A Representation of Time Series Pattern Based on Skeleton Points

Yeping Peng

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

Keywords: Time series; Algorithm; Piecewise; Skeleton points.

Abstract: Time series is the characteristic of a large amount of data, high dimension and fast data updating, so it is difficult to directly carry out data mining in the original time series. Drawing on the basic idea of linear segmentation of time series, a piecewise linear representation based on key points is proposed in this paper. Linear straight-line segments composed of skeleton points are used to approximate the time series. Skeleton points are regarded as the segmentation points to reflect the main features of time series. Experiment shows that this method can reduce the dimension of time series and minimize the overall error.

1. Introduction

Time series is a sequence consisting of sampled values of a physical quantity of an objective object at different time points arranged in chronological order. Typical of important high-dimensional data, it is widely used in economic management and engineering fields. Time series data mining can obtain useful information related to time contained in the data and realize knowledge extraction. As time series is characteristic of high dimensionality, complexity, dynamics, high noise, and easy to achieve large-scale characteristics, directdata mining in time series will not only cost a lot in storage and calculation, but also may affect the accuracy and reliability of the algorithm.

The pattern representation of time series is an Abstraction and representation of generalized feature, which is a new description of time series at a higher level.The pattern representation of time series not only has the function of compressing data and keeping the basic form of time series, but also has the ability of de-noising. The commonly used pattern representation of time series mainly includes frequency domain representation, symbol representation, principal component analysis representation and piecewise linear representation. Among various pattern representation of time series, piecewise linear representation is relatively simpler. It is simple and intuitive, and most representations support dynamic incremental updating of time series. Its basic idea is to replace the original time series approximately with K straight lines. [1]

The piecewise linear representation based on skeleton points proposed in this paper is simpler and more intuitive, and the fitting error between the piecewise linear representation and the original time series is smaller. It can achieve better results in simulation time series and real time series.

2. Definition of Time Series and Relevant Concepts

2.1 Time Series

If X = <x_1 = (t_1, $v_1(t_1)$), x_2 = (t_2, $v_2(t_2)$)…xn = (tn, vn (tn))>, then record timet$_i$ is strictly incremental(i< j < = >t_i<t_j)and vi is the observed valueat time t_i.

Time series includes large amount of data and fast data updating. It is very difficult to mine frequent patterns directly from the original time series. In order to characterize the main form of communication time series, neglect the minute details, and compress the time series for smaller storage and calculation costs, this paper mainly studies the changing patterns and rules of time series over a period of time. It is not the value of a single sequence point in time series, so this paper uses piecewise linear representation of time series, which divides a long time series into several relatively short but not overlapping subsequences. On the basis of piecewise linear representation, the sequence is divided into several line segments, each line segment represents a pattern. [2]

DOI: 10.25236/isrme.2019.070

Piecewise linear representation is more in line with human's intuitive experience, and its calculation speed is larger. In order to eliminate the influence of local fluctuations on sequence comparison and reduce the complexity of sequence description, the sequence can be described as a series of line segments. As an approximation of the original sequence, the starting and ending points of line segments are local maxima and minima under given conditions in the sequence, also known as skeleton points or inflexion points.

2.2 Skeleton Points in Series

Skeleton points are points, among all the extremum points of time series, that greatly affectthe shape of original dataand have the maximal errors after being expressed linearly in the region they are located in.

The skeleton points of a series include minimal and maximal ones.

2.3 Minimal Point

Given constant Rand time series{<x_1 = (a_1, t_1)…x_n = (a_n, t_n)>}, x_m (1<m<n)is called a minimal point whenxmmeets two conditions, i.e., there are subscriptsi andj and 1≤i <m<j≤n that make the following statements tenable:

(1) am will be the minimum value ina$_1$, a$_2$…a$_j$;

(2) $a_i/a_m \geq R$and$a_j/a_m \geq R$.

Here R is a controllable parameter. The larger the R value, the fewer the relative important points selected, the coarser the description of time series segmented. On the contrary, the smaller the R value, the more important points selected and the finer the description of segmented data. Therefore, R can adjust the fluctuation amplitude of time series to control the precision of data mining. [3]

2.4 Maximal Point

Given constant R and time series{<x_1 = (a_1, t_1)…x_n = (a_n, tn)>}, x_m(1<m<n) is called a maximal point whenxmmeets two conditions, i.e., there are subscripts i and jand 1≤i<m<j≤n that make the following statements tenable:

(1) am is the maximum value in a$_1$, a$_2$…a$_j$;

(2) $a_i/a_m \leq R$ and$a_j/a_m \leq R$.

Here R is a controllable parameter. The larger the R value, the fewer the relative important points selected, the coarser the description of time series segmented. On the contrary, the smaller the R value, the more important points selected and the finer the description of segmented data. Therefore, R can adjust the fluctuation amplitude of time series to control the precision of data mining. [3]

2.5 Definition of Time Series Segmentation

Connecting the skeleton points of time series, a series of line segments are obtained.

S = {<(t_1, v_1), (t_2, v_2)>, <(t_2, v_2),(t_3, v_3)>…<(t_n-, v_n-1), (t_n, v_n)>}, wherein v_i, v_{i+1} are the starting and ending values of piecei; nis the number of linear pieces in time series.

3. Piecewise Representation of Time Series

In order to fully reflect the overall characteristics of the original data and highly compress the time series, the piecewise point selected in this paper is the skeleton point of the time series sequence. In the selection process, the starting point and the ending point are first taken as the skeleton points of the initial sequence, and the sequence is scanned and compared to obtain the skeleton point set, and then the sequence is linearized piecewise. [4-5]

The algorithm for obtaining skeleton points of time series is described as follows:

(1) Input time series S;

(2) Output sequence skeleton points.

The algorithm steps are as follows:

(1) Initialize the set of skeleton points in sequence;

(2) Initial and termination points are added to the set of sequence skeleton points.

(3) Compute time series extremum points in{x_1 = a_1, t_1)…x_n= (a_n, t_n)}and find out sequence skeleton points according to the definition of sequence skeleton points.

The algorithm has the following three characteristics:

(1) It only needs scanning the sequence once, and the piecewise process only needs simple comparison, without complex least squares operation.

Inter-complexity is $O(_n)$ (n is sequence length);

(2) Support online selection of sequences;

(3) Only one R is needed for input parameters.

4. Experiment and Results

The experiment uses Ma Data dataset, which is a simulation dataset used by MA et al. to detect abnormal patterns. The time series of the simulation dataset is generated by the following stochastic processes:

X (t) = sin (40π /Nt) + $e_1(t)$ + $e_2(t)$

t =1, 2…N (N =1 200)

$e_1(t)$ and $e_2(t)$are defined as follows:

$$e_1(t) = \begin{matrix} n_1(t) \;\; t \in [600,620] \\ 0 \end{matrix}$$

Here, $n_1(t)$ meets normal distribution(0, 0.5).

$$e_2(t) = \begin{matrix} 0.4 \times \sin\left(\frac{40\pi}{Nt}\right) \;\; t \in [820,870] \\ 0 \end{matrix}$$

Figure 1 is the comparison of Ma_Data dataset before and after segmentation based on skeleton points.

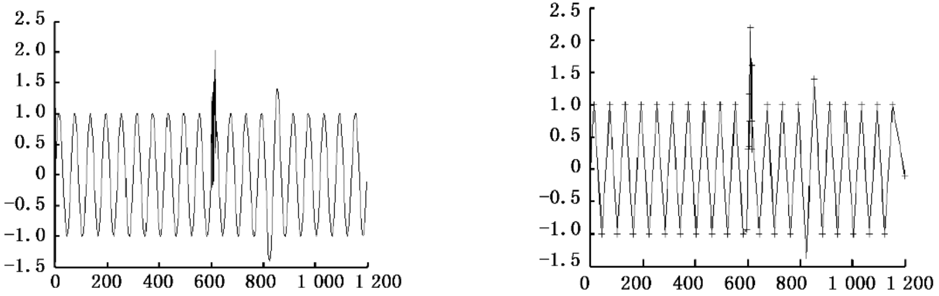

Figure 1 Comparison of Ma_Data Dataset before (Left) and After (Right) Segmentation Based on Skeleton Points

This time series consists of 1200 data points. By using the time series linear piecewise algorithm adopted in this paper, the time series can be represented only by 45 line segments composed of 46 skeleton points in Figure 1. [6]

5. Comparison with PAA algorithm

Piecewise clustering of time series is similar to Piecewise Aggregate Approximation (PAA), which method is to divide the time series into equal-width segments, and each sub-segment is represented by the average value of the time series on the sub-segment. In the contrast experiment with PAA, the performance of the algorithm by using the actual data sets of 8 time series from different fields. The linear representation method of time series based on skeleton points proposed in this paper is abbreviated as SP. [5]

The experiment result is compared with the performance of the algorithm by using the actual data sets of 8 time series from different fields provided by KEOGH et al. The linear representation method of time series based on skeleton points proposed in this paper is abbreviated as SP.

At the same compression rate (compression rate is 80%), the comparison between time series segmentation algorithm based on skeleton points and PAA fitting error is shown in Table 1.

Table 1 Comparison of Fitting Errors between SP and PAA At Same Compression Rate

Algorithm	Chaotic	Ocean	Darwin	Earthquake	Sunsport	Powerplant	Speech	Tide
PAA	1.76	0. 31	4. 22	3. 48	2. 65	1. 07	2. 39	3. 22
SP	1. 63	0. 38	2. 41	2. 10	3. 52	2. 23	1. 52	2. 29

As can be seen from Table 1, the fitting error of the proposed method (SP) is smaller than that of PAA in five of the eight time series.

6. Summary

The pattern representation of time series can compress the data of time series, retain the main form of time series and remove the interference of details. It can better reflect the characteristics of time series and improve the efficiency and accuracy of data mining. The linear representation of time series is an important means to realize the pattern representation of time series. Experiments show that the proposed method is based on skeleton points. The linear representation of time series is effective and efficient in segmentation of time series, and can describe the overall characteristics of time series well. Through this algorithm, time series researchers can be provided with a basis for describing time series, and on the basis of this algorithm, time series patterns can be further excavated.

Acknowledgement

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

References

[1] F. L. Chung, T. C. Fu, T. Y. Ng Vincent, R. W. P. Luk, An evolutionary approach to pattern-based time series segmentation, IEEE Transactions on Evolutionary Computation, 8, 5 (2004) 471-489.

[2] F. Liu, G. D. Guo, An improved algorithm for time-series pattern discovery, Journal of Minnan Normal University (Natural Science), 24, 4 (2011) 27-33.

[3] Y. Liu, D. C. Pi, C. M. Chen, Similarity measurement based on key points of time series with different length, Computer Engineering & Applications, 50, 20 (2014) 1-4.

[4] Q. Liu, S. Li, Y. Fang, et al. An Effective Similarity Measure Algorithm for Time Series Based on Key Points, International Conference on Intelligent Human-Machine Systems and Cybernetics, (2016) 17-20.

[5] E. Keogh, S. Kasetty, On the need for time series datamining benchmarks: a survey and empirical demonstration. Data Mining and Knowledge Discovery, 7, 4 (2003) 349-371.

[6] S. Yang, Y. Zhang, Key point based data analysis technique, ICIC, (2007) 444-455.