# A Presentation of Time Series Algorithm Based on K-Mean and Its Application in Clustering

**Yeping Peng**

Software College of Jishou University, Zhangjiajie, Hunan Province, China

**Keywords:** Time series; Algorithm; Similarity; K-mean; Clustering

**Abstract:** As the data volume of a time series database is much greater than that of an ordinary database, some general data mining tools cannot be applied to time series directly to generate satisfactory result. Therefore, a piecewise linearization representation algorithm is proposed for time series in this paper, which can greatly improve the computation speed of similarity measurement. Based on the piecewise linearization representation, a similarity calculation method is proposed, which is insensitive to various deformation of time series. The k-mean clustering algorithm is applied to time series represented by piecewise linearization for a desired result.

## 1. Introduction

Time series refers to a series of data recorded at a certain time interval, such as changes in stock prices, daily sales of supermarkets, changes in temperature and astronomical observation records. Time series data exist in many fields, such as scientific research records, medical records and business transactions, and reflect the quantitative relationship of some aspects of some complex systems. With the continuous development of science and technology, the storage capacity of computers and storage devices is increasing, and the time series database is also growing, so it is more and more necessary to study the time series. However, because the amount of time series data is too large, and it is difficult to define appropriate similarity measurement formulas for time series in different fields, it is difficult to cluster time series using common data mining tools. In order to solve the above problems, it is one of the effective methods to improve the computational efficiency to adopt a more effective description method for time series. [1] At present, there are many methods to improve the efficiency of time series expression, which can be roughly divided into the following three types:

(1) Mapping time series from time domain to frequency domain through Fourier transform or wavelet transform, and representing the original time series data with very few low frequency coefficients, this method is highly efficient in data concentration, but sensitive to noise and not intuitive.

(2) Piecewise linearization. The central idea is to approximately replace the original time series with K straight line segments. This method can achieve the purpose of data compression, and allows scaling on the time axis, but requires that the number of straight-line segments K be given in advance. The choice of K value is a key factor, too small will lose useful information, too large will produce too much redundant information.

(3) Landmarks technology defines some significant points as turning points, and uses these turning points to replace the original time series data. In a curve, when the n-th derivative of a point is zero, the point is defined as the n-th turning point. Because the importance of high-order turning point is relatively small, and it is difficult to obtain high-order derivatives for actual data, so generally only low-order turning point is chosen, and the order of turning point is different for different professional fields. In this paper, a piecewise linearization method for structural adaptation is proposed. It combines piecewise linearization with turning point technology. The first-order turning point is regarded as the piecewise linearization piecewise point, and the number of linearized piecewisekcan be automatically generated by an error criterion. Because piecewise linearization reduces the amount of data greatly, it can be clustered by common clustering tools. In

  DOI: 10.25236/isrme.2019.068

this paper, the most common clustering algorithm k-means method is used to cluster time series represented by piecewise linearization. [2]

## 2. Piecewise Linearization of Time Series

Piecewise linearization is to divide the time series data into several straight-line series, and to replace the original data approximately by a series of straight-line segments. It can effectively compress and filter data, and its presentation can be directly perceived through the senses with strong visualization. In addition, it is characteristic of high compression rate, which enables high computing speed of similarity, especially for time series with large amount of data.

### 2.1 Piecewise Linearization Method

The key to piecewise linearization is how to select the appropriate number of straight-line segments k and approximate the original time series with k straight-line segments. If the value of k is too small, useful information will be lost; if it is too large, too much redundant information will be generated. In order to avoid the disadvantageous factors caused by the artificial selection of the value of k, this paper combines the turning point technology with piecewise linearization, uses the first-order turning point as the endpoint of the straight-line segment, and combines the corresponding error adjustment criteria. This method can automatically generate the number of straight-line segments k. [3]

Assuming that time series S contains n points and that each value of S is a function of time, then S is expressed by the following formula:

$S_t = y(t),\ t = 1, 2 \ldots n$(1)

In the formula, $S_t$ is the value of time series S at time t. By obtaining the turning point of the first order of Y(t) the extreme point of the curve is obtained. The turning point is defined as ($t_i$, y ($t_i$)). The time series point between the two turning points can be approximated by a straight line. The right end of the former line is the left end of the latter line.

Let two turning points be $(t_i, y(t_i)), (t_{i+1}, y(t_{i+1}))$respectively, then the slope of the approximate straight line is α:

$$\alpha = \frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i} \quad (2)$$

The approximate error in this area is$\beta_j$ (j = 1,2…M), wherein M is the number of y(t) between $t_i$ and $t_i$+1.

$$B_j = y(t_{i+j}) - y(t_i) + \alpha \cdot (t_{i+j} - t_i) \quad (3)$$

The approximate error Bi of the whole line segment is defined as:

$$B_j = \sum_{j=1}^{m} \frac{\beta_j}{M} \quad (4)$$

The right end of each line segment is determined by the following method. Complete in the following steps:

Sep 1 If the condition$P_1 \le B_i \le P_2$ is satisfied, where$P_1 < P_2$, then $(t_{i+1}, y(t_{i+1}))$ is adopted as the right end point of the straight-line segment;

Step 2 Otherwise, if$B_i < P_2$, take an original data point as the right endpoint in the middle of $(t_{i+1}, y(t_{i+1}))$ and its nearest turning point at the left end;

Step 3 If $B_i < P_1$, then judge whether the next turning point satisfies the above conditions until the conditions are met.

This method is expressed as$MAX - MIN(P_1, P_2)$. The values of P1 and P2 depend on different subjects. When the original data of time series changes greatly, the values of P1 and P2 are also relatively large. After the above segmentation, the optimal number of straight lines will be automatically generated.

## 3. Time Series Similarity Definition Based on Piecewise Linearization

After piecewise linearization, the original time series will be approximated by a series of straight lines connected at the beginning and end. The following is the definition of the symbol for the full text. An original undivided time series is represented by the capital letter Sin italic, and the piecewise linearized time series is represented by the capital letter **S** in bold italic. S is a time series consisting of four variables with length K, i.e.,

$S \equiv \{STL, STR, SYL, DYR\}$(5)

Where in, the line segment in number i in S is represented by the left endpoint (STL, SYL) and the right endpoint (STR, SYR) of the line.

### 3.1 Various Deformations of Time Series

Various distortions and distortions are often encountered in the research of time series similarity algorithm. The more inclusive a similarity algorithm is to all kinds of deformation, the stronger the function of this algorithm is proved. Assuming that a time series is represented by a function y (t), all kinds of deformation can be defined as the following functions. [4]

(1) Noise

NO(y) is defined as NO(y(t)=y(t) + e(t), where in e(t) is random noise.

(2) Shifting

SH(y) is defined as SH(y(t)=y(t)+$c_s$, where in $c_s$ is a constant.

(3) Amplitude scaling

AS(y)is defined as AS(y(t)=$c_A$y(t), where in $c_A$ is a positive constant.

(4) Time scaling

TS(y) is defined as TS(y(t)=y ($c_t$t), where in $c_t$ is a positive constant.

(5) Linear drift

LD(y) is defined as LD(y(t)=y(t)+L(t), where L(t) is a linear equation.

Of course, more deformation can be combined from these functions. For the case of noise, piecewise linearization can directly filter out the noise. Similarity measurement needed in many professional fields should be insensitive to translation. This paper puts forwards an effective measurement method that can cope with such deformation. Based on piecewise linearization, a simple similarity measurement formula is defined, which is insensitive to the first four types of deformation mentioned above. [5]

### 3.2 Similarity Distance Formula

Firstly, the amplitude of each endpoint of the sectioned time series S is normalized. If the value of an endpoint is Y, then the normalized value is:

$$Y = \frac{Y - Y_{MIN}}{Y_{MAX} - Y_{MIN}} \quad (6)$$

In the formula: $Y_{MIN} = MIN(SYL, SYR), Y_{MAX} = MAX(SYL, SYR)$. The formula for the similarity distance between time series A and B are defined as follows:

$$D(A, B) = \sum_{i=1}^{K} \frac{(AYR_i - AYL_i)}{(ATR_i - ATL_i)} - \frac{(BYR_i - BYL_i)}{(BTR_i - BTL_i)} \quad (7)$$

If the number of the original data points of time series isn, the computing time of similarity is only K/n of the original data after piecewise linearization.

## 4. K-Mean Clustering Algorithm for Time Series Based on Piecewise Linearization

Cluster analysis is an important human behavior. As early as childhood, one would learn how to distinguish different kinds of animals or plants by constantly improving the subconscious clustering model. Clustering analysis has been widely used in many fields, including pattern recognition, image processing and data analysis. Of course, clustering can also be applied to time series analysis, such as market research and earth observation database. However, due to the large amount of data in time series, and the clustering analysis technology mainly focuses on distance-based analysis, the

direct application of clustering algorithm for ordinary data to time series will greatly increase the computational cost. By piecewise linearization of time series, the amount of data is greatly reduced, so it is more suiTable for the application of common data mining algorithm.

K-means algorithm takes kas parameter and divides N objects into k clusters, which makes the similarity between clusters higher, while the similarity between clusters is lower. The calculation of similarity is based on the average value of objects in a cluster.

The processing flow of K-means algorithm is as follows. Firstly, k objects are randomly selected, and each object initially represents the average value of a cluster. For each remaining object, it is assigned to the nearest cluster according to the distance from each cluster center. The distance measurement here adopts formula (7) and then recalculates the average value of each cluster. This process repeats until the criterion function converges. The criterion function is usually defined as the square error criterion.

$$E = \sum_{i=1}^{k} \sum p \in C_i |p - m_i|^2 (8)$$

In the formula E is the sum of the square errors of all objects,p the object of Cluster $C_i$,$m_i$the average value of Ci (the dimension of p and $m_i$ is k, i.e., the number of straight-line segments after time series segmentation), which criterion tries to make the generated result cluster as compact and independent as possible.

But the disadvantage of this method is that the number of clusters k must be given in advance. In order to solve this problem, according to the specific research object, this paper first proposes a change of k, which is divided into 2-6 classes, and calculates the above calculation for each k until convergence, and calculates the distance between the average values of all clusters, which is recorded as DC. After calculating all k values, the minimum k of E/DC value is selected as the optimal number of clusters. This method makes the cluster compact and the distance between clusters maximum. [6]

## 5. Conclusion

In this paper, a piecewise linearization algorithm for time series with adaptive structure is proposed. The number of linearized segments k can be automatically generated by an error criterion. Experiments show that this method greatly improves the computing speed of similarity measurement, and thus greatly improves the speed of subsequent clustering algorithms. At the same time, based on the piecewise linearization representation, this paper proposes a similarity calculation method. This method is insensitive to the four kinds of deformation of the time series mentioned in the paper, and the effect is good.

## Acknowledgement

This research was financially supported by the General Project of Hunan Provincial Department of Education (No. 16C1312).

## References

[1] H. Zhang, T. B. Ho, M. S. Lin, An evolutionary K-means algorithm for clustering time series data, International Conference on Machine Learning & Cybernetics,(2004) 282-1287.
[2] H. T. Liu, Z. W. Ni. Clustering Method of Time Series Based on EMD and K-means Algorithm[J]. Pattern Recognition & Artificial Intelligence, 22, 5 (2009)803-808.
[3] C. W. Tsai, C. S. Yang, M. C. Chiang, A Time efficient Pattern Reduction algorithm for k-means based clustering[J]. Information Sciences, 181, 4 (2007)716-731.
[4] Y. Li, H. Wu, A Clustering Method Based on K-Means Algorithm, Physics Procedia, 25 (2012) 1104-1109.
[5] C. S. Peng, H. Wang, S. R. Zhang, et al, Landmarks: a new model for similarity-based pattern querying in time series databases, IEEE Conference on Data Engineering, (2000) 33-42.
[6] Y. Zheng, A Data Mining Algorithm Based on Improved K-Means Clustering. Applied Mechanics & Materials, (2014)2028-2031.