International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Motif Discovery Comparison using Multivariate Rhythm Sequence Technique and Dynamic Time Warping (DTW) in Time Series Data

Kumar R[1], Capt. Dr. Santhosh Baboo[2]

[1]Manonmaniam Sundaranar University

[2]Dr.D.G Vaishnav College, Chennai, India

Abstract: ***Motif extraction is the process of looking for recurring patterns in time series sequences. Dynamic time warping is an algorithm that finds the distance between pairs of sequences and can be used to find clusters in a set of sequences. In this paper, we review the Multivariate Rhythm Sequence Technique of motif discovery and discussed ultrafast subsequence method Dynamic Time Warping(DTW) to identify the subsequence in the discovered motif. Dynamic Time Warping distance is applied both in ECG and Video data and comparative results of two different sequencesaredepicted. We demonstrated our method with the most extensive set of multi-dimensional time series data and experiments are shown.***

Keywords: Motif Discovery, Dynamic Time Warping, time series, motifs, uniform scaling, Subsequence

1. Introduction

Time series motifs are repeated patterns found within the huge volume of data. For many data mining applications, detection of such repeated pattern is significant. This paper concentrates only on temporal sequences and ignores spatial or numeric sequences. The techniques for finding patterns in sequences been used in various applications such as biological sciences, speech recognition, computer science, and quantitative psychology. A time series is a collection of observations made sequentially in time.People measure things that change over time such as blood pressure, annual rain fall etc. working with time series data is difficult due to data handling problems such as different data formats, different sample rates and noise and missing values. There are plenty of tasks are associated with time series data that are clustering, Classification, motif discovery, rule discovery, visualization and novelty detection.The term sub-sequence may have a different meaning in some contexts. According to some computer science texts [2] said a sub-sequence does not have to contain consecutive elements from the parent sequence. Most of data mining algorithms require similarity comparisons as a subroutine, and in spite of the consideration of dozens of alternatives, there is increasing evidence that the Dynamic Time Warping (DTW) measure is the best measure in most domains [1].

Kruskal and Liberman (1983) describe the dynamic time warping (DTW) method that does not increase the distance between two sequences for compressions and expansions of an event. At the same time, this method accounts for insertions, deletions and replacements—are making it an ideal choice for determining the similarities between two category coded sequences that have been obtained by recording the events in an episode over time.

The Euclidean distance metric has been widely used [5], in spite of its known weakness of sensitivity to distortion in time axis [4]. A decade ago, the Dynamic Time Warping (DTW) distance measure was introduced to the data mining community as a solution to solve this weakness [3]. Two-time series that are similar can be compared that are locally out of phase to align in a non-linear manner. DTW is the best known method for time series problems in a variety of domains with $O(n^2)$ time complexity, including bioinformatics [6], medicine [7], engineering, entertainment [8], etc.

The rest of the paper is organized as follows. In Section 3, we give an overview of Multi-Variate Rhythm Motif Mining and its related work. The next sectiondescribes about Dynamic Time Warping (DTW) and comparative results of motif discovery of ECG and Video data. Section 5 suggests some avenues for future researches, and Section 6 gives conclusions and directions for future work.

2. Back Ground and Related Work

A time series x is a sequence of n ordered values such that $x=(x_1,x_2,\ldots,x_n)$ and $x_t\in$for any $t \in [1,n]$. it is assumed that two consecutive values are equally spaced in time or the interval between them can be disregarded without loss of generality. Each value $x_t$is an observation. The Dynamic Time Warping (DTW) algorithm is a most useful distance measure for time series analysis. Empirical evidence suggested that the simple nearest neighbour algorithm using DTW outperforms more "sophisticated" time series classification methods in a wide range of applications [9].

We begin by reviewing background of DTW; Dynamic Time Warping is a distance measure that originated from within the speech recognition community. DTW is a distance measurement between time series that determines similarity between time series and widely used for time series classification. Euclidean distance is an efficient distance measurement that can also be used. The Euclidian distance between two-time series is simply the sum of the squared distances from each nth point in one-time series to the nth point in the other. The main disadvantage of using Euclidean distance for time series data is that its results are

very unintuitive. For example, if two-time series are identical, but one is shifted slightly along the time axis, then Euclidean distance consider them to be different from each other. Dynamic time warping (DTW) was introduced [10] to overcome this limitation.

DTW replaces the one-to-one point comparison, used in Euclidean distance, with a many-to-one (and vice-versa) comparison. The main feature of this approachis that it allows to recognize similar shapes, even if they present signal transformations represented in Fig. 1.

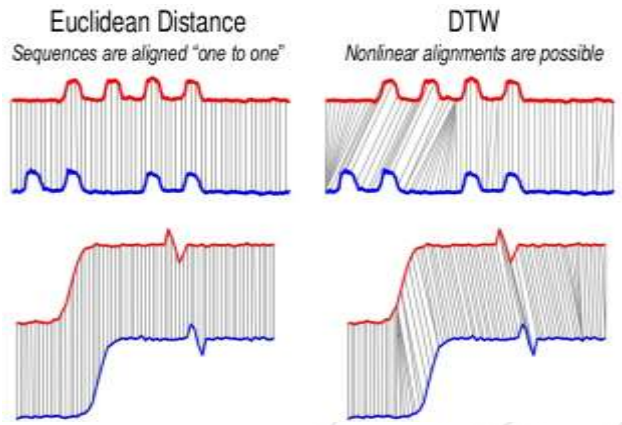


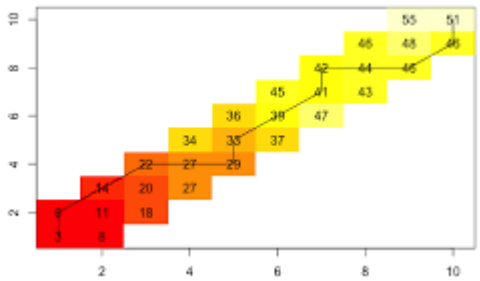
Figure 1: difference between ED and DTW

Recent work strongly suggests that DTW is the best distance measure for many data mining problems [11]. Authors stated in [13] literature search of more than 800 papersa distance measure that outperform by a statistically significant amount is DTW. Recent independent work has empirically confirmed this with exhaustive experiments [12].

DTW allows a one-to-many mapping between data points that enables meaningful comparison between two-time series that have similar shapes but are locally out of phase. Finding the warping path W, construct the distance matrix between the two-time series X and Y. Each element (i, j) in this matrix is the Euclidean distance between the point i^{th} of X and j^{th} of Y. Warping path is a set of contiguous matrix that defines the alignment between X and Y. The k^{th} element of W is defined as w_k= (*i*,). The following constraints are followed to start and finish in diagonally opposite corner cells of the matrix, the subsequent steps must be in the adjacent cells, and all the cells in the warping path must be monotonically spaced in time. This DTW is called unconstrained DTW. Among all the warping paths possible, we are only interested in the path that minimizes the differences between the two-time series data (ECG and Video Data).

$$DTW(X,Y) = \sqrt{\sum_{k=1}^{K} w_k}$$

In the above equation, DTW is a variant decided on how much limit the warping path can deviate from the diagonal. This limit is known as the warping window width (w). For example, in Figure 2.warping path is highlighted.



Many researches shows that DTW deals directly with sequences of different length even contain the phrase of different lengths [14][15] showing examples of such comparison.DTW can be used to measure similarity between sequences of different lengths. Some of these papers further suggest that the simple 4S solution to DTW similarity search is not useful because it requires that sequences of different lengths to be re-interpolated to the same length, and use this fact to motive new approaches.

3. Motif discovery using Multivariate Rhythm Technique

We introduced a simultaneous Multivariate Rhythm Sequence Technique (MRST) to find the rebound repeated motifs and their instance in every document automatically as well as simultaneously.

The video footages from non-dynamic cameras and data location bounded to the motif-mining server. The high semantic representation gave an advantage of event counting and analysis. We used sample images and videos from New York City traffic data for experiments and the results shown better performance than the existing motif mixtures analysis in the time series.

We are getting the video sequence with different movements by various peoples or various objects present in the video of footage. We use long term recording to study about the independent activities and find out motif presence in the video automatically.

Normally the time series has different types. Different type of the time series has the same characteristics of being unification of the multiple actions or movements. For example, we assume the time series prepared in the building for the electricity lining and consumption of the water in a particular building. In this experiment, we can find out the motif of water consumption and short circuit in the particular building.

The overview of the fusion approach for motif discovery in time series data is depicted in this section. In order to achieve the efficiency, fused approach is introduced by employing approaches such as MRST and Fast Motif

4. Dynamic Time Warping(DTW)

Problem Formulation: In our work, the dynamic time warping problem is stated as follows:

Given two-time series X, and Y, of lengths |X| and |Y|,$X = x_1, x_2 \ldots . x_i, \ldots . x_X$ $Y = y_1, y_2 \ldots . y_i, \ldots . y_Y$

Construct a warp path W, W= $w_1, w_2 \ldots . w_i, \ldots . w_K \max(|X|,|Y|) \leq K < |X|+|Y|$ where K is the length of the warp path and the k[th]element of the warp path is $w_k = (i, j)$ where i is an index from time series X, and j is an index from time series Y. Warp path starts at w1 = (1, 1) beginning of each time series and end at $(w_k = (X,Y)$K = (|X|, |Y|). it ensures that every index of both time series is used in the warp path. There is also a constraint on the warp path that forces i and j to be monotonically increasing in the warp path, it also representing the warp path do not overlap. Every index of each time series must be used. The optimal warp path is the warp path is the minimum-distance warp path, where the distance of a warp path Dist(W) is the distance (typically Euclidean distance) of warp path W, and Dist(wki, wkj) is the distance between the two data point indexes (one from X and one from Y) in the kth element of the warp path.

DTW is a dynamic programming approach is used to find this minimum-distance warp path. Instead of attempting to solve the entire problem all at once, solutions to sub-problems (portions of the time series) are found, and used to repeatedly find solutions to a slightly larger problem until the solution is found for the entire time series. A two-dimensional |X| by |Y| cost matrix D, is constructed where the value at D(i, j) is the minimumdistance warp path that can be constructed from the two time series X'=x1 ,...,xi and Y'=y1 ,...,yj . The value at D(|X|, |Y|) will contain the minimum-distance warp path between time series X and Y. Both axes of D represent time.

The x-axis is the time of time series X, and the y-axis is the time of time series Y. Cost matrix is found with a minimum-distance warp path traced through it from D(1, 1) to D(|X|, |Y|).

If the warp path passes through a cell D(i, j) in the cost matrix, it means that the ith point in time series X is warped to the j th point in time series Y.

If X and Y were identical time series, the warp path through the matrix would be a straight diagonal line. To find the minimum-distance warp path, every cell of the cost matrix must be filled. The rationale behind using a dynamic programming approach to this problem is that since the value at D(i, j) is the minimum warp distance of two time series of lengths i and j, if the minimum warp distances are already known.

5. DTW for Comparing two motif Sequences(ECG and Video Data)

The dynamic time warping (DTW) technique is the best method that shows the distance between two sequences. It also provides a fairly accurate measure of how dissimilar the sequences are. A distance of 0 implies that the sequences have an identical ordering of their constituent elements because the DTW algorithm accounts for motif code compressions and expansions. Overall aim of this article is to find the distances that to find patterns in the different motif sequences by seeing how similar or dissimilar pairs. Later it is used to group of similar motifs. These motifs would indicate the different ways that a specific event in the videois played out over time or identify a sequence played significant role in the reading of ECG data.

Dataset of one day of electrocardiograms (ECGs) sampled at 256Hz. We created this data by concatenating the ECGs of more than 10 people with 8,518datapoints. The short video footage is created, coloured and superimposed in an image. Real time activity in the traffic timeis taken as video data setto analyse the results and motif recover by the method is the real time image.

Multivariate Rhythm Sequence Technique proposed in [16] used to discover the motif from the above two data sets. The following Figures 2 and 3 shows the sample motif discovered using MRST Technique.

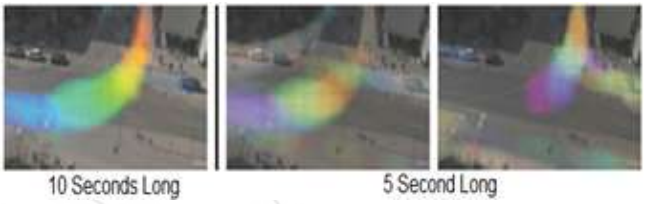


Figure 2: Motif of 10 seconds from Video Data

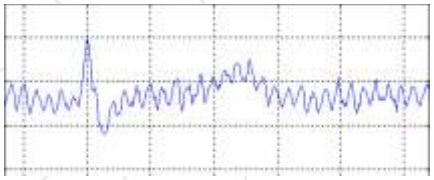


Figure 3: Motif of ECG data

Based on the extracted motifs from two different Sequence video or ECG data, we started comparative analysis of two different sequence using Dynamic Time Warping Technique. The Results of Dynamic Time Warping is shown in the following table. Error rates of both data sets are depicted in table 1.

Table 1: Error rates of DTW

Data Set	DTW error rate
ECG Data	1.80
Video Sequence	30.83

We then compared the DTW technique with the popular distance measure Euclidean Distance Measure to compare the discovered motifs. Results of such comparison is shown in the following Table 2.

Table 2: Time taken to measure distance(Msec)

Data Set	Euclidean	DTW
ECG Data	45	8670
Video Sequence	18	1112

Results shows that Dynamic Time Warping Technique gives much better results than Euclidean distance on motif discovery and DTW is slow when compared to Euclidean distance.

6. Conclusion and Future Work

The overall goal was to demonstrate motif discovery in sequences of behaviouralinformation such as video or ECG records. MRST technique is used to extract the motifs from two different data sets both be used to classify groups of sequences and find the patterns or structures that characterize them. Distance algorithms such as DTW used to compare individual sequences and seeing how they cluster together. To improve the DTW on these data sequences, fast approximations can be used. For long time series, shape based similarity gives very poor results. We need to measure similarly based on high level structureor long time series, shape based similarity will give very poor results. We need to improve this by applying similarity measure on high level structure. We have a plan to approximate the time series with some compressed or downsampled representationin our future .

References

[1] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. J. Keogh. 2008. Querying and mining of time series data: experimental comparison of representations and distance measures. PVLDB 1, 2, 1542-52.

[2] Gusfield, D. (1997). Algorithms on strings, trees, and sequences: Computer science and computational biology. Cambridge University Press.

[3] Berndt, D. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. AAAI Workshop on Knowledge Discovery in Databases, pp. 229-248.

[4] Keogh, E. (2002). Exact indexing of dynamic time warping. In 28th International Conference on Very Large Data Ba Hong Kong. pp. 406-417

[5] Keogh, E. and Kasetty, S. (2002). On the Need for Time Seires Data Mining Benchmarks: A Survey and Empirical Demonstration. In the 8th ACM SIGKDD, pp. 102-111

[6] Aach, J. & Church, G. (2001). Aligning gene expression time series with time warping algorithms. Bioinformatics. Vol. 17 pp. 495-508

[7] Caiani, E.G S., Pieruzzi, F., Crema, C., Malliani, A., &Cerutti, S. (1998). Warped-average template technique to track on a cycle-bycycle basis the cardiac filling phases on left ventricular volume. IEEE Computers in Cardiology, pp. 73-76

[8] Zhu, Y. &Shasha, D. (2003). Warping Indexes with Envelope Transforms for Query by Humming. SIGMOD 2003. pp. 181-192.

[9] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, "Experimental comparison of representation methods and distance measures for time series data", Data Min. Knowl. Discov., vol. 26, no. 2, pp. 275-309, 2013.

[10] Kruskall, J. & M. Liberman. The Symmetric Time Warping Problem: From Continuous to Discrete. In Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, pp. 125-161, Addison-Wesley Publishing Co., Reading, Massachusetts, 1983

[11] Ding, H., Trajcevski, G., Scheuermann, P., Wang, X. and Keogh, E. Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures. Proc' of the VLDB Endowment, 2008, 1542-52.

[12] Paparrizos, J. and Gravano, L. k-Shape: Efficient and Accurate Clustering of Time Series. In Proceedings of the 2015 ACM SIGMOD, 1855-1870.

[13] Rakthanmanon, T., et al. Searching and mining trillions of time series subsequences under dynamic time warping. In Proc' of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, 262-270.

[14] Bozkaya, T, Yazdatani, Z, and Ozsoyoglu, Z. Matching and Indexing Sequences of Different Lengths. CIKM-97

[15] Park, S., Chu, W, Yoon, J., and Hsu, C (2000). Efficient searches for similar sub-sequences of sequence databases. In ICDE-00.

[16] R Kumar & Capt. Dr. S Santhosh Baboo (2017) Discovery of Non-Persistent Motif Mixture using MRST (Multivariate Rhythm Sequence Technique), Global Journal of Computer Science and Technology: Software & Data Engineering, Volume 17, Issue 1 Version 1.0, ISSN 0975-4350