Hindawi
Discrete Dynamics in Nature and Society
Volume 2024, Article ID 9878510, 1 page
https://doi.org/10.1155/2024/9878510

Retraction

Retracted: Time Series Symbolization Method for the Data Mining K-Means Algorithm

Discrete Dynamics in Nature and Society

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] G. Wang, "Time Series Symbolization Method for the Data Mining K-Means Algorithm," *Discrete Dynamics in Nature and Society*, vol. 2023, Article ID 5365673, 11 pages, 2023.

Hindawi
Discrete Dynamics in Nature and Society
Volume 2023, Article ID 5365673, 11 pages
https://doi.org/10.1155/2023/5365673

Research Article

Time Series Symbolization Method for the Data Mining K-Means Algorithm

Guisheng Wang

Tongling University, Tongling 244000, Anhui, China

Correspondence should be addressed to Guisheng Wang; wgs1869@tlu.edu.cn

Received 10 December 2021; Revised 10 January 2022; Accepted 24 November 2022; Published 15 April 2023

Academic Editor: Ahmed Farouk

Time series is a data type frequently encountered in data analysis. With the current depth and breadth of the data and the improvement in computer processing capabilities, the dimensionality and the complexity of time series are getting higher and higher. Time series symbolization is to cluster and assign complex and lengthy time series in the form of symbols to achieve the purpose of reducing the dimensionality of the sequence or making the sequence easier to process. Considering the excellent performance of the K-means algorithm in data mining and processing, as well as in the allocation algorithm for clustering, we plan to develop a simple method for the symbolization of time series for the K-means algorithm and hope that this method can realize the high-dimensional time series dimensionality reduction, processing of the special points in time series, and so on. Based on this, this article proposes an improved sans algorithm based on the K-means algorithm and discusses the representation method and the data processing of time series symbolization. Experimental results show that this method can effectively reduce the dimensionality of high-dimensional time series. After dimensionality reduction, the information retention rate contained in the elevation of the sequence can reach more than 90%, which is very effective for the detection of outliers in low-dimensional sequences.

1. Introduction

Mobile Internet is the current trend of social development. Driven by the rapid development of information technology, in the face of increasing time series data, data storage and processor functions have increased, and data storage and data processing capabilities have been continuously enhanced. Data mining is a technology that crosses multiple disciplines. It effectively analyzes and discovers the knowledge in these data, observes the data generation mechanism, that is, analyzes intuitive data, and then makes correct decisions to avoid risks. However, data mining exploration relies on practical applications to advance in many aspects. In the past, only simple single-table data could be processed. After development, it can automatically discover multiple modes of massive data and extract useful information and knowledge from the sequence facade. The extracted information can be divided into intuitive information and intrinsic information, and the reliability of the discovered knowledge is relatively high. Data mining tools for specific fields are also relatively specific, but can complete reliable data extraction, which also promotes the continuous maturity of data analysis technology, thereby improving the level of decision-making. At the same time, big data analysis is closely related to the processing of time series data.

Time series come from various application fields in real life and can reflect the change characteristics of a certain attribute value of an object in time sequence. This type of data is called time series. The collection of time series is easily affected by many external factors, but data mining focuses on the combination of practice and theory. For such a large amount of data, there is a huge value. In these data, a large amount of useful information that can reflect the characteristics of the system is contained, so time series forecasting has an important practical application value. Time series forecasting methods have experienced a development process from linear models to nonlinear models, and time series

data mining is widely used in fields such as finance. Cluster analysis is one of the important techniques of data mining, and its uncertainty is expressed in numerical values. This method requires an overall comparison of the global data, breaks through the normal distribution assumptions of traditional algorithms for datasets, fully takes into account the particularity of data and needs, and discovers time series patterns and converts them into easy-to-understand knowledge. It is essential to give full play to the value of information. Due to the increase in sequence-related research studies, it has attracted more and more attention from the relevant researchers from all over the world. In particular, the demand for in-depth processing of multimedia and other information is increasing. Clustering algorithms are more effective for data mining and processing massive data. Nowadays, the time sequence in life is becoming more and more multidimensional and complicated. If the analysis of the time sequence is not symbolized, the algorithm will become complicated and will take up a lot of processing time. Therefore, the current research study on the symbolization of time series is the study of algorithm simplification, which is the saving of time and resources.

It is usually unrealistic to analyze the original unprocessed time series because the original time series does not have strong laws, and it is difficult to extract effective information from the chaotic sequence. In the environment of big data, data mining work tends to be important. In response to this, many scholars at home and abroad have performed relevant research studies. Because the K-means algorithm is widely used, there are also many studies on it at home and abroad. Xia et al. introduced the different types of clustering algorithms and introduced the classic K-means algorithm and the canopy algorithm in detail. Then, combined with the map-reduce computing model and the spark cloud computing framework, after using the canopy algorithm to optimize the initial value of the K-means algorithm, the parallel canopy-K-means algorithm is introduced. They proposed a parallel adaptive canopy-K-means algorithm, which can be used in a cloud computing framework to adaptively determine the distance threshold parameter T2 based on the statistical methods. The experimental results show that the proposed method is effective [1]. Yang et al. redefined the density of points based on the number of their neighbors and the distance between the points and their neighbors. On this basis, they proposed an initial cluster center selection algorithm that can dynamically adjust the weight parameters. The adaptability of the algorithm to datasets with various characteristics has been proved [2]. Jing and Wang believed that with the widespread use of social software, there are more and more tags-related research studies and applications. Based on the randomness and personalization of user tags, in order to better compare the clustering results, they proposed the clustering corresponding results matrix (CCR matrix), which is expected to become an effective tool to capture the evolution of the social tag system [3]. Time series research studies are also very hot topic of discussion. Osmanoglu et al. observed that time series analysis is applied to interferometric phase measurement. When the observed motion is greater than one-half of the radar wavelength, it will circle. However, no single algorithm can provide the best results in all situations. Since time series analysis of InSAR data is used in various applications with different characteristics, each algorithm has inherent unique advantages and disadvantages. They proposed several algorithms developed for time series analysis of InSAR data, using a set of sample results to measure the rate of subsidence in Mexico City [4]. Zhu and others believed that phase picking is a key step in microseismic data analysis. However, due to the lack of relevant methods for S phase picking, they proposed a method called time series segmentation clustering (TSSC), which is based on the K-means algorithm to select the *S* phase. After experiments, statistical analysis shows that this method is feasible compared with the results of manual picking [5]. Deklel et al. believed that symbolic regression is usually performed using evolutionary algorithms such as genetic programming (GP). In order to build a symbolic model from examples, they proposed a new symbolic regression method based on the artificial neural networks. Experiments show that although this idea is universal and can be used for general symbolic purposes, it is only applicable to symbolic regression in Boolean domains [6]. In order to provide a new spatial clustering process for time series data, Rasidah Ali and Ku-Mahamud proposed a clustering algorithm that introduced data transformation, using X-means data splitting to study the spatial homogeneity of time series rainfall data. The results show that data conversion using X-means data splitting in hierarchical clustering is better than other conversion techniques and is more consistent between training and test datasets based on similarity measures [7]. Amir et al. proposed a new rule-based automatic method based on crop phenology. For research purposes, Sentinel-2 data with a spatial resolution of 10 meters obtained in the red and near-infrared bands during the rice growing season in three regions of Iran were used. Experimental results show that although the rice fields have extensive intraclass temporal phenological variation, the algorithm performs well in detecting them. For Marvdasht, Dargaz, and Qazvin, the obtained kappa coefficients are 0.73, 0.94, and 0.70, respectively [8].

This article proposes a K-means clustering method, sans algorithm, and anomalous subsequence search algorithm based on the time series symbolization problem, followed by optimization. The research study on this aspect of the predecessors is also very thorough, and they have their own understanding of the problem of time series symbolization. Compared with the predecessors, this article focuses on the following innovations: (1) For data mining technology, many scholars have their own algorithm research studies and a large number of statistical software research studies, but this article discusses the role of data mining on the symbolization of time series, which is rarely performed by predecessors. (2) For the K-means clustering algorithm, most scholars use it for data analysis and data statistics, and there are many optimization algorithms based on the K-means algorithm, but most of them are also based on data analysis. This article is the courage to optimize the related optimization of time series symbolization. (3) For the

symbolization of time series, this article uses computer binary to assign values, which can make the expression more brief, and is suitable for many statistical software, and can be analyzed on multiple platforms.

2. Time Series Symbolization Method Based on the K-Means Algorithm

2.1. Data Mining K-Means Algorithm. The K-means algorithm is a basic and most widely used division method among clustering analysis methods. It is a method of discovering clusters and cluster centers in unclassified labeled data [9]. Its main advantage is that the algorithm is simple and fast. If the resulting clusters are dense and the difference between clusters is obvious, it works best [10]. For processing large datasets, the algorithm is relatively scalable and efficient. They are composed of n feature attributes, among which the numerical feature dissimilarity is measured by Euclidean distance, as shown in the following equation:

$$\sin\left(a_i, a_j\right) = \sqrt{\sum_{k=1}^{q} \left|h_{ki} - h_{kj}\right|^2}. \tag{1}$$

Generally, the following cost function is minimized as the clustering criterion [11], where β is the clustering criterion, and the formula is as follows:

$$\beta = \sum_{j-1}^{a} \sum_{i-1}^{b} y_i d\left(x_j, Q\right). \tag{2}$$

Supposing M and N are the difference degree function, then the formulas are

$$d(M, N) = \sum_{j-1}^{a} \sigma\left(m_i, m_j\right), \tag{3}$$

$$\sigma\left(m_i, m_j\right) = 1 - \begin{cases} 0, & \left(m_i = m_j\right), \\ 1, & \left(m_i \neq m_j\right). \end{cases} \tag{4}$$

This left-different degree definition assigns the same importance to all possible values of an attribute [12].

The data matrix is represented by the following formula:

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}. \tag{5}$$

The dissimilarity matrix is shown by the equation that follows:

$$\begin{bmatrix} 0 & & \\ c(2,1) & \ddots & \\ c(n,1) & c(n,2) & 0 \end{bmatrix}. \tag{6}$$

The Euclidean distance formula is represented as

$$d(a,b) = \sqrt{\left(a_{11} - b_{11}\right)^2 + \left(a_{12} - b_{12}\right)^2 + \cdots \left(a_{nn} - b_{nn}\right)^2}. \tag{7}$$

The Manhattan distance formula is as follows:

$$d(a,b) = \left(a_{11} - b_{11}\right)^2 + \left(a_{12} - b_{12}\right)^2 + \cdots \left(a_{nn} - b_{nn}\right)^2. \tag{8}$$

The similarity is shown by the equation as

$$x(a,b) = \frac{A + B}{W + B + A + Z}. \tag{9}$$

The Jaccard coefficient represents the similarity between the two data objects [13], as shown by the following equation:

$$x(a,b) = \frac{A + B}{W + B + A}. \tag{10}$$

Ensure that the number of states in each sequence variable is the same [13], as shown by the following equation:

$$Z = \frac{R - 1}{M - 1}. \tag{11}$$

The error sum-of-squares criterion is shown by the equation as follows:

$$J_A = \sum_{a=1}^{C} \sum_{b=1}^{M} \left\|x_a - m_b\right\|^2, \tag{12}$$

wherem_{b}represents the mean value of the samplem_{i}, which is shown by the formula that follows:

$$m_b = \frac{1}{n} \sum_{a=1}^{b} x_a. \tag{13}$$

The weighted average squared distance sum criterion is represented as

$$J_b = \sum_{b=1}^{C} P_t S_t. \tag{14}$$

The average squared distance within a class between data samples is shown by the following equation:

$$S_t = \frac{2}{n(n-1)} \sum_{a \in A} \sum_{b \in B} \|X - x\|^2. \tag{15}$$

The distance between classes and the criterion is to describe the distribution of the distance between the different classes [14]. Specifically, there are two types of general interclass distance and criterion Jb1 and weighted interclass distance and criterion Jb2, as shown by the following formulas:

$$J_m = \sum_{i=1}^{c} \left(m_i - m\right)^T \left(m_i - m\right), \tag{16}$$

$$J_n = \sum_{i=1}^{c} P_i \left(n_i - n\right)^T \left(n_i - n\right). \tag{17}$$

The flowchart of the K-means algorithm is shown in Figure 1.

2.2. Time Series. Time series often appear in our daily lives. It can be said to be a relatively common mathematical model [15], but many people do not pay special attention to it. In mathematics, it is generally used to give time to a single column of data. Generally speaking, the data those change with time are recorded in a chronological order, which is referred to as the time series. The values of these variables change over time, for example, recording daily air quality, recording daily temperature, daily humidity, and so on, or the length of each hour of the day, the traffic flow of each hour, and so on. These are relatively simple time series that can be counted, but because the amount of statistical data is generally very large now [16], the amount of data is also increasing, and the time is not a fixed hour or one day, which causes the dimensionality of the time series to become higher, and the data becomes more complicated, and it is no longer a single simple sequence [17]. What we are most concerned about is actually the extractable information contained in the sequence, and this is also because the amount of data is large, and we cannot analyze the information contained and hidden one by one, so various methods are used to analyze the trend and extract the most important information [18].

For the time series, we commonly use the ARIMA model and the AR and MA models, which are commonly used models for analyzing time series in mathematics. Since this article focuses on the study of the symbolization of time series, it will not go into detail and will only explain some important characteristics of the time series [19].

(1) Stationarity

For a time series, stationarity is one of its important characteristics. For a nonstationary series, it is necessary to make the series stationary to obtain sufficient information from its trend [20]. The general test of stationarity is through its autocorrelation graph and unit root test. These two are used more and can be easily implemented in various statistical software. For the autocorrelation graph, if the curve has a downward trend and a slow tailing trend, then it is nonstationary, and a progressive difference is needed to make the sequence stable [21].

(2) Differential processing

For nonstationary series, it is generally differentiated to make it stationary. Differences are divided into period-by-period differences and seasonal differences [22]. It should be noted that the sequence cannot be differentiated, because after the difference, although the sequence is stable, the information in the sequence will also be lost, causing the sequence to be invalid and there will be no trend. For a sequence without any trend, no matter what analysis is performed, the information cannot be extracted [23].

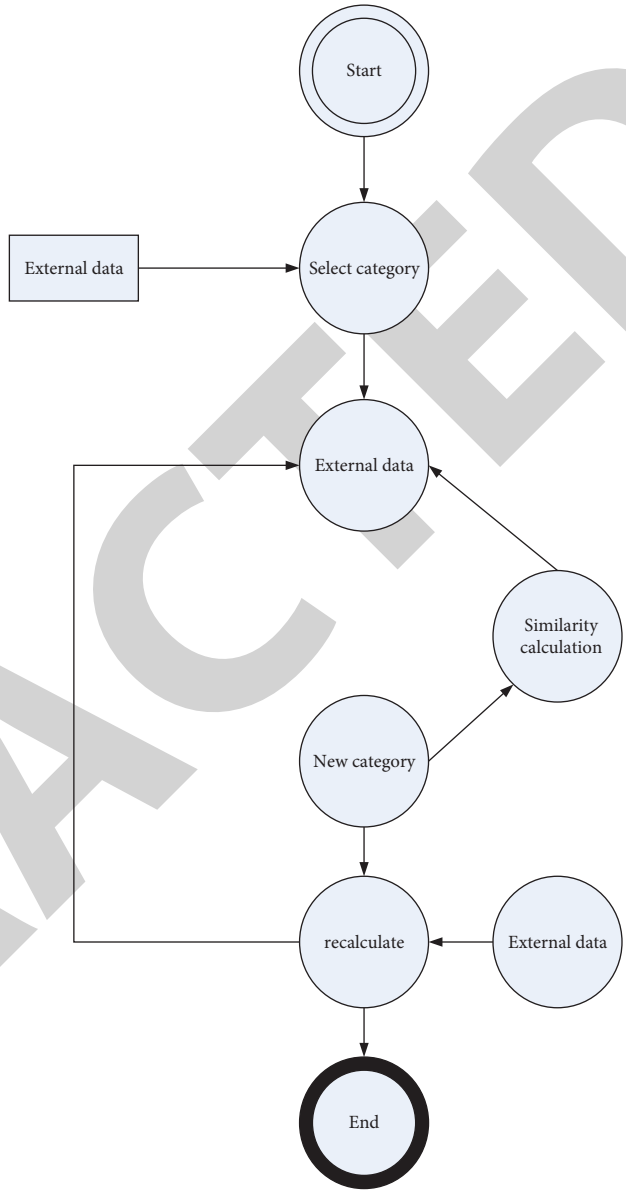

Figure 1: Algorithm flow chart.

(3) Periodicity

For time series, one month, one season, one year, and so on are usually used as a cycle, so the periodicity should be considered when analyzing it. We cannot blindly look at the trend, especially when predicting the trend of the sequence, we must take into account the influence of cyclicality [24].

(4) Model order

After the preprocessing of the sequence is completed, the order of the model must be determined, and the appropriate order and model can maximize the trend of the sequence, so as to better extract effective information.

2.3. Symbolization. There are not much research studies on the symbolization of time series because the concept of

symbolization was originally in aesthetics and life sciences, and because of the powerful expressive power and simple table elements of symbols [25], it has now been integrated into social life. The symbol is a very representative symbol, which is used in mathematics as a specific symbol. There are actually many similar symbols, such as our common Roman symbols and Latin symbols, which all represent a strong algorithmic meaning. In mathematics, the expression of numbers alone is far from enough and the things that numbers represent are also limited, but these are combined with specific symbols, such as pi, e, and other common formulas. With the in-depth study of time series, there are more and more symbolic expressions [26]. This article discusses the role of time series symbolization, which is different from the general role of symbols.

3. Time Series Symbolization Experiment Based on K-Means

Due to the large amount of time series data and high dimensionality, we cannot directly perform data mining on the original data. Therefore, how to effectively reduce the dimensionality and representation of the original time series data needs to be considered. The goal here is to find a representation method that can effectively reduce the dimensionality of the data and can minimize the loss of information, and on this basis, we conduct an effective and rapid mining and analysis of time series data. The symbolic representation method is an important representation method in the dimensionality reduction representation of time series and is widely used, not only because of its simplicity, understanding, and efficiency, but also because the use of these symbolization algorithms can refer to related algorithms in the fields of text processing, information retrieval, and biology to process the symbolized data [27].

3.1. The Slope-Based Symbolic Representation Algorithm Sans. The basic idea of the SANS algorithm is to first use a sliding window of length m to obtain $n-m+1$ subsequences of length m on a time series T of length n and symbolize each subsequence. According to the size of the parameter segment number c, the subsequence is divided into c segments. For each segment, it is represented by the discretization symbol of the local trend information, and the subsequence is represented by a string of length c, thus completing the symbolization process. In this article, the symbolization based on sans' is used more for the comparison with the optimization based on the K-means algorithm.

Therefore, the SANS algorithm here discretizes the slope value according to the distribution of specific data to ensure good results on different data and to make the mapped symbols more reasonable. Two different ways are used in the discretization method, which are explained as follows.

In Figure 2, the left side is the discretization method of equal size and the right side is the discretization method of equal probability. Due to the difference between the two methods, the discrete results are also different. It can be seen that in a graph of equal size, the split point is set as a breakpoint, and the obtained symbol distance represents the difference of local trends in different time series. In an equal probability graph, the boundary of each point set is taken as the breakpoint, and three different sizes are obtained. However, there are collections with the same number of datasets. In this kind of a collection, the differences between the datasets are not big, and the global trend of changes is more considered. Therefore, for data discretization, various methods have their own characteristics, and one can choose according to their characteristics and needs of the algorithm.

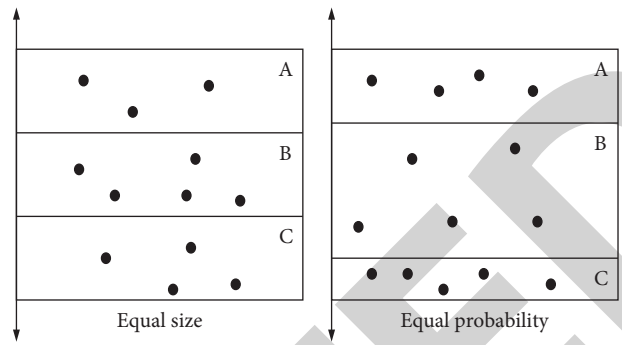

Figure 2: Discretization effects of different methods.

3.2. Similarity Measurement of Sans Algorithm. To divide the area with equal probability, the following discretization methods need to be discussed.

In Table 1, it represents the distance between equal probability symbols, that is, a, b, c, and d, and the probability of falling in the dataset is equal. As can be seen from Table 1, the maximum distance between a and d is 8.9, the minimum distance between a and b is 2.8, and in one-dimensional space, the distance between a and b is the same as between b and a, and the others are similar.

In Table 2, it is the distance between symbols of equal size, that is, a, b, c, and d. The probability of falling in each dataset is not equal, and the size of each dataset is equal. It can be seen from Table 2 that the maximum distance between a and d is 9.9 and the minimum distance between c and b is 2.5. It is not the same as the equal probability distance. It can be found that in the equal-size discretization method, the distance between the data will be larger than the equal probability.

From the analysis of Tables 1 and 2, it can be seen that in the two distance measurement methods, the distance calculation between symbols is determined based on the difference between the mean values of the slope values represented by the divided symbols, and the obtained symbol distance represents the difference of the local trend of different time series.

In Figure 3, it can be seen that as t increases, the dataset generally shows an exponential decline, but there is an inflection point when t is 1200. At this time, the size of the dataset is stable and will not float. It will not show an upward trend up to 2400, but the upward trend does not last long, reaching the apex of a small upward trend at 2500. After that, it continued to decline to the lowest point. Using the nearest

Table 1: Distance between equal probability symbols.

	a	*b*	*c*	*d*
A	0	2.8	5.8	8.9
B	2.8	0	3	6.1
C	5.8	3	0	5.9
D	8.9	6.1	5.9	0

Table 2: Distance between equal-sized symbols.

	a	*b*	*c*	*d*
A	0	3.4	7.4	9.9
B	3.4	0	4	6.5
C	7.4	4	0	2.5
D	9.9	6.5	2.5	0

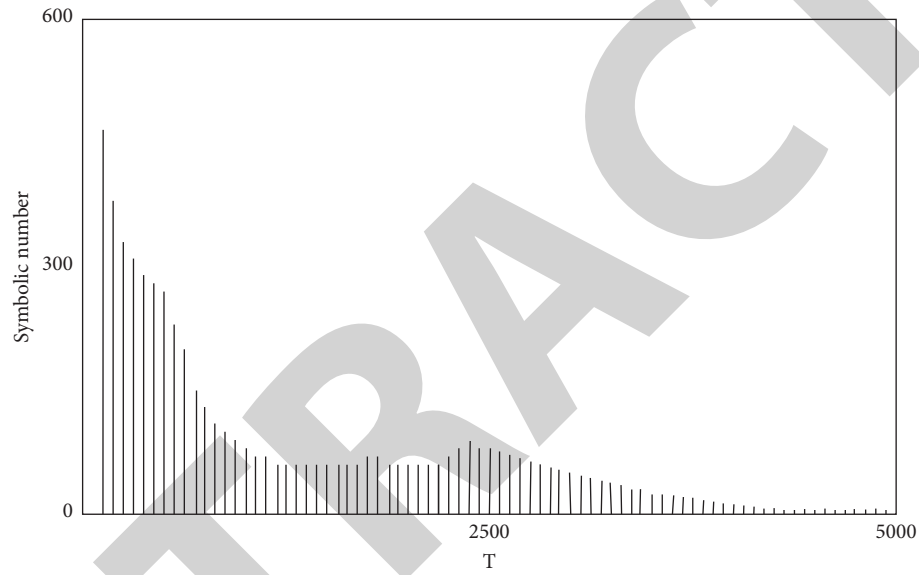

Figure 3: The size of the character set produced by different *t*.

neighbor distance of the sequence to measure the degree of abnormality of the sequence, the time complexity will be reduced to 0 in the optimal case. After constructing the abovementioned dataset, the nearest neighbor distance is more likely to be greater for subsequences that appear less frequently, and it is more suitable for datasets with different characteristics. For the value of *t*, the impact is greatest at the turning point. Therefore, the algorithm can show different properties by selecting different *t*, so selecting the appropriate *t* value is also an important aspect of the symbolization algorithm.

3.3. Symbol Assignment. To symbolize a given time series database, it is necessary to manually give the clustering parameter *k* (as a parameter to determine the running time), but how to determine the selection of the clustering series *m* is a problem that needs to be discussed.

Figure 4 shows the clustering symbolization process of level 2 and 2 means. By assigning each value of the sequence according to the binary value, the subsequent division is carried out. The specific principle is as follows.

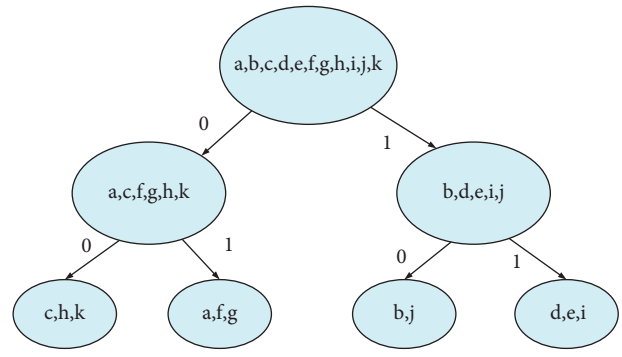

Figure 4: 2-level 2-mean clustering symbolization.

Clustering is a method to explore the internal structure and essential characteristics of the data. The characteristics of the data in the subsequence can also be well expressed. Therefore, by splicing these fragments to get the entire time series, the segmented subsequence will contain less data. The smaller the amount of data in each segment, the finer the feature representation. However, by carefully observing the

local changes in each segment, it can reflect the global change trend of the time series. This is applicable to various time series feature representation methods, but it has great limitations in the application of nonstationary signals. It can only roughly reflect the global trend of the time series, and the linear representation method will have large errors. This segmentation method is not as reasonable as manual segmentation and can only be used for the data preprocessing part of the subsequent clustering analysis. Each clustering takes the subspace as the shape of the object cluster, which is very random. Through linear fitting and combining the respective advantages of these three methods and by analyzing one by one, it can more objectively reflect the distribution of the objects in space, can better dig out potential patterns of time series, and can connect close-core objects and their neighborhoods.

It can also be seen from Table 3 that the assignment of the symbolized result is quite effective. For binary language, this is a very simple and effective way of assigning and distinguishing. After classification, there are three datasets of 0, 1, and 2, and only performing correlation distance analysis on it will be much better than the original dataset.

4. Symbolic Algorithm Diagnosis and Optimization

4.1. Related Technologies

4.1.1. Wavelet Data Preprocessing. Since the actual collected signal has a lot of noise, which has a greater impact on the accuracy of the clustering algorithm, this article first uses wavelet transform to remove the signal noise and then improves the accuracy of the subsequent processing process, as shown in the following equation:

$$D = \int_{-\infty}^{\infty} \frac{|A(w)|^2}{w} dw < \infty. \tag{18}$$

After performing the wavelet transform, it can be represented by the equation as

$$QT(a,x) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) w\left(\frac{t-1}{a}\right) dt, a > 0. \tag{19}$$

The equivalent frequency is represented as

$$QT(a,x) = \frac{\sqrt{a}}{2\pi} \int_{-\infty}^{\infty} x(t) w(\mathrm{aw}) e^t dw. \tag{20}$$

Among them, similar to the basic properties of the Fourier transform, in the processing of data, transformation is performed, and the main difference between the wavelet function system and the harmonic function system is the translation and expansion of the wavelet function.

As shown in Figure 5, after wavelet transform processing, the fluctuation frequency of the signal sequence becomes lower and more stable. The statistic reaches its maximum value at the 736$^{\text{th}}$ sample. After being processed by the wavelet transform, the sample also reached the maximum value at the 736$^{\text{th}}$ place, and there was a mean change point at the total sample, and the sequence did not change in trend after processing. Therefore, the wavelet transform has a very good effect. There is no missing data after the transformation of the data processing, and the amount of information is still not reduced, but the data processing is much simpler. Compared with the situation where the original sequence has a lot of complicated and redundant data, the current sequence fluctuates smoothly after processing, the peak value is preserved, and the transformation can be effectively carried out.

Table 3: Symbolized results of 2-level2-mean.

Sequence	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*
Label	01	10	00	10	10	01	01	00	10	10	00
Symbol	1	2	0	2	2	1	1	0	2	2	0

4.1.2. Support Vector Machine. The mechanism of the support vector machine is to find an optimal classification hyperplane that meets the classification requirements, so that the hyperplane can maximize the isolation between positive and negative examples while ensuring classification accuracy.

In Figure 6, the squares and circles represent the two types of data samples, H is the optimal classification line of the two types of samples, and H ensures that the data is accurately classified and the classification interval is the largest. If a two-dimensional plane is transformed into a three-dimensional space, then H represents the optimal classification plane. The linear discriminant function of the optimal classification surface is shown in the following equation:

$$f(a) = \mathrm{sgn}(wa + b) = \mathrm{sgn}\left\{\sum_{p=1}^{m} x_p^* y_p\left(x_p, x\right) + b\right\}. \tag{21}$$

In formula (21), sgn() is the judgment function, and the experience level of the new test statistic proposed in this article is not well controlled when the sample size is 300 and the parameter is 0.2; that is, the long memory parameter has a certain influence on the experience level.

4.2. Symbolization Algorithm Diagnosis. The so-called diagnosis is outlier detection, among which three common outliers are shown in Figure 7.

In Figure 7, the three types of outliers are additive anomaly points, innovative anomaly points, and time series anomalies. For additive abnormal points, they are expressed as prominent points. In the sequence, there will usually be data recording errors or relatively high abnormal point detection errors, which is very unfavorable for the analysis of the sequence. The normal series fluctuates within a certain range, but the appearance of outliers will greatly affect the stationarity of the series. For innovation anomalies, this usually occurs in abnormal fluctuations in the sequence, which will cause the overall trend of the sequence to change, which will make the sequence degenerate, and the time series will be distorted. In this way, the time series will lose its

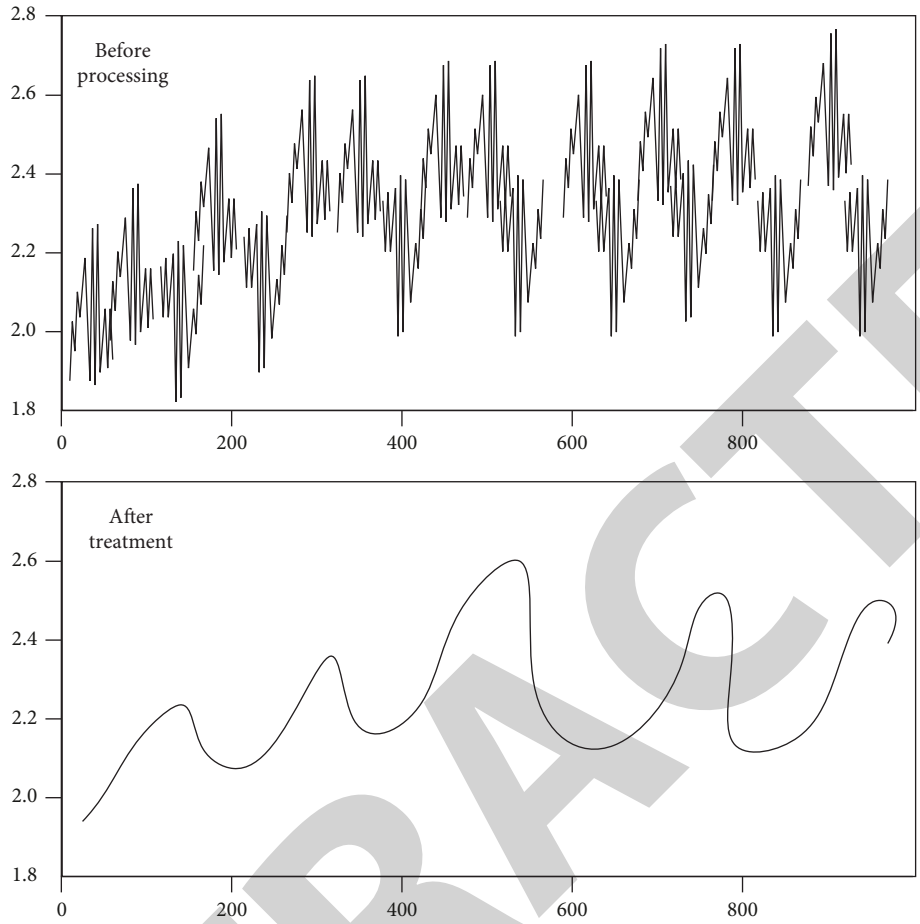

Figure 5: Signal sequence before and after wavelet transform processing.

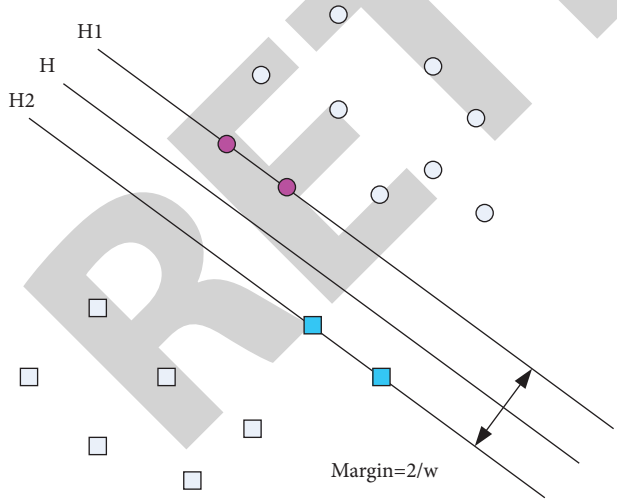

Figure 6: Support vector machine.

analytical significance. For anomalous segments of a time series, it is manifested as an anomaly at the peak of the sequence, which is smaller than the anomaly at an additive anomaly point. This usually occurs in a small and short sequence, which will have a greater impact on the analysis of a small sequence.

4.3. Symbolization Algorithm Optimization. This chapter presents an algorithm for finding anomalous subsequences based on pruning optimization, and the algorithm first symbolizes the time series based on the SANS algorithm in the previous chapter, then clusters the obtained symbols, and finally optimizes the pruning and optimization of the abnormal subsequence search based on the information of the symbol clustering. The basic idea of the abnormal subsequence detection algorithm in this article is specifically introduced, as shown in Figure 8.

It can be seen in Figure 8 that the optimized algorithm time is significantly reduced. Although the algorithm does not assume that the dataset is in accordance with the standard normal distribution, it is reduced by at least 0.6 s. When the parameters increase, the processing time of the optimized algorithm becomes lower, but as the parameters increase, the processing time will slowly become higher, which increases the amount of calculation that needs to be performed. This is different from the original algorithm. As the parameters of the original algorithm become higher, the running time will become lower. This shows that the optimized algorithm has a very good performance at low parameters, but the follow-up time may be higher than the original algorithm, but this is also out of consideration, because the sequence used in daily life generally does not exceed the scope of this experiment.

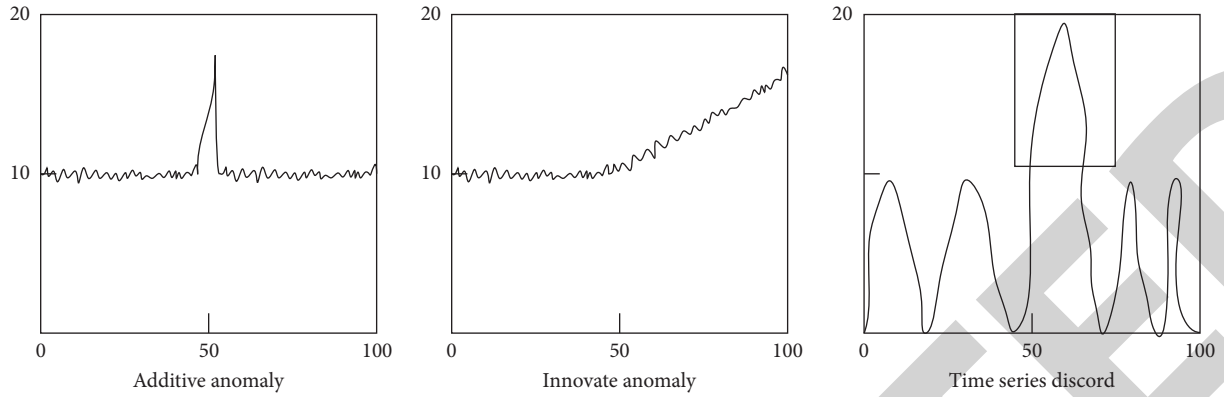

Figure 7: Three different outliers.

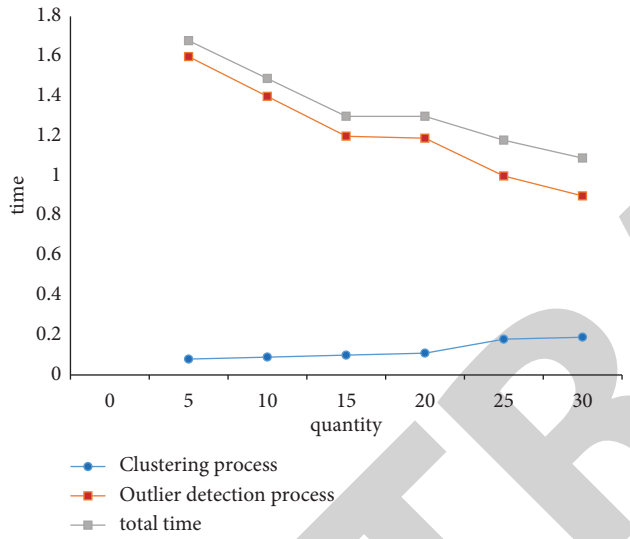

Figure 8: The influence of parameters on the running time of the two algorithms.

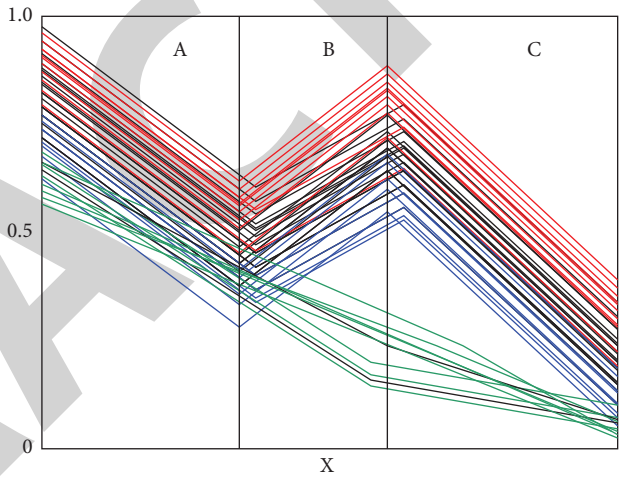

Figure 9: Graph of clustering results.

After that, do K-means clustering on the time series, as shown in Figure 9.

In Figure 9, the clustering effect for the three datasets is still satisfactory. The equal probability discretization method is used. Because of the same size, the distance between the datasets is too large, which is not conducive to clustering. It can be seen that in the A and C parts, the aggregation degree of the three sets is relatively high, but in part B, there is an obvious stratification. This is actually the error of the dataset when the K-means clustering algorithm is an equal probability.

Figure 10 shows the accuracy of the algorithms based on the different K-values before and after optimization. It can be seen from the data in the figure that, among them, the algorithm after symbolization is simpler and the calculation speed is faster, and it has obvious advantages for exploring the unlabeled data structures. It can be realized by only taking the first few coefficients. The use of equal-width windows to intercept time series has a larger amount of data than the ordinary static data, making it difficult to obtain good clustering results, or even

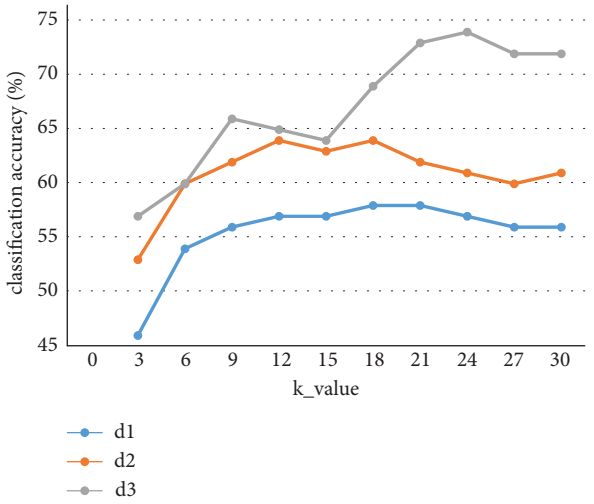

Figure 10: Classification accuracy of different algorithms.

clustering. The empirical potential of the optimized algorithm statistics is mostly lower than the original statistics algorithm, but as the sample size increases and the window width parameter increases, the gap gradually decreases.

The K-means clustering algorithm proposed in this article aims at the problem that its dimensionality is too high, which is not conducive to mining, it cannot reflect the local characteristics of the time domain, it is difficult to obtain a good feature representation, the performance is poor, and the computational complexity is high, to discover its latent patterns through pattern separation, which are used in time series feature extraction and clustering methods. If neither method rejects the null hypothesis of no change point, then it can be considered that there is no change point in the data. This method is a model whose evaluation is well-adapted to the data and can be used as a strong support for subsequent data analysis.

5. Conclusions

For the problem of time series symbolization, this article is based on the K-means clustering method to symbolize it and to optimize the algorithm. There are a lot of relevant research studies on time series, which is also because many analyses in daily life are used. Therefore, in this article, a more in-depth study of the problem of symbolization optimization and the comparison and analysis of its running time and accuracy are carried out. The result is also expected. However, many aspects of time series symbolization need to be studied in-depth. Due to the limited space of this article, there is a lack of relevant introduction to the underlying logic and operating principles of the optimized algorithm. We hope to continue the related research study in the later period.

Data Availability

No data were used to support the study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] D. Xia, F. Ning, and W. He, "Research on parallel adaptive canopy-K-means clustering algorithm for big data mining based on cloud platform," *Journal of Grid Computing*, vol. 18, no. 2, pp. 263–273, 2020.

[2] J. Yang, Y. Ma, X. Zhang, S. Li, and Y. Zhang, "An initialization method based on hybrid distance for k-means algorithm," *Neural Computation*, vol. 29, no. 11, pp. 3094–3117, 2017.

[3] Y. Jing and J. Wang, "Tag clustering algorithm LMMSK: improved K-means algorithm based on latent semantic analysis," *Journal of Systems Engineering and Electronics*, vol. 28, no. 2, pp. 374–384, 2017.

[4] B. Osmanoglu, F. Sunar, S. Wdowinski, and E. Cabral-Cano, "Time series analysis of InSAR data: methods and trends," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 90–102, 2016.

[5] X. Zhu, B. Chen, X. Wang, and T. Li, "Time series segmentation clustering: a new method for S-phase picking in microseismic data," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, no. 99, pp. 1–5, 2022.

[6] A. K. Deklel, A. M. Hamdy, and E. M. Saad, "Multi-objective symbolic regression using long-term artificial neural network memory (LTANN-MEM) and neural symbolization algorithm (NSA)," *Neural Computing & Applications*, vol. 29, no. 4, pp. 935–942, 2018.

[7] N. Rasidah Ali and K. R. Ku-Mahamud, "Spatial clustering algorithm for time series rainfall data using X-means data splitting," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 14, no. 6, pp. 221–226, 2017.

[8] R. Amir Moeini, A. Davoud, S. Hamid Salehi, and N. Hamed, "Developing an automatic phenology-based algorithm for rice detection using sentinel-2time-series data. Selected topics in applied earth observations and remote sensing," *IEEE Journal of*, vol. 12, no. 5, pp. 1471–1481, 2019.

[9] M. Ochodek, M. Staron, and W. Meding, "Simsax: a measure of project similarity based on symbolic approximation method and software defect inflow," *Information and Software Technology*, vol. 115, pp. 131–147, 2019.

[10] A. Zhang and G. Shi, "A fast symbolic SNR computation method and its Verilog-A implementation for Sigma-Delta modulator design optimization," *Integration*, vol. 60, pp. 190–203, 2018.

[11] V. Makarov and N. Romaniuk, "Symbolic algorithm of the functional-discrete method for a sturm-liouville problem with a polynomial potential," *Computational Methods in Applied Mathematics*, vol. 18, no. 4, pp. 703–715, 2018.

[12] A. Euldji, A. Tienti, and A. Boudghene Stambouli, "A novel modelling approach of RLC electrical circuits for symbolic circuit analysis by the direct topological method," *Arabian Journal for Science and Engineering*, vol. 45, no. 3, pp. 1897–1909, 2020.

[13] H. Liu, H. X. Lin, Jiang, X. Mao, Q. Liu, and B. Li, "Estimation of mass matrix in machine tool's weak components research by using symbolic regression," *Computers & Industrial Engineering*, vol. 127, no. JAN, pp. 998–1011, 2019.

[14] J. T. Jeng, C. M. Chen, S. C. Chang, and C. C. Chuang, "IPFCM clustering algorithm under euclidean and hausdorff distance measure for symbolic interval data," *International Journal of Fuzzy Systems*, vol. 21, no. 7, pp. 2102–2119, 2019.

[15] V. Yakhno and M. Altunkaynak, "Symbolic computation of the time-dependent electric and magnetic fields in bi-anisotropic media with polynomial inputs," *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 31, no. 5, pp. e2339–e2339.19, 2018.

[16] K. Nabeshima and S. Tajima, "Algebraic local cohomology with parameters and parametric standard bases for zero-dimensional ideals," *Journal of Symbolic Computation*, vol. 82, no. SEP.-OCT, pp. 91–122, 2017.

[17] J. Abbott, "Fault-tolerant modular reconstruction of rational numbers," *Journal of Symbolic Computation*, vol. 80, pp. 707–718, 2017.

[18] X. Li, L. Gao, X. Xu, J. Shao, F. Shen, and J. Song, "Kernel based latent semantic sparse hashing for large-scale retrieval from heterogeneous data sources," *Neurocomputing*, vol. 253, no. Aug.30, pp. 89–96, 2017.

[19] T. Chen, T. L. Lee, and T. Y. Li, "Mixed cell computation in Hom4PS-3," *Journal of Symbolic Computation*, vol. 79, pp. 516–534, 2017.

[20] A. R. Adem, "Symbolic computation on exact solutions of a coupled Kadomtsev–Petviashvili equation: lie symmetry analysis and extended tanh method," *Computers & Mathematics with Applications*, vol. 74, no. 8, pp. 1897–1902, 2017.

[21] H. Yoshida, G. Li, T. Kamiya et al., "KLOVER: automatic test generation for C and C++ programs, using symbolic execution," *IEEE Software*, vol. 34, no. 5, pp. 30–37, 2017.

[22] F. M. Gomes, F. M. Pereira, A. F. Silva, and M. B. Silva, "Multiple response optimization: analysis of genetic programming for symbolic regression and assessment of desirability functions," *Knowledge-Based Systems*, vol. 179, no. SEP.1, pp. 21–33, 2019.

[23] M. Wisniewski and S. Deniziak, "BMB synthesis of binary functions using symbolic functional decomposition for LUT-based FPGAs," *Journal of Parallel and Distributed Computing*, vol. 120, no. OCT, pp. 16–22, 2018.

[24] E. S. Selima, Y. Mao, X. Yao, A. M. Morad, T. Abdelhamid, and B. I. Selim, "Applicable symbolic computations on dynamics of small-amplitude long waves and Davey–Stewartson equations in finite water depth," *Applied Mathematical Modelling*, vol. 57, no. MAY, pp. 376–390, 2018.

[25] C. Kyrkou, C. S. Bouganis, T. Theocharides, and M. M. Polycarpou, "Embedded hardware-efficient real-time classification with cascade support vector machines," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 1, pp. 99–112, 2016.

[26] J. Kirkpatrick, R. Pascanu, N. Rabinowitz et al., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences of the USA*, vol. 114, no. 13, pp. 3521–3526, 2017.

[27] H. Palangi, L. Deng, Y. Shen et al., "Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 694–707, 2016.