Comparative Analysis of Motif Discovery Methods in Time Series

Sukriti Parnami* and Veenu Mangat**
* M.E. IT Research Scholar, UIET, Panjab University, Chandigarh
sukritiparnami@yahoo.com
** M.E. Supervisor, Assistant Professor, IT, UIET, Panjab University, Chandigarh
veenumangat@yahoo.com

Abstract: Due to wide use of Information Technology, substantial amount of information is being gathered for exploratory analyses, business operations and online networking in the big data era. Due to the gathering of information for several events at distinctive time periods, huge datasets are formed. A time series can be defined as a series of numeric values obtained at various points, occurring after regular intervals. An interesting research problem in time series is motif discovery. Motif discovery subroutine can be utilized as a part of algorithms for classification and summarization. In this paper, comparative analysis of various motif discovery methods in time series datasets has been done and a method for improving the same has been suggested.

Keywords: Data mining, motifs, match, subsequences, time series.

Introduction

Due to wide use of Information Technology, substantial amount of information is being gathered for exploratory analyses, business operations and online networking in the big data era. Due to the gathering of information for events at distinctive time periods, huge datasets are formed. A time series can be defined as a series of numeric values obtained at various points, occurring after regular intervals. E.g.: time series sequence may represent network flow, exchange rates or weather conditions over time. The data is substantial in size, has high dimensionality and needs to be updated after continuous period of time.

A major reason for representation of time series data is to minimize the dimension of original data. The most common method is to sample the data. However, if the rate is low, this method has the disadvantage of disfiguring the shape of time series sampled. Another technique is to utilize the mean estimation of every fragment for representation of the set of data points. For a given time series $X = (x_1, x2_{,....},x_m)$ and n, reduced dimension, the compressed time series $Q = (q_1,q_2,\ldots.q_n)$ after dimensionality reduction can be obtained by

$$q_k = \frac{1}{e_k - s_{k+1}} \sum_{i=s_k}^{e_k} x_i \qquad \text{(i)}$$

where the starting and ending data points are denoted by s_k and e_k of the kth point in the time series X, respectively (Figure1.). The technique is known Piecewise Aggregate Approximation (PAA).

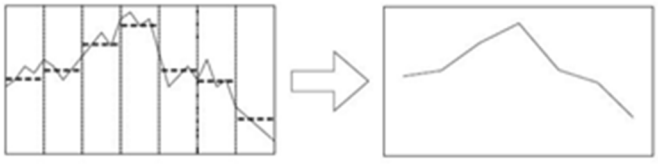

Figure 1: Dimension reduction by PAA

Another approach for representation of time series is to convert the numeric data in time series to symbols. First discretization of the time series into fragments is done, and then each fragment is converted into a symbol. A common technique for this is symbolic aggregate approximation (SAX) the reduced time series from PAA is converted to string of symbols. The y-axis is divided into equiprobable regions. Symbols are used to represent regions and every segment is mapped to a symbol based on the region in which it lies. The compressed series, q utilizing PAA is changed over to a symbol string S $(s_1,\ldots,s_w)$.

Based on the representation of time series, various mining tasks can be found and can be grouped into fields such as clustering and pattern discovery, classification, rule discovery and summarization [1]. Among these, detection of previously

unknown, frequently occurring patterns is more interesting problem. Such patterns are called motifs. In Figure2. An example of motif discovered in an electroencephalogram (EEG) time series dataset is shown. An efficient motif discovery algorithm can be used as a tool for summarizing and envisioning enormous time series databases. Likewise, it can be utilizes as a subroutine as a part of different mining tasks, for instance:

- Motif discovery is required for the discovery association rules. Motifs are mentioned as primitive shapes in [2] and frequently occurring patterns in [3].
- A few algorithms for classification work by developing prototypes of every class [4, 5]. These prototypes may be viewed as motifs.
- Many time series inconsistency/interestingness discovery algorithms basically comprise of demonstrating normal behavior with a set of motifs, and identify future patterns that are not at all like these motifs [6].
- In robotics, a method has been introduced by Oates et al. [7], to permit an autonomous agent to sum up from a set of qualitatively diverse experiences gathered from sensors. These "experiences" are seen as motifs.
- Much work on discovering estimated occasional patterns in time series can be considered as an effort to find motifs that happen at compelled intervals [8].

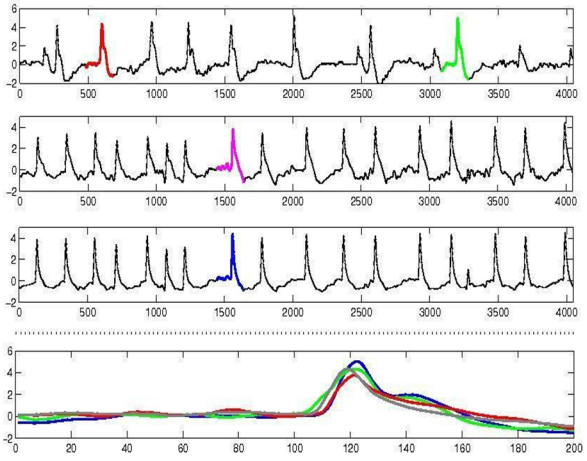

Figure 2: Example of motifs in their original context (up), and in the same referential for similarity observation (down) for EEG time series data

Definitions and Notations

Time Series
A time series can be defined as an ordered list T = $\{t_1,t_2,\ldots t_m\}$ of real-valued variables, where m represents the length of time series.

Subsequences
For a given time series T of length m, a subsequence $T_{i,k}$ of T is a time series of length k < m, which starts from position I, i.e., $T_{i,k} = \{t_i,t_{i+1},\ldots,t_{i+k-1}\}$, $1 \leq i \leq m-k+1$.

The distance between two subsequences is often compared by Euclidean Distance (ED).

Euclidean Distance
For two time series (or subsequences) $X=\{x_1,x_2,\ldots x_n\}$ and $Y=\{y_1,y_2,\ldots y_n\}$ both of length n,
Euclidean Distance (ED) between them is given by:

$$D(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad \text{(ii)}$$

Non-Trivial Match
For a given time series T containing two subsequences $S_{i,k}$ and $S_{j,k}$ with the length k, If their distance $D(S_{i,k}, S_{j,k})$ is smaller than the specified threshold r, then $S_{j,k}$ is the matching subsequence of $S_{i,k}$,
If the two subsequences begin at essentially different positions, they are said to be a nontrivial match. We say that the starting positions are essentially different if there exists x, such that. i < x <j and the distance D $(S_{i,k}, S_{x,k})$ is larger than r.
Counting of trivial matches should be excluded by the algorithms.

K motifs
For a given time series T, subsequence length n,range r, the most significant motif in T (known as 1-Motif) is a subsequence

S_1 that has the highest count of non-trivial matches. The most significant motif in T (called K-Motif) is the subsequence S_k having the highest count of non-trivial matches and satisfies D (S_k,S_i)> 2r, for all $1 \leq i \leq K$.

Motif Type: Motif subsequence is a subsequence that matches the motif type within a time series *T*. A motif type often has multiple motif subsequences.
Taking the CBF dataset as an example, three similar subsequences of a particular motif type are detected, as seen in Figure3.

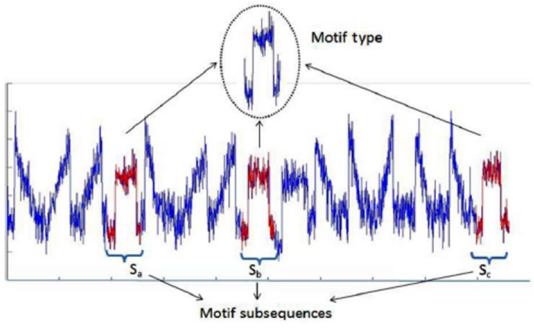

Figure 3: Motif type and three subsequences S_a, S_b and S_c

Various Motif Discovery Methods

Primarily motif discovery can be divided into exact and approximate methods. Exact methods are able to find motif exactly but are often inefficient while dealing with larger dataset. Approximate ones map segmentations into low dimensional words to reduce their computational complexity and execution time.

Exact Motif Discovery Method

J. Lin et al. [9], the paper introduce a Brute-Force (BF) algorithm for exact discovery of motifs, which require comparisons quadratic in the length of time series. For a subset t, of size n in time series BF compares it with each subset s in subsequence S to find the closest match. Its time complexity is O (m^2), where m represents the length of the time series.

Table 1: Brute Force Algorithm

Algorithm	Brute Force Motif Discovery
Procedure	$[L_1,L_2]$=*BruteForce_Motif* (*D*)
in: *D*:	Database of Time Series
Out: L_1,L_2:	Locations for a Motif
1	*best-so-far* = INF
2	**for** *i* = 1 to *m*
3	**for** *j* = *i*+1 to *m*
4	**if** $d(D_i,D_j) <$ *best-so-far*
5	*best-so-far*=$d(D_i,D_j)$
6	L_1=*i*, L_2=*j*

Mueen-Keogh (MK) [10], a tractable exact algorithm, faster than BF by using linear ordering of data to provide useful information to guide the search for motifs. Worst case of the algorithm is quadratic but can be reduced by three orders of magnitude.

Approximate Discovery Methods

The first approximate algorithm for motif discovery was proposed by Chiu et al. [11]. It uses probabilistic and iterative approach. Here time series is converted to SAX representation. For every iteration, random positions of words are selected as wildcards and whole list is traversed. For each match collision matrix entry is incremented. At the end largest entries in collision matrix are selected as motif candidates and ach candidate is checked for validity in original data.
H. Tang and S. S. Liao [12]. A novel algorithm is introduced in this paper that doesn't require exact w value, which is the length of the pattern. The idea is to keep w value relatively small to distinguish short pattern first, and then utilize a

concatenation routine to concatenate the short patterns found to generate the entire pattern .This approach improves the widely used K-motif algorithm, it is capable of discovering patterns with different lengths

Liu et al. [13], Here the issue of locating motif has been formalized as persistent top-K motif ball issue in a m-dimensional space and heuristic methodology has been suggested that can enhance quality of motifs. Comparative time series subsequences are assembled as m-dimensional data points in a ball of range r. A motif in the data set is then a thick ball after removal of the trivial matches in an m-dimensional space. The maximum radius of k ball being

$$\sqrt{\frac{1}{1/m+1}}\sqrt{2r}.$$

The above talked about methodologies are restricted to finding pattern occurrences of same length, neglecting to capture the likenesses of events being consistently scaled along the time axis. D. Yankov et al, [14]. This paper proposes identification of time series motif for uniform scaling. Here time series motif projection algorithm of [11] is extended to capture motif under uniform scaling distance d_u

Combining Exact and Approximate Methods

B. Liu et. al. [15], the paper introduced an efficient Motif Discovery method for Large - scale time series (MDLats) by combining the advantages of both exact and approximate methods. By computing standard motifs, MDLats eliminates a majority of redundant computation in related area and reuses existing information to the maximum. Figure4. Shows the main steps of MDLats

First step is the input i.e. the initial time series which is processed to obtain subsequence symbols by using SAX [16]. and PAA [17] methods. Segmentation Compression Module compresses subsequence symbols by abstracting repetitive symbols. Then standard motifs are calculated which are used by Final Motifs Discovery Module for generation of final motifs that have similar patterns. Since all motifs are included in candidates, whole motifs with different lengths from short patterns.

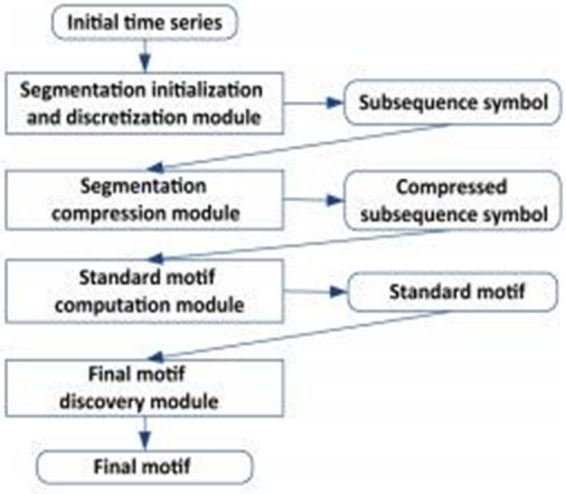

Figure 4: Main Steps of MDLats

Advantages of MDLats algorithm over the above discussed:

- It improves efficiency for motif discovery while retailing accuracy.
- By computing standard motif, it reduces majority of redundant computations and reuse existing information to the maximum by exploiting the relation between existing and newly arrived data.
- It is adaptive to different lengths of motif.
- It is scalable and is deployed in Hadoop for parallel computing.

Experimental Analysis

In this section we have compared two exact discovery methods i.e. BF [9], MK [10] and one approximate method RP [11] in terms of execution time. The results are shown in fig. 4. The dataset used in this section is constituted by random walk time series available in the MK algorithm [10] website. There are 10 sets of random walk series data, containing 10000 to 100000

time series, each of length 1024.
Note that, we have compared results with only one iteration of RP algorithm. However, as an iterative algorithm, several iterations are necessary in order to converge. Results with more than one iteration will have less execution time than the BF algorithm as BF algorithm is quadratic in nature whereas RP algorithm is linear with high constant factor. MK algorithm is up to three orders faster than BF.

Table 2: Comparison of various methods for motif discovery based on execution time

Size of Data set	BF	MK	RP
10000	83.703	3.45	53.54
20000	190.962	16.11	193.88
30000	399.765	32.58	404.41
40000	648.18	31.29	705.02
50000	965.047	49.92	1221.13
60000	1429.116	66.54	1613.53
70000	1861.126	90.03	2139.20
80000	2550	103.31	2708.53
90000	3277	108.70	3468.50
100000	4200	130.68	4357.39

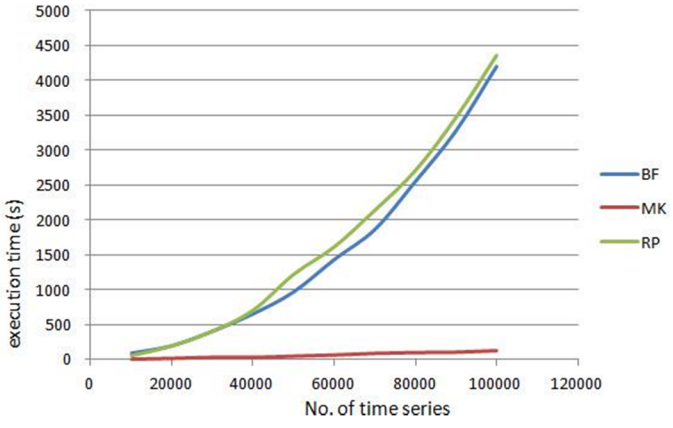

Figure 5: Execution time of the algorithms vs. No. of time series

Conclusion and Future Work

- We can improve the MDLats algorithm mentioned in [15] by optimizing the SAX representation by using algorithms like Genetic algorithm (GA), Particle Swarm Optimization (PSO).
- Motifs should be evaluated in such a way that they could be ranked in terms of their significance.
- Motifs can be used to describe the time series data in other data mining tasks like:
    - Classification
    - Abnormality Detection
    - Forecasting

Acknowledgement

The authors would like to thank UIET and Panjab University authorities for providing resources to carry out this research work.

References

[1] T. C. Fu, "A review on time series data mining" ,Engineering Applications of Artificial Intelligence, 24 (1) (2011), pp. 164–181
[2] G. Das, K. Lin, H. Mannila, G. Renganathan and Smyth, "Rule discovery from time series" in proceedings of 4th International Conference on Knowledge Discovery and Data Mining. New York, NY, Aug 27-31. pp 16-22, 1998.
[3] F. Hoppner, "Discovery of temporal patterns - learning rules about the qualitative behavior of time series", in the proceedings of fifth

European Conference on Principles and Practice of Knowledge Discovery in Databases. Freiburg, Germany, pp 192-203. 2001.
[4] E. Keogh and Pazzani, “An enhanced representation of time series which allows fast and accurate classification clustering and relevance feedback”, International Conference on Knowledge Discovery and Data Mining. New York, NY, Aug 27-31. pp 239-243
[5] M. Hegland, W. Clarke and Kahn, “Mining the MACHO dataset, Computer Physics Communications”, vol.142, pp. 22-28, 2002.
[6] D. Dasgupta and S. Forrest, “Novelty detection in time series data using ideas from immunology”, In Proceedings of the fifth International Conference on Intelligent Systems, 1999.
[7] T. Oates, M. Schmill and P. Cohen, "A Method for Clustering the Experiences of a Mobile Robot that Accords with Human Judgments", In Proceedings of the 17th National Conference on Artificial Intelligence, 2000, pp 846-851.
[8] J. Han, G. Dong and Y. Yin, “Efficient mining partial periodic patterns in time series database”, in Proc. of the fifteenth International Conference on Data Engineering, Sydney, Australia, pp 106-115,1999.
[9] J. L. E. L. S. Lonardi and P.Patel, ”Finding motifs in time series,” in Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, Edmonton, AB, Canada, 2002, pp.53-68.
[10] A. Mueen et al., “Exact discovery of time series motifs,” in Proceedings of SIAM International Conference on Data Mining ,2009,pp. 473-484.
[11] B. Chiu, E. Keogh, and S. Lonardi, “Probabilistic discovery of time series motifs,” in Proc. 9th International Conference Knowledge Discovery Data Mining, Washington, DC, USA , 2003, pp. 493-498.
[12] H. Tang and S.S. Liao, “Discovering original motifs with different lengths from time series,” Knowledge Based Systems., vol.21,pp.666-671.2008
[13] Z.Liu .et.al.,"Locating motifs in time series data ,"in Proceedings of SIAM International Conference on Data Mining,2009,pp.473-484.
[14] D. Yankov et al.,"Detecing time series motifs under uniform scaling” in Proc. 9th ACM SIGKDD International Conference Knowledge Discovery Data Mining, 2003, pp. 493-498.
[15] B. Liu, J. Li, C. Chen, W. Tan, "Efficient motif discovery for large-scale time series in healthcare "IEEE Transactions on Industrial Informatics., vol. 11, no. 3, pp. 583-590, June 2015
[16] J. Lin et al.,"A symbolic representation of time series, with implications of streaming algorithms," in Proc 8th ACM SIGMOD Workshop Res. Issues Data Mining Knowledge Discovery,2003,pp. 2-11
[17] E. Keogh et al., "Dimensionality reduction for fast similarities search in large time series databases," J. Knowledge Inf. Syst., vol. 3 pp.263-286,2001