Available online at www.sciencedirect.com

SciVerse ScienceDirect

Physics Procedia 25 (2012) 1097 – 1103

Physics Procedia

2012 International Conference on Solid State Devices and Materials Science

A New Method for Piecewise Linear Representation of Time Series Data

Jiajie Zhou,Gang Ye,Dan Yu

State Key Laboratory of Software Development Environment Beihang University Beijing,100191 ,China

Abstract

In various methods of modeling of time series, the piecewise linear representation has the advantage of being simple, straightforward and supporting dynamic incremental update of time series. This paper proposed a new method of Piecewise Linear Representation of Time Series based on Slope Change Threshold (SCT). Detailed experiments on real datasets from various fields show that STC representation, compared with several other Piecewise Linear Representations, can be easily calculated and has a high degree of fitting.

Keywords: Time series; Piecewise linear representation; Data mining; Fitting

1 Introduction

Time series is an ordered set of being arranged in chronological order for the observations, widespread in banking, transportation, industry and other fields. Over time, these complex data are continually accumulating and the direction is toward the mass. It is a concerned problem for data analysts all the time to excavate effectively the potential knowledge from the complex and mass data. However, the time series data usually fluctuate frequently and exit a lot of noise. So data mining in the original sequence data directly will not only cost highly in the storage and computation, but also probably affect the accuracy and reliability of the data mining algorithms.

Therefore, many time series models are proposed, which can transform original series to new series. Modeling may not only compress the data, but also keep the main form and ignore fine changes. Accordingly, it can help improve the efficiency and accuracy of the data mining algorithms, which will provide policy support for data analysts.

This paper is organized as follows: The second section describes related work for piecewise linear representation of Time Series Data. The third section introduces the thought, procedure and analysis of

doi:10.1016/j.phpro.2012.03.205

the algorithm, Piecewise Linear Representation of Time Series based on Slope Change Threshold (SCT). Experiments and evaluations of STC are given in the fourth part. The fifth part is the conclusions.

2 Related Work

At present, the frequently-used piecewise linear representation methods of time series including: Piecewise Aggregate Approximation, Piecewise Linear Representation Based on Important Point, Piecewise Linear Representation Based on Characteristic Point, Piecewise Linear Representation Based on Slope Extract Edge Point.

1) Piecewise Linear Representation Based on Important Point: Pratt and Fink[1] proposed a piecewise linear representation based on the important points. The important points are defined as the points which are the extreme points within the local scope and the ratio of the important point and the endpoint exceeds the parameters R. After extracting the important points from the time series, the algorithm then combines the points with the line orderly. Thus it will generate a new time series and get various piecewise linear representation with different fine and granularity by selecting different parameters R.

2) Piecewise Aggregate Approximation (PAA): Keogh [2] and Yi [3] proposed the method of the piecewise aggregate approximation independently. The algorithm divides the time series by the same time width and each sub-segment is represent by the average of the sub-segment. The method is simple, intuitionistic. It not only can support the similarity queries, all the Minkowski metric and the weighted Euclidean distance, but also can be used to index to improve query efficiency.

3) Piecewise Linear Representation based on the characteristic points: Xiao [4] proposed a method of piecewise linear representation based on the characteristic points. After extracting the characteristic points from the time series, the algorithm then combines the points with the line orderly. Thus it will generate a new time series.

4) Piecewise Linear Representation Based on Slope Extract Edge Point (SEEP): ZHAN Yan-Yan [5] brought forward a new piecewise linear representation combining slope with the characteristics of time series. The algorithm can select some change points according to the rate of slope change firstly, and then combines the points with the line sequentially. Finally it will generate a new time series.

The literatures above are analyzed as follows:

The piecewise linear representation gets some characteristics (e.g., extreme point, trend, etc.) of each section by segmenting the series mainly. The above methods not only have the advantages of simple and intuitive, but also can support dynamic incremental updates, clustering, fast similarity search, and so on. But the cost and fitting error is different.

3 Sct Algorithm

3.1.Basic Idea

This paper referenced the idea of the geometric slope in SEEP and proposed a new method of Piecewise Linear Representation of Time Series based on Slope Change Threshold (SCT).Firstly, the algorithm calculates the two segments' slope of the certain point connecting with the two adjacent points(except the two endpoints of time series).Secondly, it determines the change points by the ratio of slope. And then it combines the points with the line orderly. In this way, a new time series arises.

The key of the algorithm is determining the change points. The change points must follow the following principles:

1）The first point and last point are both change point;

2）When the slope of the line combining the certain point with its left neighboring point is zero, we look on the point as change point if the slope of the line combining the certain point with its right neighboring point is out the range of(-d，+d);

3）When the slope of the line combining the certain point with its left neighboring point is not zero, we look on the point as change point if the slope ratio of two lines is beyond the range of(1-d，1+d). The two lines refer to the line which combines the certain point with its right neighboring point and the line which combines the certain point with its left neighboring point.

Above d is a threshold parameter.

As shown in Figure 1, once A, B are both settled, the third point will fall on the line l parallel to the vertical axis. After inputting the threshold parameter d, we can calculate C 'and C". if the third point falls on the line segment C'C", we look on the point as change point. Otherwise, we believe that it is change point.

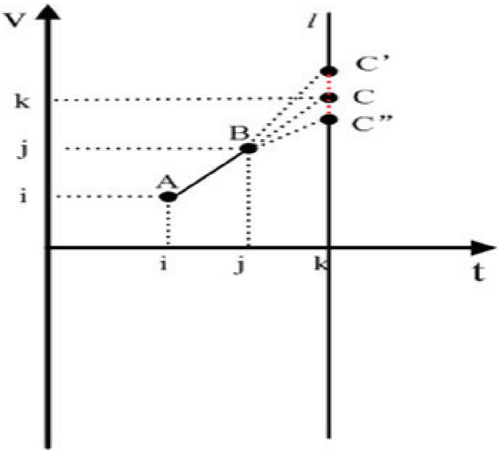


Figure.1 change point definition

3.2 Algorithm Description

Algorithm name: Piecewise Linear Representation of Time Series based on Slope Change Threshold (STC)

Algorithm input: time series X=< x_1 =(t_1，v_1)，x_2 =(t_2,v_2)，…，x_n =(t_n,v_n)>，parameter n, d (n for the length of the original time series data, d for a threshold value parameter)

Algorithm output: STC representation of time series

Algorithm Description:

```
ii=O；jj=1；kk=2；
$xx=\{(x_1，1)\}$；  //the first point of time series is change point
for($i$=1；$i<n$-1；$i$++)
  $l_1=(x[jj]-x[ii])/(jj-ii)$；$l_2=(x[kk]-x[jj])/(kk-jj)$；
if($x[ii]=x[jj]$)   //the connection slope of the first two point is zero
if(($l_2>=d$) || ($l_2<=-d$))
```

```
        then   $xx = xx + \{(x_i，i)\}$；   //the point $x_i$ join the set of $xx$
      $ii = i$；  $jj = i + 1$；  $kk = i + 2$；
  else  $jj = i + 1$；  $kk = i + 2$；
  else      / / the connection slope of the first two point is not zero
      if(((($l_2 / l_1$ )>=(1+$d$ )) || (($l_2 / l_1$ )<=(1-$d$ )))
        then  $xx = xx + \{(x_i，i)\}$；    // the point $x_i$ join the set of $xx$
      $ii = i$；  $jj = i + 1$；  $kk = i + 2$；
  else  $jj = i + 1$；  $kk = i + 2$；
    $xx = xx + \{(x_n，n))$；  //the last point is change point
  output L(X)={L($x_{i_1}$，$x_{i_2}$)，L($x_{i_2}$，$x_{i_3}$)，…，L($x_{i_{k-1}}$，$x_{i_k}$)$\,|\,$($x_{i_m}$，$m$)$\in xx$}；
```

3.3 Algorithm Analysis

The algorithm can be carried out easily and used in time series online directly. All the change points can be determined through scanning once and the time complexity is only O(n), n is the length of time sequence.

4 Experiments

4.1 Datasets

In the experiment, we compare the performance of the piecewise linear representation by using the datasets of K-Data[6] and Random_walk[6]. The information of datasets is shown in Table I.

Table I dataset description

Dataset Name	**Length(lines)**	**Dataset Name**	**Length(lines)**
Burst	9382	Ocean	4096
Chaotic	1800	Powerplant	2400
Darwin	1400	Speech	1021
Earthquake	4097	Tide	8746
Leleccum	4320	Sunspot	2899
Randomwalk	65355		

4.2 Evaluations

We use the compression ratio and fitting error to evaluate the performance of piecewise linear representation.

1） *Compression Ratio*

Assuming the time series X=< x_1，x_2，…，x_n > to be exiting，we can get a new time series X=<x_1'，x_2'，…，x_n'> after STC transformation ，where x_1'= x_1, x_n'= x_n . The compression ratio of time series is shown in formula (1).

$$\eta=(1-\frac{n'}{n})\times 100\% \quad (1)$$

2）*Fitting Error*

Assuming the time series X=< x_1，x_2，…，x_n> to be exiting，we can get a new line segment representation L(X)=<L (x_{i1}，x_{i2})，(x_{i2}，x_{i3})，…，(x_{ik-1}，x_{ik})> after STC transformation，where L(·,·) is the line segment connecting two points. And then applying linear interpolation to L(X), we can acquire the time series X^c=<x_1^c，x_2^c，…，x_i^c >. The fitting error between the original series and the new series is shown in formula (2).

$$E=\sqrt{\sum_{i=1}^{n}(x_i-x_i^c)^2} \quad (2)$$

4.3 Experimental Methods

Zhan Yan-Yan[5] experiments show that SEEP is simpler, a higher degree of fitting ,comparing with other piecewise linear representation methods. So our experiments mainly compare the performance of SEEP and STC from the following two aspects.

1)At same compression ratio, we compare the fitting error of different data sets;

2)At different compression ratio, we compare the fitting error of the same dataset applying the two algorithms.

Because the data sets are from different fields, sequence values varied widely. In order to facilitate comparison, we standardize time series firstly. The sequence values will be normalized to [0,1].

Normalization is shown in the formula (3).

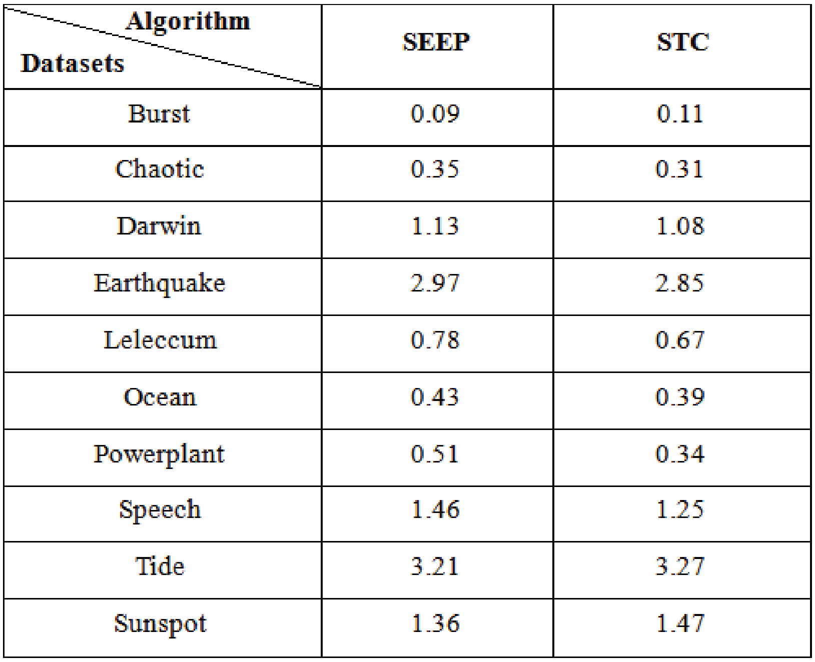
$$\mathrm{norm}(x_i)=\frac{x_i-\min(X)}{\max(X)-\min(X)} \quad (3)$$

Where, min (X) represent the minimum value of time series and max (X) represents the maximum of time series.

4.4 Result

1）At the same compression rate η = 75%, we compared the fitting error after applying SEEP and STC. The results are shown in Table 2.

Table II fitting error at the same compression rate



Algorithm Datasets	SEEP	STC
Burst	0.09	0.11
Chaotic	0.35	0.31
Darwin	1.13	1.08
Earthquake	2.97	2.85
Leleccum	0.78	0.67
Ocean	0.43	0.39
Powerplant	0.51	0.34
Speech	1.46	1.25
Tide	3.21	3.27
Sunspot	1.36	1.47

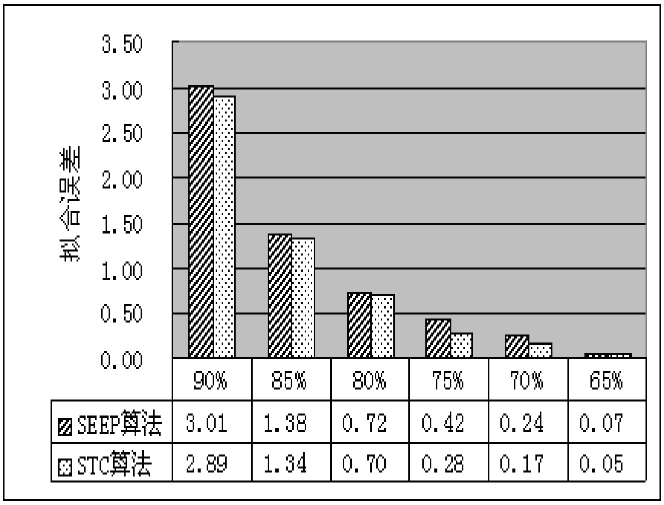
Table II shows that there are seven time series that the fitting error after applying the STC algorithm is smaller than the one after applying the SEEP algorithm in ten time series. And the fitting error after applying the two algorithms is near in other three time series.

We also see the fitting error of the two algorithms is near for the time series that the change of slope is concentrated. And, the fitting error of the STC algorithm is smaller than the one of the SEEP algorithm for the time series that the change of slope is larger.

2) At the different compression rate, we compared the fitting error after applying SEEP and STC to the same data set.

We used the data set Random_walk and the compression rate was set 90%, 85%, 80%, 75%, 70% and 65% respectively. The results are shown in Figure 2.

Figure 2 shows the comparative results after applying the two algorithms to the same dataset. As can be seen from the above figure, the fitting error of the two algorithms is reducing with the lower compression rate. And the fitting error of the STC algorithm is smaller than the one of the SEEP algorithm at the same compression rate.



	90%	85%	80%	75%	70%	65%
▨SEEP算法	3.01	1.38	0.72	0.42	0.24	0.07
▒STC算法	2.89	1.34	0.70	0.28	0.17	0.05

Figure.2 fitting error for the same dataset

5 Conclusion

Effective representation method of time series can improve efficiency and accuracy of time series data mining. In this paper, we proposed a new method of Piecewise Linear Representation of Time Series based on Slope Change Threshold (SCT). Experiment results show that the algorithm is simple, a high degree of fitting and adaptive.

Acknowledgment

We thank Keogh **(**University of California) for providing the experimental data sets from different fields.

References

[1]Prat K B, Fink E．Search for patterns in compressed time series [J]. International Journal of Image and Graphics.2002,2 (1) , pp. 89-106

[2]Keogh E J, Chakrabarti K, Pazzani M J. Sharad Mehrotra. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases [J]. Knowl．Inf．Syst，2001,3(3), pp. 263-286

[3]Yi B K. Faloustsos C. Fast Time Sequence Indexing for Arbitrary Lp Norms [C]. In: Proceedings of the 26th International Conference on Very Large Data Bases, San Francisco: Morgan Kaufmann Publishers Inc, 2000, pp. 385-394

[4]Xiao Hui, Feng Xiao-Fei, Hu Yun-Fu. A new segmented time warping distance for data mining in time series database[C]. In：Proceedings of 2004 International Conference on Machine Learning and Cybernetics，Shanghai, China, 2004, pp. 1277-1281

[5]ZHAN Yan-Yan, XU Rong-Cong, CHEN Xiao- Yun. Time Series Piecewise Linear Representation Based on Slope Extract Edge Point[J]. Computer Science, 2006,33(11), pp. 139-161.

[6]Keogh E, Folias T. The UCR Time series Data Mining Archive．http://www.cs.ucr.edu/～eamonn/ time_series_data/.Irvine, CA：University of California, Department of Information and Computer Science, 2002.