag{2}$$

where C_1 and C_2 are penalty parameters, $\varepsilon_1, \varepsilon_2 > 0$ are constant, $\boldsymbol{w}_1, \boldsymbol{w}_2 \subseteq \mathbb{R}^d$, $b_1, b_2 \subseteq \mathbb{R}$. $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are slack variables, and $\boldsymbol{e}$ is a unit column vector of dimension $d \times 1$.

By introducing the Lagrange multipliers $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, we can obtain the following dual problems of Equations (1) and (2):

$$\begin{aligned} &\max -\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{G}\left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\right)^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{\alpha} + \boldsymbol{f}^{\mathrm{T}}\boldsymbol{G}\left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\right)^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{\alpha} - \boldsymbol{f}^{\mathrm{T}}\boldsymbol{\alpha} \\ &\text{s.t.} \quad 0 \leq \boldsymbol{\alpha} \leq C_1\boldsymbol{e} \end{aligned} \tag{3}$$

$$\begin{aligned} &\max -\frac{1}{2}\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{G}\left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\right)^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{\gamma} - \boldsymbol{h}^{\mathrm{T}}\boldsymbol{G}\left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\right)^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{\gamma} + \boldsymbol{h}^{\mathrm{T}}\boldsymbol{\gamma} \\ &\text{s.t.} \quad 0 \leq \boldsymbol{\gamma} \leq C_2\boldsymbol{e} \end{aligned} \tag{4}$$

where $\boldsymbol{G} = [\ \boldsymbol{A} \quad \boldsymbol{e}\ ]$, $\boldsymbol{f} = \boldsymbol{Y} - \varepsilon_1\boldsymbol{e}$ and $\boldsymbol{h} = \boldsymbol{Y} + \varepsilon_2\boldsymbol{e}$.

Solving Equations (3) and (4), we can obtain the following regression function:

$$\boldsymbol{f}(\boldsymbol{x}) = \frac{1}{2}[\boldsymbol{f}_1(\boldsymbol{x}) + \boldsymbol{f}_2(\boldsymbol{x})] = \frac{1}{2}(\boldsymbol{w}_1 + \boldsymbol{w}_2)^{\mathrm{T}}\boldsymbol{x} + \frac{1}{2}(b_1 + b_2) \tag{5}$$

where $\left[\begin{array}{cc} \boldsymbol{w}_1^{\mathrm{T}} & b_1 \end{array}\right]^{\mathrm{T}} = \left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\right)^{-1}\boldsymbol{G}^{\mathrm{T}}(\boldsymbol{f} - \boldsymbol{\alpha})$ and $\left[\begin{array}{cc} \boldsymbol{w}_2^{\mathrm{T}} & b_2 \end{array}\right]^{\mathrm{T}} = \left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\right)^{-1}\boldsymbol{G}^{\mathrm{T}}(\boldsymbol{h} + \boldsymbol{\gamma})$.

By introducing kernel function $\boldsymbol{K}(\cdot,\cdot)$, we can easily extend the linear case to the non-linear case. The non-linear TSVR determines the primal problems by the ε_1 insensitive lower bound $\boldsymbol{f}_1(\boldsymbol{x}) = K\left(\boldsymbol{x}^{\mathrm{T}}, \boldsymbol{A}^{\mathrm{T}}\right)\boldsymbol{w}_1 + b_1$ and the ε_2 insensitive upper bound $\boldsymbol{f}_2(\boldsymbol{x}) = K\left(\boldsymbol{x}^{\mathrm{T}}, \boldsymbol{A}^{\mathrm{T}}\right)\boldsymbol{w}_2 + b_2$, which determines the final regression function.

Let $\boldsymbol{H} = \left[\begin{array}{cc} \boldsymbol{K}\left(\begin{array}{cc} \boldsymbol{A} & \boldsymbol{A}^{\mathrm{T}} \end{array}\right) & \boldsymbol{e} \end{array}\right]$, the final regression function of non-linear TSVR can be expressed as follows:

$$\boldsymbol{f}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{K}\left(\boldsymbol{x}^{\mathrm{T}}, \boldsymbol{A}^{\mathrm{T}}\right)(\boldsymbol{w}_1 + \boldsymbol{w}_2) + \frac{1}{2}(b_1 + b_2) \tag{6}$$

where $\left[\begin{array}{cc} \boldsymbol{w}_1^{\mathrm{T}} & b_1 \end{array}\right]^{\mathrm{T}} = \left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{f} - \boldsymbol{\alpha})$ and $\left[\begin{array}{cc} \boldsymbol{w}_2^{\mathrm{T}} & b_2 \end{array}\right]^{\mathrm{T}} = \left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{h} + \boldsymbol{\gamma})$.

In this section, we first give an overview of combining the above two algorithms, and then we present the pseudocode of this algorithm as shown in Algorithm 2.

Algorithm 2: Combine TSVR and the proposed algorithm.

Input: The training sample set T, the optimal parameters C_1, C_2, ε_1, ε_2, parameters r and $\delta = 5$, set iterator = 1;

Output: The regression function $f(x)$.

Step 1. Compress data using Algorithm 1;

Step 2. Remove outliers and get compressed data;

Step 3. Solving Equations (3) and (4) by the LMI (matlab toolbox);

Step 4. Construct the regression function $f(x)$ according to Equation (5);

Step 5. Compute the iterator value, if iterator < 1000, then update C_1, C_2, ε_1, ε_2, repeat Steps 3 and 4; else go to Step 6;

Step 6. Use the LMI to compute the final optimal solutions C_1, C_2 and ε_1, ε_2;

Step 7. Construct the final regression function $f(x)$ according to Equation (5).

5. Experimental Results and Analyses.

In this section, we first present the experimental design. Then, we discuss the parameter selection of different algorithms. Finally, we conduct extensive experiments on power plant datasets, artificial datasets.

5.1. Simulation analysis.

In this simulation, three types of data sets will be used to evaluate and verify the algorithm. The data type 1 is an artificial data set that satisfies the function $\sin c(x)$, which is a time series data set. The data are fetched from $[0, 2\pi]$ with an interval of 0.1; data type 2 is an artificial data set that suffices the Gaussian distribution. The data set is an ordinary structured data set; type 3 is the real data set of power plant (the data over a period).

Compare the proposed algorithm with the popular spinning door transformation (SDT) algorithm on the artificial data set that satisfies function $\sin c(x)$. The results are shown in Figure 7. As can be seen from the figure, the spinning door transformation algorithm is exceedingly sensitive to the maximum point of time series, and most of the retained data are near the maximum point while the distribution of the proposed algorithm is even; from the compression results, the compression ratio for the proposed algorithm is 0.17, while it is 0.43 for the spinning door transformation algorithm. To sum up, the

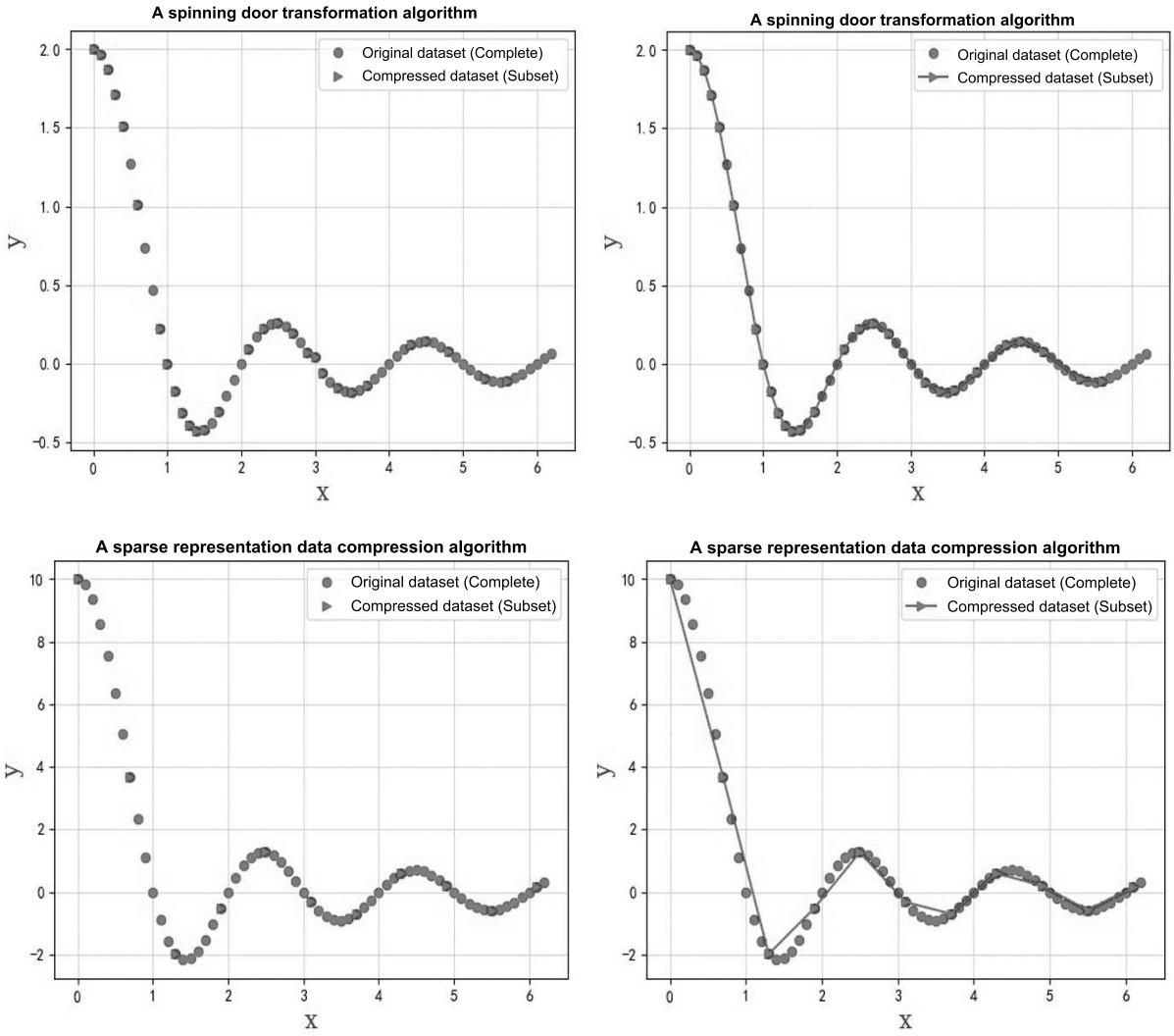

FIGURE 7. The original data and reconstitution of data

spinning door transformation algorithm is relatively stable and effective on the uniform and stationary time series data sets.

On the artificial data set that suffices Gaussian distribution, the results of the proposed algorithm are shown in Figure 8 (simulation will not be done for the spinning door transformation algorithm, as it is mainly for the time series data). Note: The red data points in the figure also belong to blue data points, and the overlapping makes the effect less distinct. As can be seen, the main idea of the proposed algorithm is to filter the similar or overlapping data for compression without altering the data structure, which is one of the merits of such algorithm. It can compress the time series data set as well as the structured data sets common in life in a flexible manner.

As stated above, compared to SDT algorithm, the compression rate of the proposed algorithm is effective and reasonable. And our algorithm is suitable for any situation, so the proposed algorithm has a wider application field than the traditional SDT algorithm. The relevant compression rate can be flexibly controlled by adjusting the parameters, so the flexibility of the proposed algorithm is better than SDT algorithm.

Part of the data of power plant were selected for the simulation experiment due to the consideration of time factor. Three sections of data H, M and L were randomly selected to carry out the simulation experiment. Figure 9 shows the analysis results of the comparison

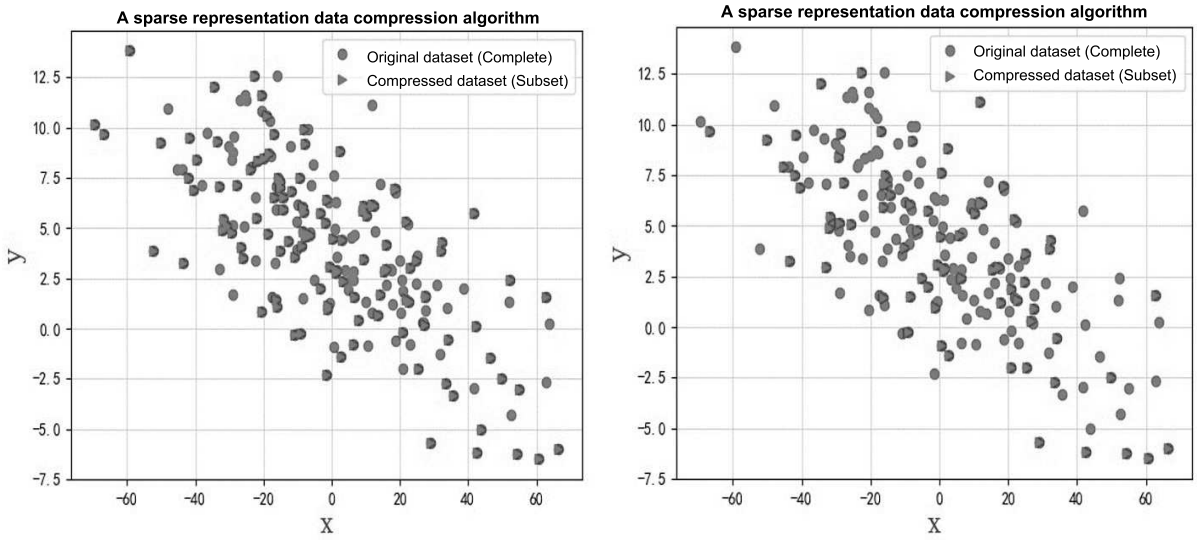

FIGURE 8. Visualization diagram of the algorithm effect with compression ratio of 0.63 (left) and 0.36 (right)

of the proposed algorithm and the spinning door transformation algorithm, among which, the left is the visualization results of the proposed algorithm for the same data set and the right is that of the spinning door transformation algorithm. The simulation results indicate that the proposed algorithm has better effect than that of the compared algorithm in terms of disordered and time series data sets with large amplitude of fluctuation. The results also demonstrate that, in terms of the three sections of data H, M and L, the compression ratio of the spinning door transformation algorithm was 0.91, 0.92 and 0.90, while that of the proposed algorithm was 0.12, 0.14 and 0.13. The difference is distinctive, revealing the defficiency of the spinning door transformation algorithm. This method, which measures the similarity of random and disordered data through slope and then conducts data filtering, has great limitations. The spinning door transformation algorithm can only be applied to the specific data for the lossy compression, which is not a good choice for the data of power plant.

In summary, on power plant data, the compression rate of our proposed algorithm is much better than that of traditional SDT algorithm. In particular, the compression rate of the traditional SDT algorithm is very low, so our proposed algorithm has wider applicability. It can be concluded from the above artificial dataset that our proposed algorithm has better applicability and stronger robustness.

5.2. **Result analysis.** To better evaluate the validity of the data reconstitution, the twin support vector regression was adopted to fit the mentioned data to assess whether the compressed data set can retain the features of the original data [23-26]. In terms of the parameters in the algorithm, the current more popular grid search was adopted in the simulation experiment to determine the optimal parameters. The selection range for the twin support vector regression penalty parameter C and the insensitive loss parameter ε is $\{10^i | i = -3, -2, \ldots, 3\}$. In terms of the selection of different data set parameters r and δ, the general practice is to select about 1% of the total number of data sets as r and δ are normally set as 5. Two regression performance metrics, namely, root-mean-square error (RMSE) and mean absolute error (MAE), have been adopted to appraise the estimated

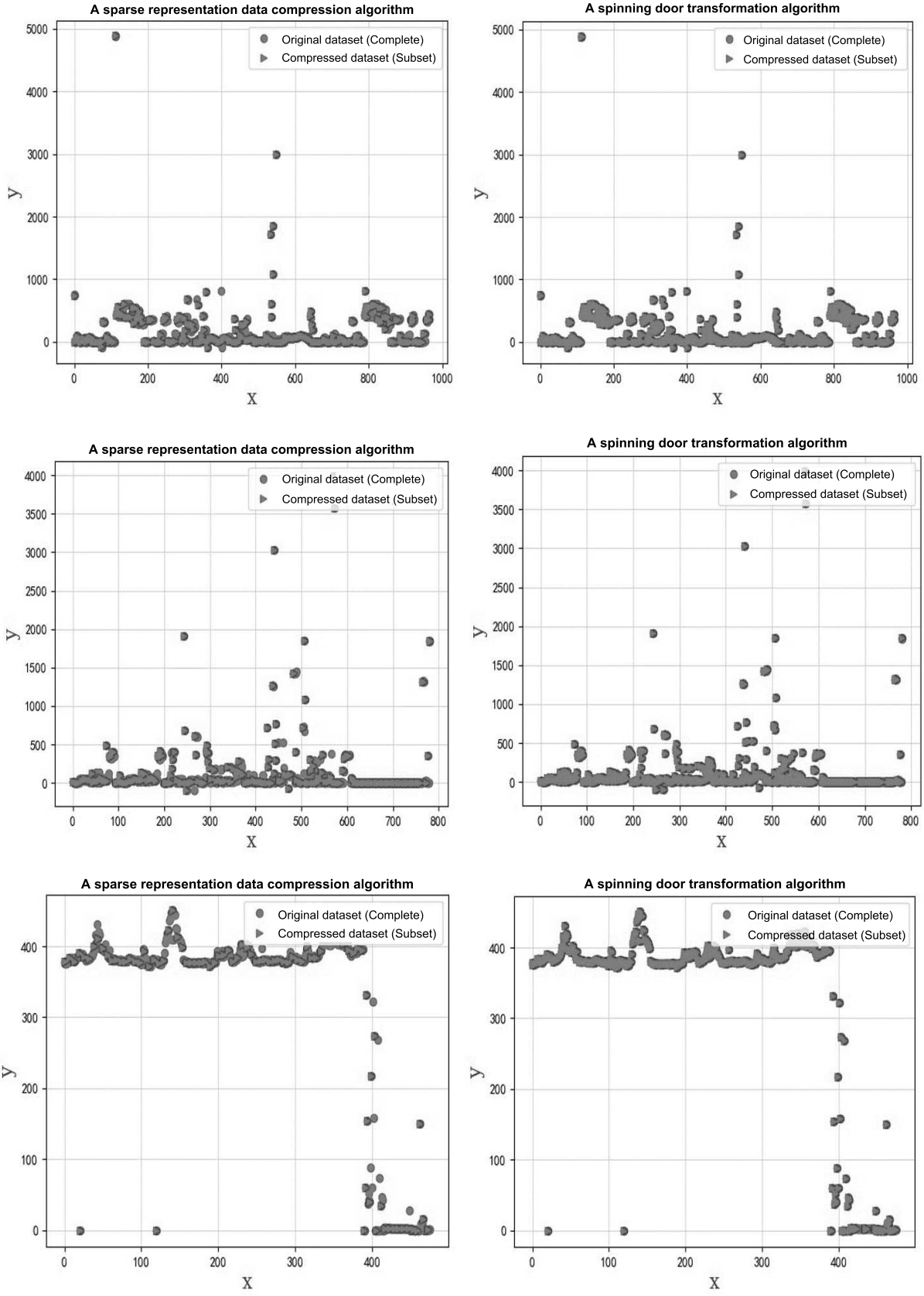

FIGURE 9. The original data and reconstitution of data

performance of the algorithm.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

where $\hat{y}_i$ signifies the predicted value, y_i is the true value, and n refers to the number of samples. The smaller the RMSE and MAE, the better the fitting effect of the algorithm. Furthermore, the experiment results made statistics of the running time of the algorithm. To make the experiment results more persuasive, 80% of each data set was taken as the training set and 20% was the test set. The standard 10-fold cross validation method has been adopted to the simulation experiment and all the results are the average of 10 independent runs. Table 2 shows the experiment results on the data set of power plant.

TABLE 2. The experiment results for data set of power plant before and after compression

Dataset	Method	RMSE	MAE	Time
H	After	**0.52**	**0.42**	**1.37**
	Before	0.60	0.48	14.06
M	After	**0.54**	**0.46**	**1.31**
	Before	0.69	0.56	13.48
L	After	**0.25**	**0.13**	**1.32**
	Before	0.38	0.22	13.57

As can be seen from Table 2, the compressed data set performed better on the twin support vector regression and the time efficiency of the algorithm was higher. The data compression has reduced the impact for the interference points in data on the fitting curve to a certain extent. In addition, the compression algorithm can filter the possible abnormal points existing in data. In terms of the running time, the data only contain the important data points after compression that the insignificant data points were eliminated; as a result, the running efficiency has been significantly improved, verifying the effectiveness of the proposed algorithm. This also illustrates from the side that it is imperative to compress the data in the process of mass data classification or regression analysis, which is not only efficient but also effective.

Therefore, both artificial dataset and power plant dataset show that the compression rate of the proposed algorithm is better than the traditional SDT algorithm. The proposed algorithm can also be better applied to traditional data mining work, not only can improve the accuracy of data mining, but also can greatly reduce the time of data mining. In particular, for the current era of big data, this work has a good basic research effect.

6. **Conclusions.** In this paper, a sparse representation data compression algorithm for power plant has been put forward. The core idea of the method is to update the locations of circular zone through the iteration of time so that the similar or overlapping data sets can be filtered and screened. During the filtering process, the maximum filtering value will be increased, avoiding the destruction of original data structure. The experiment results indicate that the compression ratio can be lower than 0.15 in terms of the real data of power plant and the data can better retain the features of original data. Therefore,

compared with the other popular compression methods, the algorithm in the paper can compress the data of power plant more effectively.

However, it should be pointed out that the zone limitation strategy is not an optimal choice. Therefore, improving zone limitation strategy by introducing other approaches with better global searching capabilities may be a solution we need to focus. Furthermore, combining the advantages of different compression algorithms to form new compression algorithm is also our future research directions. We hope these questions will be successfully addressed in the near future.

REFERENCES

[1] J. W. Hang and J. Pei, *Data Mining Concepts and Techniques*, Machinery Industry Press, 2012.

[2] X. An, Y. Shi and Y. Zang, Research on trend forecasting system utilizing big data network and information technology, *Journal of Physics: Conference Series*, vol.1982, no.1, 2021.

[3] A. Fachrizal, J. Michal, S. Tomasz et al., clustering methods for power quality measurements in virtual power plant, *Energies*, vol.14, no.18, pp.5902-5903, 2021.

[4] O. J. Khaleel, L. T. Khaiil, I. F. Basim et al., Developing an analytical model to predict the energy and exergy based performances of a coal-fired thermal power plant, *Case Studies in Thermal Engineering*, vol.28, 2021.

[5] G. Liang, W. Y. Chen and H. J. He, Research on safe and reliable data of power plant control system based on principal component analysis algorithm, *World Scientific Research Journal*, vol.7, no.4, 2021.

[6] H. J. Choi, C. W. Kim and D. Kwon, Data-driven fault diagnosis based on coal-fired power plant operating data, *Journal of Mechanical Science and Technology*, vol.34, pp.1-6, 2020.

[7] T. J. Desell, A. Z. Lyu, D. Stadem et al., Long term predictions of coal fired power plant data using evolved recurrent neural networks, *Automatisierungstechnik*, vol.68, no.2, pp.130-139, 2020.

[8] J. Michal, S. Tomasz, K. Dominika et al., A case study on data mining application in a virtual power plant: Cluster analysis of power quality measurements, *Energies*, vol.14, no.4, 2021.

[9] M.-W. Baek, M. K. Sim and J.-Y. Jung, Wind power generation prediction based on weather forecast data using deep neural networks, *ICIC Express Letters, Part B: Applications*, vol.11, no.9, pp.863-868, 2020.

[10] W. Bao and R. Zhou, Two-dimensional lifting wavelet compression algorithm for power plant vibration data, *Power Engineering*, vol.5, pp.732-735, 2007.

[11] S. N. Qu and L. Liu, Data compression algorithm for railway air interface monitoring based on waveform dictionary, *Application Research of Computers*, vol.37, pp.266-269, 2020.

[12] L. Struski, J. Tabor and P. Spurek, Lossy compression approach to subspace clustering, *Information Sciences*, vol.435, pp.161-183, 2018.

[13] M. A. Khan, J. W. Pierre, J. I. Word et al., Impacts of swinging door lossy compression of synchrophasor data, *International Journal of Electrical Power & Energy Systems*, vol.123, 2020.

[14] Y. J. Yang, X. U. Jiang, X. U. Shuai et al., Research on lossy compression algorithm in real-time database, *Computer Technology and Development*, 2012.

[15] A. Joseph and M. Abdallah, Robust IoT time series classification with data compression and deep learning, *Neurocomputing*, vol.398, pp.222-234, 2020.

[16] M. B. Begum and Y. Venkataramani, A new compression scheme for secure transmission, *International Journal of Automation and Computing*, vol.10, no.6, pp.578-586, 2013.

[17] D. A. Huffman, A method for the construction of minimum-redundancy codes, *Proc. of the IRE*, vol.40, no.9, pp.1098-1101, 1952.

[18] C. S. Wang and J. D. Wang, Data compression based on energy threshold and adaptive arithmetic coding, *Automation of Electric Power Systems*, vol.24, 2004.

[19] J. Ziv and A. Lempel, A universal algorithm for sequential data compression, *IEEE Trans. Information Theory*, vol.23, no.3, pp.337-343, 1977.

[20] D. Barman and M. B. Ahamed, Improved LZO compression technique using difference method, *International Journal of Innovative Technology and Exploring Engineering*, vol.9, no.5, pp.87-92, 2020.

[21] T. J. Kim, J. J. Yoo and J. W. Hong, Fast mode decision algorithm for scalable video coding based on luminance coded block pattern, *Optical Engineering*, vol.52, no.1, pp.17401-17401, 2013.

[22] J. K. Zhao, L. S. Mu, P. F. Zhu et al., In-network time-series data compression for electric Internet of Things, *Applied Mechanics and Materials*, vol.2111, pp.3213-3223, 2013.
[23] J. Ruan, Y. Shi and J. Yang, Forest fires burned area prediction based on support vector machines with feature selection, *ICIC Express Letters*, vol.5, no.8, pp.2597-2603, 2011.
[24] Jayadeva, R. Khemchandani and S. Chandra, Twin support vector machines for pattern classification, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.29, no.5, pp.905-910, 2007.
[25] S. F. Ding, F. L. Wu and Z. Z. Shi, Wavelet twin support vector machines, *Neural Computer & Application*, vol.25, pp.1241-1247, 2014.
[26] J. Z. Yu, S. F. Ding and F. Jian, Twin support vector machines based on rough sets, *International Journal of Digital Content Technology & Its Applications*, vol.6, no.20, pp.493-500, 2012.

Author Biography

Shuanzhu Sun received the B.Sc. degree in Production Process and Automation from North China Electric Power University, China, 1995; the M.Sc. degree in Power Plant Thermal Power Engineering from North China Electric Power University, China, 1998.

Shuanzhu Sun is currently a fourth-level employee of the Environmental Technology Center of Jiangsu Frontier Electric Technology Co., Ltd. He mainly researches on the development of energy-saving, emission-reduction monitoring technology in the power industry, big data analysis applications and related environmental protection products.

Bin Sun received the B.Sc. degree in Computer Science and Technology from Hohai University, China, 2003.

Bin Sun is currently working in the Environmental Technology Center of Jiangsu Frontier Electric Technology Co., Ltd. He mainly researches on the development and management of software projects such as Java, database, and big data analysis applications. He has a deep interest in artificial intelligence, anomaly detection and time series analysis.

Qixiang Wang received the B.Sc. degree in Computer Science and Technology from the University of Science and Technology of China, China, 1996.

Qixiang Wang is currently working in the Environmental Technology Center of Jiangsu Frontier Electric Technology Co., Ltd. He mainly researches on the network communications, database applications, big data analysis applications, software development and project management. He has a deep interest in artificial intelligence, anomaly detection and time series analysis.

Chunlei Zhou received the B.Sc. degree in Environmental Monitoring from Nanjing University of Science and Technology, China, 1995; the M.Sc. degree in Computer Application from Donghua University China, 1998.

Chunlei Zhou is currently working in the Environmental Technology Center of Jiangsu Frontier Electric Technology Co., Ltd. She mainly researches on the big data analysis about power plant, data mining technology application, related product research and development. She has a deep interest in artificial intelligence, fault diagnosis and abnormal detection.