Hindawi
Mobile Information Systems
Volume 2022, Article ID 5309556, 7 pages
https://doi.org/10.1155/2022/5309556

Research Article

Integration Mechanism of Heterogeneous Foreign Language Education Resources Based on Time Series Analysis in IIoT

Hongyue Jin

Jilin Normal University, Siping 136000, China

Correspondence should be addressed to Hongyue Jin; 1808023@jlnu.edu.cn

Received 6 December 2021; Accepted 31 December 2021; Published 21 January 2022

Academic Editor: Jianhui Lv

Industrial Internet of Things (IIoT) has attracted much attention from global researchers and has been applied into many fields, such as medical treatment, transportation, and education. This paper pays attention to an IIoT-oriented education problem and gives the corresponding solution. Heterogeneous educational resources have multisource target data, so it is necessary to integrate the repetitive data and data with the same attributes. However, due to the poor tracking effect of the model constructed by traditional methods, the mining technology loses a part of the data characteristics and affects the multisource foreign language education data integration. So this article studies the integration mechanism of foreign language heterogeneous educational resources based on time series analysis. The mechanism adopts a data cleaning and fusion method based on the time series similarity measurement. This method uses approximate symbol aggregation, European algorithm, and similar sequences with adjusted similarity weights to complete the data cleaning of foreign language heterogeneous educational resources. After that, it uses multiple heterogeneous data fusion algorithms to complete data integration. Experiments with foreign language education resources at all levels in a certain city show that the mechanism can detect abnormal data of foreign language education resources, fill in vacant data, reduce data redundancy, and integrate heterogeneous data. After the data are cleaned by multisource heterogeneous data fusion algorithm, the credibility of the measurement data is reflected, and the mean absolute percentage error is only 6.25%. The data quality is improved as a whole, and it provides reliable basic data for the application of foreign language education resources.

1. Introduction

In recent years, Industrial Internet of Things (IIoT) has been widely applied with the explosive increase of mobile devices and cloud platforms. In fact, during the process of building smart city, IIoT plays an important role, for example, improving the level of medical treatment, the efficiency of transportation, and the quality of education. Especially during the period of COVID-19, online education has become more are more significant. Therefore, the IIoT-oriented education resource allocation issue has attracted much attention from global researchers.

With the advancement of the construction of foreign language education information, most schools, education departments, and related institutions have established their own foreign language education and teaching resource systems, but such resource systems have been established due to different establishment periods and lack of unified technical specifications and education resources. The unified understanding of utilization has led to repeated investment in hardware facilities, repeated development of software platforms, and repeated construction of online courses, resulting in heterogeneity such as uneven distribution of digital foreign language education resources, low standardization, and difficulty in integration and sharing. The problem seriously hinders the effective use and reasonable distribution of digital foreign language education resources. In view of the current heterogeneity of foreign language digital education resources and the need for the rational allocation and effective use of educational resources in the industrialization of foreign language digital education, the research and realization of the integration of foreign language heterogeneous educational resources have become the current digital educational industrialization project. Only by establishing a reasonable development strategy for digital foreign language education resources can we better promote the construction of digital foreign language education resources, give full play to its maximum benefits, and serve the construction of educational information [1].

Data cleaning is a method used to detect and eliminate errors and inconsistencies in data [2]. In recent years, researchers have proposed a variety of data cleaning technologies to improve data quality, such as missing data attribution, object repeated detection, anomaly detection, logical error detection, and data inconsistency detection [3]. However, current data cleaning methods have high computational complexity and inaccurate detection of missing data. In 2017, the University of California, Riverside, Keogh, group [4] by based on changes in the field of feature representation, symbol, and piecewise linear representation to complete the feature extraction, but as a result of data in the process of feature extracting large dimensions, high computational complexity does not have scale invariance defects for multivariate time series data cleaning effect.

Similarity measurement technology is the basis of sequence data analysis. Sequence analysis can reflect the characteristics and relationships between data and judge data outliers based on the relationships mined, attracting a large number of scholars to conduct in-depth research. At present, in addition to the longest common substate distance and edit distance, similarity measurement methods mainly include Euclidean distance [5], dynamic time warping distance, and singular value decomposition and point distribution-based methods. The commonly used methods for data correction include the interpolation model [6], random replacement model, mean replacement model, and regression model [7]. When it comes to the identification of abnormal data, there may be a problem that the quoted "correct" sequence does not exist.

Data integration is essentially the collaborative processing of data from multiple parties to achieve the purpose of reducing redundancy, comprehensive complementation, and capturing collaborative information. This technology has become a research hotspot in the fields of data processing, target recognition, situation assessment, and intelligent decision making. In [8], Yu et al. studied multisensor data integration technology based on statistics and artificial intelligence (AI) methods; in [9], Lai et al. studied the organization and management of multisource heterogeneous data in mobile geographic information systems and established a multisource heterogeneous data fusion model; in [10], Premkumar and Ganesh combined wireless sensor network and data fusion technology and proposed a Kalman filter batch estimation fusion algorithm; in [11], Lasheng and Yiquang studied a massive multisource heterogeneous data fusion method in the Internet of Things environment and successfully applied it in the process of target positioning and tracking; in [12], Zhang et al. studied the intelligent maintenance decision-making architecture of the high-speed rail signal system based on heterogeneous data fusion, which improved the accuracy and effectiveness of decision making; in [13], Wen et al. studied many aspects of the digital mine construction process. Source heterogeneous data fusion technology ensures the safety, stability, and efficiency of the basic information platform in the construction of digital mines.

In view of this, in order to improve the quality of foreign language education resource data and support the large-scale collection and storage of data, this article focuses on the data cleaning and fusion in the construction of foreign language heterogeneous resource integration system and conducts a preliminary analysis of the characteristics of foreign language education multisource data. And a practical value-based foreign language education resource data cleaning and fusion algorithm based on time-series similarity measurement is proposed, and experiments have shown that it can achieve a better cleaning and fusion effect.

The rest paper is structured as follows. In Section 2, the integration of foreign language education resources is studied. In Section 3, the time series analysis-based integration mechanism of heterogeneous foreign language education resources is studied. The experimental results are presented in Section 4, and Section 5 concludes this paper.

2. The Integration of Foreign Language Education Resources

At present, the data of foreign language education resources present the characteristics of "two" (diversified data sources and data types) and "two" (high heterogeneity dimension and high overall value of data), and its greatest value lies in the realization of cross system and cross platform data exchange and sharing. The integrated application of big data in total foreign language education aims to break the "data island," establish the data governance system of foreign language education resources, and form the total data assets of foreign language education in the smart city ecosystem. The outstanding problems existing in the data integration of foreign language education resources are as follows: no unified data standard, data source is not clear, out of sync data exchange, data storage and scattered in disorder; this series of factors has resulted in unfavorable situations of low data quality, chaotic data flow, insufficient data sharing, and unsmooth data lifecycle management, which greatly restricts the height that foreign language education big data-assisted smart application terminals can achieve [14, 15].

The integration and application of foreign language resource education data focus on three aspects of "management + governance + application," and the key problems are mainly reflected in the following three aspects:

(1) Business data cannot effectively follow a unified data standard. Data standards regulate the consistency and accuracy standards of data used and exchanged within and outside regions at all levels, restrict the normative documents of data standards, and carry out data standardization control and data standard management organization to provide data for

foreign language education resources at all levels. The platform provides a unified data definition standard and logical model. However, due to the different construction ages and different structural levels of various platforms, they were not defined in accordance with a unified data standard at the initial stage of construction, which brought a lot of inconvenience to the exchange and sharing of foreign language education resource data. In response to such problems, we should first start with the top-level design of informatization and intelligent information services, formulate unified data standards, establish a scientific and standardized data application assessment and evaluation mechanism, and carry out transformations in stages and steps; source data carries out all aspects of data cleaning.

(2) The data source of foreign language education resources is not unique, and the data flow is unreasonable. The producer of the data must determine the focal point. The focal point is the uniqueness of the data source. The content of the data cannot be maintained by multiple systems at the same time; otherwise, the uniqueness and accuracy of the data source cannot be guaranteed. The flow of data is aimed at achieving exchange and sharing, and the public data platform (public data pool) completes the cross-business data interaction [16]. This type of problem requires the establishment of relevant organizational structures through administrative management methods, determining the authority for data generation, clarifying the data responsible unit, constructing a data flow relationship table, and providing a complete data flow for the data source connected to the system and the data interface released by the system. The unified combing of foreign language education resource data application requirements is managed and completed.

(3) The quality of business data is not high, and there are certain phenomena of "lack of data" and "wrong data." Data quality describes the applicability of the data, that is, the suitability of the data to meet the needs of users. Data quality measures data through multiple dimensions such as completeness, consistency, accuracy, timeliness, and legitimacy. In the business platform, data quality provides clean and structured data for it. It is a necessary prerequisite for the data platform to develop data products, provide data services, and play the value of big data. It is also a key factor in the management of foreign language education data assets at all levels and regions. Currently, data quality is generally not high in all levels and regions. On the one hand, it is necessary to improve data quality through in-depth data governance (analysis, correlation, cleaning, and exchange of multisource heterogeneous data), and on the other hand, it is necessary to establish data quality and improve the process and assessment system.

3. Integration Mechanism of Heterogeneous Foreign Language Education Resources Based on Time Series Analysis

In view of the lack of data and wrong data in the foreign language education resources described above, this paper proposes a data cleaning method based on time-series similarity measurement to detect abnormal data and fill in missing data in foreign language materials.

The data cleaning and fusion process is mainly divided into four steps: first, the approximate symbol aggregation algorithm is used to discretize and symbolize the foreign language resource data; second, the Euclidean distance algorithm is used to calculate the similarity between the symbol sequences; then, it is fitted according to the similar sequence. The curve of foreign language data completes the identification and correction of abnormal data and the filling of missing data; finally, the cleaned data are fused.

3.1. Approximate Symbol Aggregation Algorithm. In recent years, the symbolic aggregation approximation (SAX) algorithm is a new method of discretizing time series data. The basic idea of this method is to convert numerical time series data into discrete symbol sequences [17]. Through the specified mapping rules, the SAX algorithm can weaken the influence of abnormal and missing data in the time series on the local fluctuations and can also generate a smaller-sized symbol and nondigital sequences, which can improve further aggregation efficiency and strengthen the comparison of similarity in the later stage.

SAX is a sequence of equal-length partition based on the piecewise aggregate approximation (PAA). If the partition length is long, there may be a large internal difference, and the mean value is equal. The key point improvement method can be adopted to achieve the purpose. But the algorithm complexity is improved. In this paper, SAX is used to reduce a time series of arbitrary length n to a string of length N ($N < 26$), usually with English sentences of no less than 26 letters.

SAX first converts the data to PAA representation, reduces the time sequence from n dimension to N dimension, then maps all PAA coefficients to m equal probability intervals, and the last SAX symbolizes the PAA representation into a discrete string. The following is a brief execution process of SAX on the original time series $X = \{x_1, x_2, \ldots, x_n\}$.

(1) Performance of normalization processing: normalization is to convert the average value of each time series to 0 and the standard deviation to 1, which is expressed as $C = \{c_1, c_2, \ldots, c_n\}$. The i element is

$$C_i = \frac{x_i - \mu(X)}{\delta}\,(i = 1, 2, \ldots, n), \tag{1}$$

where μ is the average value of the original time series; δ is the standard deviation.

(2) The dimensionality of normalized sequence C is reduced by PAA to reduce the original time series

vector of n dimension to N dimension. In the process of dimensionality reduction, the N dimensional time series $C = \{C_1, C_2, \ldots, C_N\}$. The i element in C is calculated as

$$\begin{cases} \overline{C}_i = \dfrac{1}{t} \displaystyle\sum_{t(i-1)+1}^{ti} c_j \\ \\ j = t(i-1) \end{cases}, \tag{2}$$

where $\overline{C}_i$ is the mean value of the original time series vector divided into N segments; $t = n/N$ is called the compression rate; $1/t$ is the interval length of each segment.

After converting the time series set $\overline{C}$ into PAA, it is further converted into a discrete symbol form; that is, elements in the PAA representation form of the time series are mapped to equal probability symbols. Since the normalized time series has a highly characteristic Gaussian distribution, the "break point" β is determined by looking up the Gaussian distribution statistics table, thereby generating m equal sizes, that is, regions with the same probability distribution, where β is a series of the ordered list of values, and all β areas are $1/m$.

After querying and comparing the breakpoint β, the time series collection C is transformed into the string collection $\widehat{C}$, namely,

$$\widehat{C}_j = P_j, \text{if}, \quad \beta_{j-1} \leq \overline{C}_j \leq \beta_j, \tag{3}$$

where P is the alphabet; the j element of the N dimensional time series C is between β_{j-1} and β_j; the j element of the alphabet A can be expressed as the j element in the string $\widehat{C}$.

3.2. Similarity Measurement. Euclidean distance is one of the most widely used algorithms in similarity measurement. In the application process, the sequence to be compared is required to have the corresponding length and point, and the difference between the two sequences corresponds to each other [18]. The Euclidean distance can quickly calculate the similarity of SAX symbolic expressions with low computational complexity. The greater the distance between the SAX expressions of two foreign language education data sequences, the lower the similarity. Therefore, the similarity of the two foreign language education data time series curves is

$$S(Q,C) = \left[\sqrt{\sum_{i=1}^{n} \left(q_i - c_j\right)^2} \right]^{-1}, \tag{4}$$

where Q and C are the two time series, respectively; q_i is the i point of the Q sequence; c_j is the C point of the j sequence.

3.3. Similarity Curve Adjustment. After approximate aggregation of symbols and similarity measurement of the time series, ω similar time series A and similar time series SXA to be cleaned up are obtained, where X is the original series and ω is usually 30 time series in a month. The similar time series A is adjusted by the weighted adjustment method (fitted curve algorithm) to obtain the reference curve $\widehat{X}$ relative to the original time series X. If there is a missing value in the original time series, it is filled with the value of the corresponding point in the reference curve.

Judge whether a word in the data is abnormal by comparing within the reference curve, which is calculated by weighting similar letters in all foreign language education resources. This article uses an improved maximum threshold method to determine whether a foreign language word is abnormal data, and this method uses a more accurate weighted average of similar time series to calculate, and calculates x_k of the threshold δ_k, that is,

$$\delta_k = \max\left(A - \widehat{x}_k\right). \tag{5}$$

If x_k does not meet the following criteria, it is considered abnormal data.

$$\begin{cases} x_k > \widehat{x}_k - \delta_k, \\ x_k < \widehat{x}_k + \delta_k. \end{cases} \tag{6}$$

3.4. Integration of Foreign Language Heterogeneous Educational Resource Data. A foreign language multisource heterogeneous data integration focuses on computing structured and comparable foreign language heterogeneous data, with the goal of improving data quality and obtaining more significant data characteristics. Kalman filter is an efficient recursive filter, which uses a series of data obtained from the data measurement process to estimate the state vector of the dynamic system, which is more effective for heterogeneous structured data [19]. In this paper, the concept of information pair is introduced into the Kalman filter algorithm, that is, the distributed Kalman algorithm. The cleaned data are exchanged and merged with the adjacent data sequence. The information matrices X1 and X2 used are respectively

$$\begin{cases} \mathbf{M}^i_{k|k} = \left(\mathbf{P}^i_{k|k}\right)^{-1}, \\ \mathbf{A}^i_{k|k} = \mathbf{M}^i_{k|k} \widehat{x}^i_{k|k}, \end{cases} \tag{7}$$

where $\mathbf{P}_{k|k}$ is the posterior estimated covariance matrix at time k and x^i is the estimated state value at time k. The recursive form of the distributed Kalman filter is

$$\begin{cases} \mathbf{M}^i_{k|k} = \mathbf{M}^i_{k-1|k} + \left(\mathbf{Q}^i\right)^{\mathrm{T}} \left(\mathbf{R}^i\right)^{-1} \mathbf{Q}^i_k, \\ \mathbf{A}^i_{k|k} = \mathbf{A}^i_{k-1|k} + \left(\mathbf{Q}^i\right)^{\mathrm{T}} \left(\mathbf{R}^i\right)^{-1} \mathbf{Q}^i_k, \end{cases} \tag{8}$$

where Q and R are the covariance matrices of system noise and observation noise, respectively. In order to improve the accuracy of local fusion, the total dataset is N, and each series P_i can send its local posterior covariance k/k to the j adjacent series j and perform data fusion with local posterior covariance matrix P_j of j, and the fusion calculation is, respectively,

Table 1: Comparison of three data fusion methods.

Test dataset	Data residual rate (%)		
	Unfused data	Kalman algorithm	This paper
60	100	38	10
120	100	41	12
180	100	47	15
240	100	53	18
300	100	57	21
360	100	63	23
. . .	. . .	. . .	. . .
1440	100	95	36

$$\begin{cases} \mathbf{M}_{k|k}^{i,j} = \Pi_{i,j}\mathbf{M}_{k|k}^{i} + \sum\limits_{i,j\in N} \mathbf{M}_{k|k}^{i}, \\ \mathbf{A}_{k|k}^{i,j} = \Pi_{i,j}\mathbf{A}_{k|k}^{i} + \sum\limits_{i,j\in N} \mathbf{A}_{k|k}^{i}, \end{cases} \tag{9}$$

where $\Pi_{i,j}$ is the combined weight and positive value, satisfying any node, $\Pi_{i,j}$ can be obtained from the following formula:

$$\Pi_{i,j} + \sum_{i,j\in N} \Pi_{i,j} = 1. \tag{10}$$

The multisource heterogeneous data fusion algorithm flow is as follows:

(1) Initialize data, and time series $i, j \in N, N$ are data space

(2) Observe the status of the integrated data and update the information pair of the time series X1 through equation (8)

(3) The information pair of the time series i is transmitted to the adjacent time series j. If the time series j data are completed and safe, the time series j receives the information pair sent from the time series i

(4) Through formulas (9) and (10), the local data pair and the information pair of adjacent data are used for data fusion to obtain the fused information pair

(5) Update the local filter value

(6) Return to Step (2)

4. Experimental Results and Analysis

Section 3 describes the SAX method, similarity measure method, and data cleaning and data fusion method used in this experiment, respectively. In this paper, foreign language resources at all levels in a city were selected as the experimental dataset, and data cleaning and integration experiments were carried out on the experimental dataset. The experimental data consisted of 16 series of three data types mentioned above, with a collection interval of 1 min and a sampling frequency of 10 kHz for multisource data.

In order to prove the effectiveness of the multisource heterogeneous data fusion algorithm proposed in this paper, in addition to the traditional Kalman filtering algorithm [20], this paper also compares the case of not adding a filtering algorithm and draws corresponding conclusions. Table 1 shows the ratio of effective data obtained after data fusion using three data fusion methods. The three methods are the distributed Kalman filter algorithm without data fusion, the Kalman filter algorithm, and the edge calculation. The Kalman algorithm of each fused iteration of the dataset and the neighbor nodes are smaller, and the effect is better. With the rapid increase in the size of the dataset, the integration of data efficiency declines rapidly. This is because there are many series, large resource consumption, larger dataset, low data fusion ratio, and high data redundancy. The multisource heterogeneous data fusion algorithm can effectively reduce redundant data information so as to get closer to the actual effective data value. Compared with the unfused data method, it reduces a lot of resource consumption and can align different data more effectively and resolve feature conflicts between data.

Table 2: Results of data cleaning of heterogeneous foreign language education resources.

Algorithm	Wrong comparison	
	MAPE (%)	RMSE
Gaussian filtering	18.23	1998.04
Wavelet threshold	12.34	1634.98
This paper	6.25	1139.42

Table 3: Comparison of data integration results of heterogeneous foreign language education resources.

Algorithm	Wrong comparison		
	RMSE	MARE	t test
Kalman algorithm	2.76	0.028	2.97
This paper	2.12	0.013	2.16

Select the data sequence with no more than 1% gap as the experimental set, and set the data to be empty according to a random ratio, then use the method in this article to clean the data, compare the gap prediction value with the original value, and calculate the average absolute percentage error and mean square root error and standard root mean square error. Mean absolute percentage error (MAPE) is the average absolute value of the relative error ratio, which can reflect the credibility of the measurement data, namely,

$$E_{\text{MAPE}} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%. \tag{11}$$

In addition to the traditional Gaussian filter algorithm [21], this paper also compares the data cleaning of wavelet threshold value, as shown in Table 2. The average absolute percentage error deviation of the algorithm proposed in this paper is 6.25%, which shows that the prediction results in this paper are relatively accurate, and the root mean square error is within the numerical range of the standard point, which meets the requirements.

At the same time, the error analysis of the integrated foreign language education data indicators is carried out, and the accuracy of the algorithm in this paper is evaluated according to the RMSE, MAPE, and t test [22]. Table 3 reflects the RMSE, MAPE, and t-test results of the two algorithms from a quantitative perspective. From the analysis of the data results, the error of the algorithm in this paper tends to be flat or reduced, which proves that the algorithm in this paper has good performance and can meet the accuracy requirements.

5. Conclusions

In the era of AI education, building a reusable and sharable educational data model is one of the urgent problems to be solved in the development of foreign language education. The model needs to standardize the multisource and heterogeneous foreign language education data in the AI education environment, so as to achieve a high degree of sharing of heterogeneous foreign language data. In this paper, an IIoT-oriented environment is considered, and a method based on time-series similarity measurement is proposed to clean and integrate heterogeneous data of foreign language education resources. SAX by piecewise reduce the dimension of time series data, so as to achieve the aim of reducing noise and is calculated using Euclidean distance similarity measurement method, and in smaller time complexity to find a similar set of time series, using the maximum threshold method to detect outliers, the reference curve is obtained by the weighted adjustment method, according to the reference curve filling vacant values. It is more accurate than the traditional method using the maximum threshold of the average value of similar days. The data after cleaning adopt the multisource heterogeneous data fusion algorithm. Through data fusion between adjacent series, redundant data can be better fused to ensure data quality and provide more practical value for the high sharing of heterogeneous foreign language education resources.

This research still has a lot of work that needs to be further explored, such as the way to increase the integration of different types of foreign language education resources and to integrate different types of foreign language education resources such as online teaching, online vocational training, online examination, and synchronized online education guidance. The same platform is the future research and development direction of the foreign language education resource integration platform. I believe that foreign language education can be improved through different angles and ways to improve its fairness, universality, and sharing. This is a beautiful goal that we all have been working on together.

Data Availability

All data used to support the findings of the study are included within the article.

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

This paper was supported by the 13th Five-Year Social Science Research Project of the Education Department of Jilin Province (no. JJKH20200438SK) and the Social Science Fund Project of Jilin Province (no. 2020C113).

References

[1] P. Xia, "Application scenario of artificial intelligence technology in higher education," in *Proceedings of the 2019 International Conference on Applications and Techniques in Cyber Intelligence*, Huainan, China, June 2019.

[2] T. Odia, S. Misra, and A. Adewumi, "Evaluation of hadoop/mapreduce framework migration tools," in *Proceedings of the Asia-Pacific World congress on Computer Science and Engineering*, Nadi, Fiji, November 2015.

[3] Q. Ma, S. Tian, J. Wei, J. Wang, and W. W. Y. Ng, "Attention-based spatio-temporal dependence learning network," *Information Sciences*, vol. 503, pp. 92–108, 2019.

[4] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate LSTM-FCNs for time series classification," *Neural Networks*, vol. 116, pp. 237–245, 2019.

[5] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, no. 99, pp. 1662–1669, 2017.

[6] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.

[7] Y. Gu, W. Lu, L. Qin, M. Li, and Z. Shao, "Short-term prediction of lane-level traffic speeds: a fusion deep learning model," *Transportation Research Part C: Emerging Technologies*, vol. 106, pp. 1–16, 2019.

[8] H. Yu, N. Rao, and I. S. Dhillon, "Temporal regularized matrix factorization for high-dimensional time series prediction," in *Proceedings of the Neural Information Processing Systems*, Barcelona, Spain, December 2016.

[9] G. Lai, W. Chang, Y. Yang, and H. Liu, "Modeling long and short term temporal patterns with deep neural networks," in *Proceedings of the Fourty First International ACM sigir conference on research and development in information retrieval*, Ann Arbor MI USA, July 2018.

[10] M. S. Premkumar and S. H. Ganesh, "A median based external initial centroid selection method for k-means clustering," in *Proceedings of the World Congress on Computing and Communication Technologies*, pp. 143–146, Tiruchirappalli, India, February 2017.

[11] C. Lasheng and L. Yuqiang, "Improved Initial clustering center selection algorithm for K-means," in *Proceedings of the 2017 Signal Processing: Algorithms, Architectures, Arrangements, and Applications(SPA)*, pp. 275–279, Poznan, Poland, September 2017.

[12] R. Zhang, J. Huang, and T. Kumar, "Preventive leak detection for high pressure gas transmission networks," in *Proceedings of the AAAI Workshop*, pp. 64–70, San Francisco, C A, U S A, February 2017.

[13] Q. Wen, J. Gao, X. Song, L. Sun, H. Xu, and S. Zhu, "RobustSTL: a robust seasonal-trend decomposition algorithm for long time series," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, H I, USA, February 2019.

[14] D. Rajan and J. J. Thiagarajan, "A generative modeling approach to limited channel ECG classification," in *Proceedings*

of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2571–2574, Honolulu, H I, USA, July 2018.

[15] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, "Boosting deep learning risk prediction with generative adversarial networks for electronic health records," in *Proceedings of the IEEE International Conference on Data Mining*, New Orleans, LA, USA, November 2017.

[16] H. Wang, C. Luo, and X. Wang, "Synchronization and identification of nonlinear systems by using a novel self-evolving interval type-2 fuzzy LSTM-neural network," *Engineering Applications of Artificial Intelligence*, vol. 81, no. 7, pp. 79–93, 2019.

[17] S. Lin and G. C. Runger, "GCRNN: group-constrained convolutional recurrent neural network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4709–4718, 2018.

[18] Z. Gong, H. Chen, B. Yuan, and X. Yao, "Multiobjective learning in the model space for time series classification," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 918–932, 2019.

[19] J. Wang, T. Sun, B. Liu, Y. Cao, and H. Zhu, "CLVSA: a convolutional LSTM based variational sequence-to-sequence model with attention for predicting trends of financial markets," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Macao China, August 2019.

[20] F. Amato, M. Laib, F. Guignard, and M. F. Kanevski, "Analysis of air pollution time series using complexity-invariant distance and information measures," *Physica A: Statistical Mechanics and Its Applications*, vol. 547, Article ID 124391, 2020.

[21] X. Zhao, M. Ji, N. Zhang, and P. Shang, "permutation transition entropy: measuring the dynamical complexity of financial time series," *Chaos Solitons and Fractals*, vol. 139, Article ID 109962, 2020.

[22] X. Guo, Y. Wang, N. Zhou, and X. Zhu, "Optimal weighted two-sample t-test with partially paired data in a unified framework," *Journal of Applied Statistics*, vol. 48, no. 6, pp. 961–976, 2021.