MOTIFS DISCOVERY FOR STREAMING TIME SERIES

Qi Zhang, Yang Gao, Jiecai Zheng, Lin Chen, Xueqing Li

School of Computer Science and Technology, Shandong University, Jinan, China

{zhangqisdu, gaosdu}@126.com, {jc Zheng, chenlin, xqli}@sdu.edu.cn

Abstract

The motif discovery approach is used to measure the correlation of the pair of consecutiveness in time series, which also aims to find all subsequences which are similar to the given one. However, along with the arrival of Industry 4.0 era, massive numbers of detecting instruments in various fields are continuously producing a plenty number of time series streaming data, the high dimensionality and continuousness of streaming time series give rise to the potential threat for searching effectiveness. For these reasons, we come up with a novel motifs discovery approach for streaming time series based on piecewise linear representation with turning points and skyline index. As the experimental results suggest, our approach is more effective than some other traditional methods.

Keywords: Motifs, Online Segmentation, Indexing and Retrieval, Time Series, Data Mining

1. INTRODUCTION

Along with the coming of industry 4.0 era, the evolving of IoT (Internet of Things) has stimulated the deployment of massive numbers of detecting instruments (sensors etc.) in various industry fields, which are continuously producing a plenty number of time series data (e.g., the operational data of the satellite on orbit, the real-time testing data from detecting instruments etc.). Accordingly, the underlying value in time series data are discovered by data mining techniques. Such as motifs discovery revealing the regularity and uncertainty of time series. Since the streaming time series is continuously generated in a high speed and potentially forever (Luo,2015), it is difficult to do in-depth motif discovery for the streaming time series.

Motifs can be defined as kinds of "frequently occurring patterns in time series" and motifs discovery is a similarity searching approach for searching all of the subsequences which are similar to the given one in a time series object. Figure 1 illustrates motifs discovered in an astronomical database (Lin,2002).

Figure 1. An astronomical time series contains 4 near identical subsequences

While the discovery of the motifs for streaming time series, several vital issues need to be addressed: high dimensionality, continuity of the data and the efficiency of similarity calculating (Li,2013). Furthermore, we need to consider the optimization of the spatial indexes.

At present, most existing motifs discovery approaches do not work well for our problems. Some methods process on static data sets, and may incur a high cost if applied directly to large time series data. There do exist works that address the high-dimensional problem for time series data. For

instance, (SonNT,2015) propose a finding motifs method by using the Middle Points and Clipping (MP_C) and skyline index. The MP_C method can represent the sequence by some kinds of middle points, which is inevitably reducing the accuracy compared with the local extreme points of the streaming time series. To solve this problem, (Hu,2016) propose a motif discovery approach based on piecewise linear representation with important data points (PLR_IDP) which extracts the "turning points" of the given database and retains all of the temporal features during the process of dimensionality reduction for large time series.

In our paper, we propose a novel motifs discovery approach for streaming time series. First of all, we use an online piecewise linear segmentation to extract the "turning points" of the streaming time series, which retains all these important points for dimensionality reduction in large database. What's more, we define the lower bounding region for similarity measure and prove that our proposed region does not introduce any false dismissals during the searching. Last but not least we utilize the skyline index (Li,2004) for improving the efficiency of the motifs similarity searching.

Contributions of our paper are listed as follows:

1. We propose a novel motifs discovery approach with the continuous segmentation algorithm (Hu,2016) and skyline index.
2. We make use of skyline index for indexing and searching time series data.
3. The experimental results authenticate the effectiveness of our approach.

The rest of the paper is organized as follows. Section 2 introduces the related works. The motifs discovery approach is described in section 3. The experiment results on a real world industry time series database are presented in section 4. Section 5 concludes this paper.

2. RELATED WORKS

Many approaches ((Mueen,2009), (Chiu,2003)) for motif discovery considering the motif length n and the similarity threshold r . That is

Figure2. The major steps of the continuous segmentation based on turning point

to say, given a time series D and a query time series Q , the K-motif discovery algorithm is aim to retrieve those time series $S \hat{=} D$, whose similarity calculation can be transform into a distance measure of those sequences as follow:

Since the serious dimensionality curse problem in high dimensional space, it becomes infeasible to directly index time series using traditional spatial indexes. Specifically, when the dimensionality becomes higher, the performance of the similarity searching with multidimensional index can be even worse than that of a linear scan.

For this reason, researchers have concentrated their focus to data representation and indexing techniques to improve the efficiency of similarity searching.

Piecewise linear representation (PLR) (Keogh,2001) is more consistent with human visual experience than other time series data representation methods, which also has lower index dimension, faster calculation speed for similarity searching. However, just like a coin has two sides, the PLR methods do exist some shortcomings, such as the fitting errors of single point, local or whole error is too large (Zhou,2008). To improve the accuracy of the representation, (Ji, 2016) considered the global fitting error of the time series, as well as the fitting error of the single point in a certain segment, which can produce more accurate segmentation and representation, but the entire data set should be scanned first. To solve this problem, our team proposed a novel online continuous segmentation and representation algorithm based on turning points (CS_TP) for streaming time series, which can partition the time series into some continuous subsequences and maintain a high similarity between the processed subsequences and the raw streaming data.

Meanwhile, the researchers for indexing technology have also made progress. (Li,2006) proposed the skyline index and (SonNT,2015) proposed an algorithm for discovering motifs in time series data, which makes use of the MP_C

dimensionality reduction method and skyline index. (Hu,2016) proposed a novel k-motifs discovery approach based on piecewise linear representation with important data points (PLR_IDP), which can retain all of the temporal features during the process of dimensionality reduction for large time series data and speed up the process of k-motifs retrieval.

3. METHODOLOGY

In this section, we make a detailed description on how to discover k-motifs for streaming time series. Our approach can be divided in the following three major parts: continuous segmentation based on turning points (CS_TP), skyline index for CS_TP and k-motifs discovery algorithm.

3.1 Continuous segmentation based on turning points

CS_TP is an online time series data segmentation and representation method, which considers not only the continuously generated characteristics but also the accuracy of the streaming time series.

the definition of TP (turning point) is given as follow:

1. For a time series $T = (\dots, a_i, \dots, a_j, \dots)$ where $a_i = (x_i, y_i)$ is a raw data point in time series. If a_i meets one of the following two inequations, it can be defined as the Level 1 TP:

$$\frac{y_i - y_{i-1}}{x_i - x_{i-1}} < \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \text{ or } \frac{y_i - y_{i-1}}{x_i - x_{i-1}} > \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \quad (1)$$

$$\frac{y_i - y_{i-1}}{x_i - x_{i-1}} < \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \text{ or } \frac{y_i - y_{i-1}}{x_i - x_{i-1}} > \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \quad (2)$$

2. In order to eliminate noise interference, we will continue to optimize the coarse selection of the Level 1 TPs and to smooth the fitting curves with parameter μ and ρ , which can be specified by users. For a time series $T = (\dots, a_i, \dots, a_j, \dots)$ where $a_i = (x_i, y_i)$ and $a_j = (x_j, y_j)$ are TPs in

the level 1. If a_j meets the following condition, it can be defined as the Level 2 TP:

$$\frac{|p_{j-1} - p_j|}{|p_{j-1} - p_j| + |p_j - p_{j+1}|} \geq \frac{|p_{j-1} - p_j|}{(|p_{j-1} - p_j| + |p_j - p_{j+1}|)/2} \geq \text{thres} \quad (3)$$

To ensure the effectiveness and the efficiency of the segmentation algorithm, we introduce two segmentation criteria for a potential segment, which are the maximum error for single point (ME_SP) and the maximum error for entire segment (ME_ES). The CS_TP algorithm can utilize the TPs in level 1 and ME_SP for initial segmentation (shown as from Figure2(a) to Figure2(b)) and the TPs in level 2 and ME_ES to evaluation the segmentation error (shown as from Figure2(b) to Figure2(c)). After the above two steps, CS_TP can merge all of the subsequence by standing on the holistic view (shown as from Figure2(c) to Figure2(d)).

3.2 Continuous segmentation based on turning points

After the Piecewise Linear Representation method CS_TP for dimensionality reduction, a group of time series data exist into a two-dimensional time-value space. The minimum bounding rectangle (MBR) defined in this space, which is used to bound a group of sequence data by the traditional R-tree index, whose overlaps are too large to affect the search efficiency.

So as to minimize the overlaps between MBRs, we adopt to use skylines to bound a group of time series data. However, the straightforward adoption of the bounding regions defined by raw data is impractical, due to its representation, which is too costly for long time series searching.

For this reason, only the TPs selected by CS_TP should be stored in an index structure for each MBR and the regions defined by the TPs should entirely enclose the original raw data bounding regions. To meet this goal, we utilize the skyline to bound a group of time series data, which can be called CS_TP Bounding Region (CS_TP_BR), is defined as follows:

There is a group T of time series n of length m . $T = (t_1, t_2, \dots, t_n)$, the CS_TP_BR of T specifies a region surrounded by top and bottom skylines. The CS_TP_BR R of T is described as follow:

$$R = (\text{top_skyline}, \text{bottom_skyline}) \quad (4)$$

where

$$\text{top_skyline} = \{ \text{top_skyline}_1, \text{top_skyline}_2, \dots, \text{top_skyline}_m \} \quad (5)$$

$$\text{bottom_skyline} = \{ \text{bottom_skyline}_1, \text{bottom_skyline}_2, \dots, \text{bottom_skyline}_m \}$$

and, for $1 \leq i \leq m$

$$\begin{aligned} \text{top_skyline}_i &= \max\{p_1, p_2, \dots, p_n\} \\ \text{bottom_skyline}_i &= \min\{p_1, p_2, \dots, p_n\} \end{aligned} \quad (6)$$

We use the CS_TP_BR to lower bound the distance between a query object and a group of time series data.

Formally, given a query q and an CS_TP_BR R , the lower bounding distance function $D_{\text{CS_TP_BR}}()$ is defined as follows:

$$D_{\text{CS_TP_BR}}(q, R) = \sqrt{\sum_{i=1}^m (\text{top_skyline}_i - \text{bottom_skyline}_i)^2} \quad (7)$$

The $D_{\text{CS_TP_BR}}()$ distance function can satisfy the lower bound property of the whole time series. The convergence relation between the query q and the PLR_IDP_BR R which contains a group of time series data x can be described as follow:

$$D_{\text{CS_TP_BR}}(q, R) \leq \|q - x\|, \forall x \in R \quad (8)$$

For any $x \in R$, the value of $x[i]$ is between $T_{\max}[i]$ and $T_{\min}[i]$. That is to say,

$$\text{Dist}(q, x) \leq \|q - x\| \quad (9)$$

According to the above deduction, not only can we determine the lower bound property of the $D_{\text{CS_TP_BR}}()$ distance function, but we can also use the CS_TP_BR representation in a multidimensional index to improve the efficiency of the similarity searching.

3.3 Indexing and Retrieval

In previous section of the paper, we have proved that the lower bound property of $D_{\text{CS_TP_BR}}()$ can guarantee no-false-dismissals for similarity searching and the feasibility of indexing. The main steps of indexing building are presented as follow:

Step1. Find a proper leaf node for inserting the new CS_TP sequence.

Step2. When the leaf node is full, we should split the leaf node and transmit the change upward.

Step3. According to these changes, we should adjust CS_TP bound regions and propagate the transformation upward again as necessary.

Step4. If the continued upward propagation causes the root to split, we should create a new root containing the two resulting nodes.

After the accomplishment of indexing building, we propose our motifs discovery algorithm. First of all, streaming time series will be segmented and transformed into lower dimensionality by CS_TP method. Then for each segment s_i , the algorithm finds all its non-trivial matches within a range R and inserts these matches into the skyline index.

The algorithm is described as follows:

Input: time series data $T = (t_1, \dots, t_n)$, t_i is a subsequence of T

Output: K_{ml} is a list of k -motifs

Variables:

M_List : the list of non-trivial matching subsequence.

ϵ : threshold for matching

CS_TP_BR: The region of the lower bound.

Algorithm: K-Motifs retrieving based on Skyline Index

```

Begin
//ti should be transformed into tp by using CS_TP
CS_TP_BR ← Calc_BR(tp, tpn);
While (Skyline index I ≠ null)
M_List = Find_Similar_Neighbors(tp, CS_TP_BR, ε)
if (M_List == null) break;
else
for (i=1; i ≤ M_List.length; i++)
checkForNon_trival(M_List, K ml);
End for
if (K ml.length < k)
Insert M_List into K ml in decreasing order
else if (K ml.length > M_List.length)
Remove the kth value of K ml and Insert M_List
end while
Update_Skyindex();
//propagate the change upward
End Function
    
```

4. EXPERIMENTAL EVALUATION

In this section, we evaluate the performance of our approach: CS_TP&Skyline based on CS_TP and Skyline index. The assessment can be divided into two steps:

1. We compare CS_TP method with PLR_IDP to determine effect of data representation.
2. We compare our approach with two popular algorithms: Random Projection(RP) and PLR_IDP with r-tree index (PLR_IDP&R), which is nearly the latest time series motif discovery proposal up to date. We are concerned not only the accuracy of these techniques, but also the efficiency by using different datasets, which contain different sizes and different lengths of motifs.

4.1 Dataset and Evaluation Metric

Table 1. Average Representation Error of Different Methods

threshold	CS_TP		PLR_IDP	
	global error	number of segments	global error	number of segments
20	270.51	1351	199.83	1436
50	440.86	1257	380.76	1287
100	1093.79	1138	896.17	1215
150	2305.07	906	2089.64	1102
200	3927.19	789	3578.01	993

According to the experiment results, the CS_TP has better pruning power (less number of segments) than the PLR_IDP which has higher accuracy in segmenting time series. However, when evaluating the different segmenting algorithms for streaming time series, we focus more on the efficiency of segmentation rather than the accuracy of

We select some kinds of typical industry time series datasets, which include Electrocardiograph dataset (ECG) provided by the UCR homepage (Keogh,2006), the monitoring data of Jinan municipal steam heating system(JMSHSD), the monitoring data of Dong Fang Hong satellite (DFHD).

In our experiment, we use the pruning power p (Keogh,2001) to determine the effectiveness of the dimensionality reduction. The pruning power p can be defined as follow:

$$p = \frac{\sum_{tp} \sum_{all}}{\sum_{all} \sum_{tp}} \quad (10)$$

Where $\sum_{tp}$ denotes the number of subsequence which should be used, $\sum_{all}$ is the number of subsequence in dataset. The range of p is from 0 to 1, that is, when the value of p is lower, the pruning power is better.

The running times and the efficiency-ratio (Li,2012) are also used to evaluate the efficiency of the k-motifs discovery algorithm. The Efficiency-ratio is defined as follow:

$$\text{efficiency - ratio} = \frac{T_a}{T_b} \quad (11)$$

Where T_a denotes the time consumed by the Euclidean distance function which is called by the proposed algorithm. T_b is the time consumed by the Euclidean distance function by the brute-force algorithm. Similar to the above analysis, the range of the efficiency ratio is from 0 to 1, the k-motifs discovery algorithm with lower efficiency ratio is better.

4.2 Performance of the Experiment

Evaluating the pruning power in similarity search

In these experiments, the number of the data items is no less than 3000. When the fitting error threshold is set from 20 to 200, the average representation error and the average number of segments have been shown as follow in Table1

the segments. In other words, the real-time capability is more important in streaming time series segmentation and representation.

In Figure 3, we present the pruning powers comparison between CS_TP and PLR_IDP with different datasets.

Figure3. The pruning powers of CS_TP and PLR_IDP based on different datasets

Evaluating the efficiency in K-motifs discovery

We compare our approach: CS_TP&Skyline with RP and PLR_IDP&R, from the running times and Efficiency-ratio in Figure 4 and Figure 5. All of the experiments compare the three algorithm with ECG, JMSHSD and DFHD which contains no less than 50,000 data records.

Figure4(a) presents the running times of the three algorithms. Figure4(b) highlights the performance results of the CS_TP&Skyline and the PLR_IDP&R.

Figure5(a) shows the efficiency-ratio of the three algorithms. Figure5(b) highlights the efficiency results of the two approaches: CS_TP&Skyline and the PLR_IDP&R.

Figure4. The running times of these algorithms on different datasets

Figure5. The Efficiency-ratio of these algorithms on different datasets

Through the analysis and experiment results, we can see that CS_TP&Skyline and PLR_IDP&R are more efficient than RP. However, CS_TP&Skyline is more efficient than PLR_IDP&R.

The high performance of our approach can be summarized as two critical factors as follow:

1.the online continuous segmentation strategy

The CS_TP algorithm can partition and represent the streaming time series synchronously with the receiving data,

in other words, this kind of online segmentation algorithm is no need for waiting the completion of the scanning of the entire data set compared with the PLR_IDP algorithm.

2. the index optimization by skyline

the similarity search for matching motifs using skyline index, which is more effective than R-tree index by its less overlaps between MBRs. The overhead of the computation cost in our method is cheaper than that in random projection algorithm, which has to repeat the random projection so many

times before obtaining convergent results that computational cost is increased.

5. CONCLUSIONS

In this paper, we propose a motifs discovery approach for streaming time series, which improves the efficiency for similarity searching by considering the dimensionality reduction, continuous segmentation and the optimized index for streaming time series. Experimental results show the effectiveness of our approach.

Finally, due to the high data volume of streaming time series, the maximum capacity of dataset processed in our approach is no more than 100GB. we will attempt to address this big data problem with map-reduce in the future.

6. ACKNOWLEDGMENT

The authors would like to acknowledge the support provided by the National Natural Science Foundation of China (61402263), the National High Technology Research and Development Program of China (2014AA01A302), the special funds for Novel Software Technology project(KFKT2015B02), the Science and technology Project of Shandong Province (2015GGX 101009).

REFERENCES

- Luo, G., Yi, K., Cheng, S. W., & Li, Z. (2015). Piecewise linear approximation of streaming time series data with max-error guarantees. *IEEE, International Conference on Data Engineering*, pp.173-184.
- Lin, J., Keogh, E., Lonardi, S., & Patel, P. (2002). Finding Motifs in Time Series, pp.53-68.
- Hai-Lin, L. I., & Guo, C. H. (2013). Survey of feature representations and similarity measurements in time series data mining. *Application Research of Computers*, 30(5), pp.1285-1291.
- Yi, B. K., & Faloutsos, C. (2000). Fast Time Sequence Indexing for Arbitrary Lp Norms. *Proceedings of the 26th International Conference on Very Large Data Bases*, pp.385-394.
- Hu, Y., Ji, C., Jing, M., & Li, X. (2016). A K-Motifs Discovery Approach for Large Time-Series Data Analysis.
- Ji, C., Liu, S., Yang, C., Wu, L., Pan, L., & Meng, X. (2016). A piecewise linear representation method based on importance data points for time series data.
- Li, Q., Moon, B., & López, I. F. V. (2004). Skyline index for time series data. *IEEE Transactions on Knowledge & Data Engineering*, 16(6), pp.669-684.
- Hu, Y., Ji, C., Jing, M., & Li, X. (2016). A Continuous Segmentation Algorithm for Streaming Time Series, *Proceedings of IEEE ICPADS*
- Mueen A., Keogh E., Zhu Q., Cash S., Westover B. (2009). Exact discovery of time series motifs. In *Proceedings of SIAM international conference on data mining*, pp. 473-484
- Chiu, B., Keogh, E., & Lonardi, S. (2003). Probabilistic discovery of time series motifs. *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* pp.493-498.
- Keogh, E., Chakrabarti, K., Pazzani, M., & Mehrotra, S. (2001). Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, 3(3), pp.263-286.

Zhou, D. Z., & Min-Qiang, L. I. (2008). Time series segmentation based on series importance point. *Computer Engineering*, 34(23), pp.14-16.

Keogh E, Zhu Q, Hu B, Hao Y, Xi X, Wei L, Ratanamahatana CA.(2006).The UCR time series classification/clustering homepage.

Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2001). An online algorithm for segmenting time series. *ICDM*, pp.289-296.

Y. Li, J. Lin, and T. Oates. (2012). Visualizing variable-length time series motifs. In *SDM*, pp. 895-906.

Authors

Qi Zhang is a PhD student in the school of computer science and technology, Shandong University. He specializes in data mining, machine learning, business intelligent systems, and their applications.

Yang Gao is a graduate student in the school of computer science and technology, Shandong University. He specializes in research and development of business intelligence systems, data mining, machine learning and their applications.

Jiecai Zheng is a PhD in the school of computer science and technology, Shandong University. He specializes in research and development of business intelligence systems, data integration, machine learning and their applications.

Lin Chen is a PhD in the school of computer science and technology, Shandong University and works in the office of information technology, Shandong University. He specializes in research and development of business intelligence systems, IT operation and maintenance, data governance, human computer interaction and virtual reality.

XueQing Li is a professor and vice-president in the school of computer science and technology, Shandong University. He specializes in research and development of business intelligence systems, human computer interaction and virtual reality, image processing, machine learning and their applications.