Discover Motifs in Multi Dimensional Time-Series Using the Principal Component Analysis and the MDL Principle

Yoshiki Tanaka and Kuniaki Uehara

Department of Computer and Systems Engineering, Kobe University
1-1 Rokko-dai, Nada, Kobe 657-8501, Japan
{yoshiki, uehara}@ai.cs.scitec.kobe-u.ac.jp

Abstract. Recently, the detection of a previously unknown, frequently occurring pattern has been regarded as a difficult problem. We call this pattern as "*motif*". Many researchers have proposed algorithms for discovering the motif. However, if the optimal period length of the motif is not known in advance, we can not use these algorithms for discovering the motif. In this paper, we attempt to dynamically determine the optimum period length using the MDL principle. Moreover, in order to apply this algorithm to the multi dimensional time-series, we transform the time-series into one dimensional time-series by using the Principal Component Analysis. Finally, we show experimental results and discuss the efficiency of our motif discovery algorithm.

1 Introduction

Many researchers have been studying the extraction of various features from time-series data. One of these problems, efficient extraction of previously defined patterns has been received much attention. And, according to [5], "This problem may now be essentially regarded as a solved problem. However, a more interesting problem, the detection of previously unknown, frequently occurring patterns is still regarded as a difficult challenge". They call this pattern as motif[5]. The term "motif" can be defined by the characteristic that it includes the subsequences with similar behavior (temporal variation) appeared frequently in the time-series. Motif extraction is useful to discover association rules from time-series data[1], or to cluster the time-series data[2], etc.

Many researchers have proposed algorithms for discovering a motif[3][4][5]. Among them, EMMA algorithm[5] has the widest application range that can discover motifs efficiently. The algorithm extracts the motifs with various period lengths. However, we need to discover the "true motifs" in the motifs. Therefore, the computation time of extracting the motifs increases. That is, if the optimal period length is not known in advance, the algorithm is not directly applicable.

We illustrate an example of different lengths of the period in Fig. 1. The motifs in Fig. 1(a) are considered to be valid since they are almost similar. On the other hand, if the period length is longer than that of Fig. 1(a), EMMA algorithm cannot discover the motifs shown in the Fig. 1(b). Similarly, the algorithm may extract too many irrelevant motifs using a shorter period length. Hence, we need to determine an optimum period of length of a motif. One idea is to solve the problem by exhaustively applying the algorithm for all possible period lengths, but, it seems to be impractical.

In our approach, we improve the algorithm that can be applied to the multi dimensional time-series data, because, in the real world, spatial-temporal data can be represented as the multi dimensional time-series but not in the form of one dimensional time-series data.

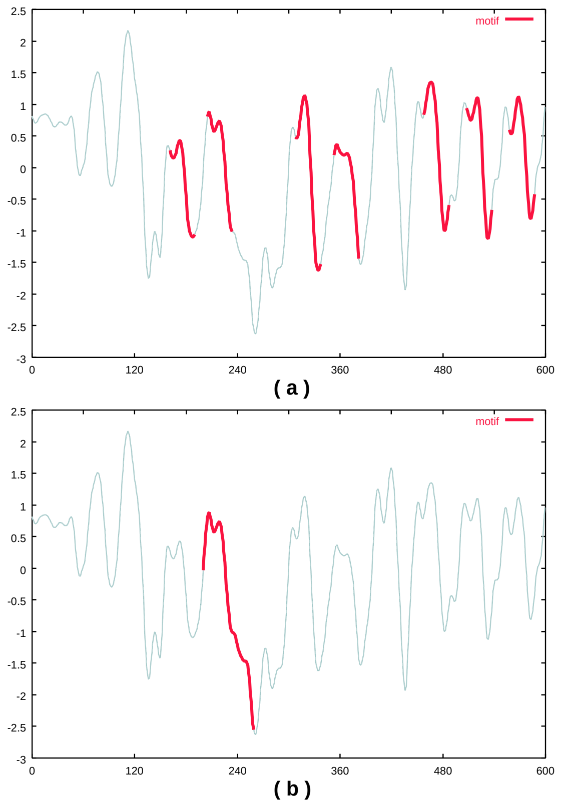


Fig. 1. Motifs discoverd from the same time-series using EMMA algorithm. Fig. 1 period length of motif (a) using 30 ,(b) using 60.

For example, in case of motion capturing system, we can obtain 3 dimensional time series data. Here, an actor puts on 18 markers which reflect infra-red ray (Fig. 2(a)). He performs some actions with the markers being surrounded with 6 cameras. The cameras record the actor's action as video images and calculate 3-dimensional locations of the markers. Finally, we obtain the 3 dimensional time-series data as in Fig.2(b). The figure represents the movement of the right hand while pitching a ball.

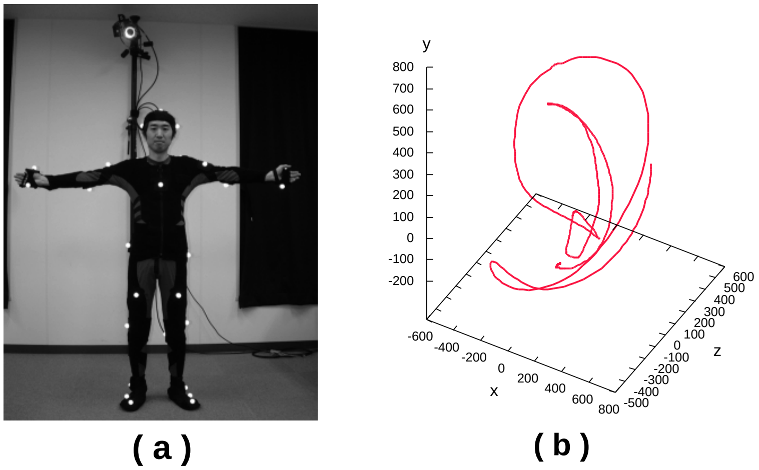


Fig. 2. (a) The motion capture system and the actor who puts on markers. (b) An example of the 3 dimensional time-series data obtained from the motion capture system.

In the work[6], Mori et al. reported the motion recognition from the 3-dimentional time series data obtained from the motion capture system. The recognition process in this study, requires temporal segmentation. The task is to divide the time series into subsequences called primitive motion at the points where velocity changes. However, in this approach, there remains an important problem that it has no fundamental basis for dividing time-series. To solve this problem in this work, a motif is extracted as a primitive motion from the time-series data. Then, we improve the accuracy and efficiency of motion recognition.

In this paper, we attempt to determine the optimum period length of motif dynamically, and discover motif that a human can recognize intuitively from the multi dimensional time-series data. First, we use Principal Component Analysis to transform multi dimensional time-series data to one dimensional time-series data. Second, based on the MDL (Minimum Description Length) principle [7] we discover optimum period lengths of motifs, that are the candidates of a motif. Finally, to discover the motif among the candidates, we employ simplified EMMA algorithm with the optimum period length. The advantage of our algorithm lies in that it can reduce the computation time for finding the motif. It can also find precise motifs than that of EMMA algorithm, from the view point of human intuition.

2 Dimensionality Reduction of Multi Dimensional Time-Series

To discover motifs from multi dimensional time-series data, we need to solve several problems. Among these problem, one significant problem is the requirement of huge amount of calculation time. Another significant problem is the complexity to discover motif directly from multi dimensional time-series. For this reason, no researcher could propose an appropriate algorithm yet.

To solve these problems, we transform multi dimensional time-series into one dimensional time-series data. We discover the motifs from the one dimensional time-series using existing motif discovery algorithm. However, in the transformation, we must minimize the loss of some information of the original multi dimensional time-series. For this purpose, we focus on the PCA (Principal Component Analysis) [8]. It is widely used in the statistical field recently.

The PCA is an effective method to find the features of the data expressed with some observed variables. For example, in a statistical field, this analysis is used to determine the data of two or more stock prices for the indexing purpose. We illustrate specific method to apply PCA to the time-series data. For example, a m dimensional time series C of length n can be represented as follows:

$$\begin{aligned} C &= c_1, c_2, \cdots, c_t, \cdots, c_n \\ c_t &= (x_{1t}, x_{2t}, \cdots, x_{mt}) \end{aligned} \tag{1}$$

In order to apply the PCA, we need to calculate a covariance matrix for the time-series by using the following equation:

$$\begin{bmatrix} \sum x_{1t}x_{1t} & \sum x_{1t}x_{2t} & \cdots & \sum x_{1t}x_{mt} \\ \sum x_{2t}x_{1t} & \sum x_{2t}x_{2t} & \cdots & \sum x_{2t}x_{mt} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{mt}x_{1t} & \sum x_{mt}x_{2t} & \cdots & \sum x_{mt}x_{mt} \end{bmatrix} \tag{2}$$

Each eigenvalue λ_i is ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. The eigenvector is represented as $[e_{1\lambda_i} e_{2\lambda_i} \cdots e_{m\lambda_i}]$. Then, the i-th principal component of time-series pc_{t,λ_i} is calculated by means of $x_1, x_2, \cdots, x_m$ respectively.

$$\begin{aligned} pc_{t,\lambda_i} = {} & e_{1\lambda_i}(x_{1t} - \bar{x_1}) + e_{2\lambda_i}(x_{2t} - \bar{x_2}) + \\ & \cdots + e_{m\lambda_i}(x_{mt} - \bar{x_m}) \end{aligned} \tag{3}$$

Most of the variance in the data is explained by considering only the first principal component [8]. As it accounts for most of the information in the data, we use the first principal component to effectively transform the multi dimensional time series to one dimensional time-series. Finally, we obtain one dimensional time-series $\dot{C}$ as follows:

$$\dot{C} = \dot{c_1}, \dot{c_2}, \cdots, \dot{c_t}, \cdots, \dot{c_n} \tag{4}$$

$$\begin{aligned} \dot{c_t} = {} & e_{1\lambda_1}(x_{1t} - \bar{x_1}) + e_{2\lambda_1}(x_{2t} - \bar{x_2}) + \\ & \cdots + e_{m\lambda_1}(x_{mt} - \bar{x_m}) \end{aligned} \tag{5}$$

In Eq. 5, the $\dot{C}$ is a linear combination of the original variables. Hence, iterational component of the significant dimensional data can be included in the first principle component. Therefore, we can assume that the discovered motif from $\dot{C}$ is same as that of the original multi dimensional time-series C.

3 Detecting an Optimum Period Length and Candidates for a Motif

To make our motif discovery algorithm useful in the various fields, it is necessary to dynamically determine a period length of a motif. In this section, we illustrate an algorithm that detects an optimum period length of motif based on the MDL principle. MDL principle is proposed by Rissanen [7]. It is used to estimate the optimality of a stochastic model. The “stochastic model” is specified to presume the “immanent structure” of the given data in various fields. The principle states that the best model to describe a set of data is that model which minimizes the description length of the entire data set. Here, for the time series data, we regard the best model as the motif. In other words, the motif minimizes the sum of the description length of a given time series data and the description length of the motif itself. Based on this idea, we introduce the algorithm to detect an optimum period length and candidates for a motif.

3.1 Transforming Time-Series into a Sequence of Symbols

We use the MDL principle for extracting an optimum pattern that is expected to be a motif. However, there is an underlying problem that the same patterns hardly appear in the time-series. In addition, we want to extract a pattern without being influenced by the “noise” of the time-series. For these reasons, we transform the time-series data into a sequence of symbols that represents the behavior excluding the noise. Then, we detect an optimum pattern of symbols in the sequence of symbols.

In order to transform time-series data into a sequence of symbols, we use dimensionality reduction algorithm based on a PAA (Piecewise Aggregate Approximation) representation [5]. Here, we show the visualization of this transformation algorithm in Fig. 3.

First, we obtain subsequences by shifting the analysis window of T_{min}, the minimum period length of motif (Fig. 3a). Second, each subsequence is transformed into a sequence of “PAA symbols” (Fig. 3b). The PAA representation is a vector expression that uses the average value in each small segment. A time series $C = c_1, \cdots, c_n$ of length n can be represented as a w-dimensional space by a vector $\bar{C} = \bar{c}_1, \cdots, \bar{c}_w$:

$$\bar{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_i \tag{6}$$

In order to transform the vector of w dimension into a sequence of “PAA symbols”, we need to determine “breakpoints”. These breakpoints determine the range of the PAA value for assigning unique PAA symbol.

According to [5], breakpoints are determined as follows: “We can simply determine the breakpoints that will produce equal-sized area under Gaussian curve, because the normalized time-series has the feature that it has highly Gaussian distribution. Breakpoints are a sorted list of numbers $B = \beta_1, \cdots, \beta_{a-1}$ such that the area under a $N(0, 1)$ Gaussian curve from β_i to $\beta_{i+1} = 1/a$ (β_0 and β_a are defined as $-\infty$ and ∞, respectively). Then, all PAA coefficients that are below the smallest breakpoint are mapped to

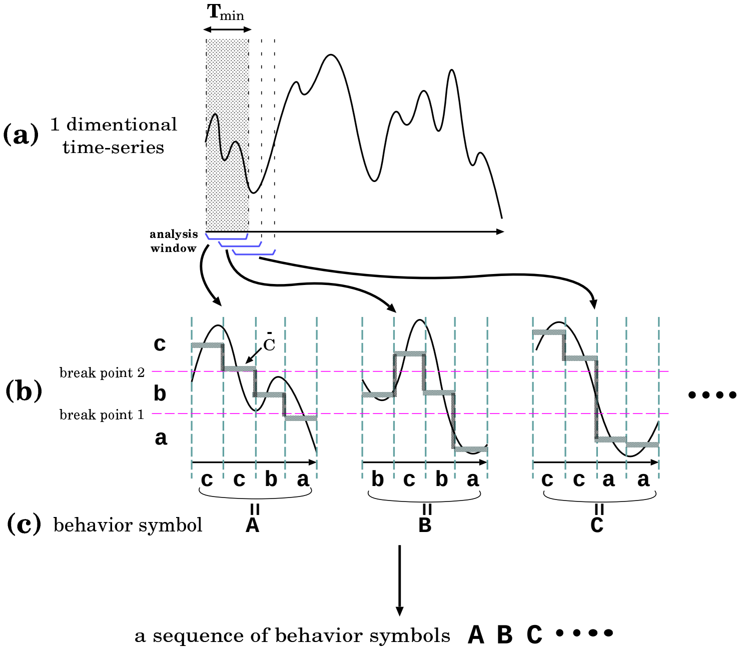


Fig. 3. Visualization of the algorithm of transforming time-series into a sequence of symbol. (a) we obtain subsequences by shifting the analysis window. (b) Each subsequence is transformed into a sequence of PAA symbols, that is based on PAA representation $\bar{C}$. (c) "behavior symbol" is assigned for every pattern in the order of PAA symbols.

a PAA symbol "**a**". All coefficients greater than or equal to the smallest breakpoint and less than the second smallest breakpoint are mapped to the PAA symbol "**b**", etc".

Finally, "behavior symbol" is assigned for every subsequence of PAA symbols. For example, the behavior symbol "A" is assigned to the subsequence "CCBA" in Fig. 3(c). Here, from a view point of our definition of the motif, we can say that discovering pattern from this sequence of behavior symbols $\tilde{C} = \tilde{c_1}, \cdots, \tilde{c_{n_a}} (n_a = n - T_{min} + 1)$ is same as discovering motif from original time-series.

3.2 Estimating Extracted Motif Candidate Based on MDL Principle

To estimate the optimality of the extracted pattern from the sequence $\tilde{C}$ using the MDL principle, we need to define a description length of sequence of symbols. We assume that n_p is the length of a subsequence SC appeared in the sequence $\tilde{C}$ and s_p is the number of unique symbols used in SC. First, we need $log_2 n_p$ bits to encode the number of symbols of SC. Then, encoding the labels of all n_p symbols requires $n_p log_2 s_p$ bits. Hence, the description length of SC is defined as follows:

$$DL(SC) = log_2 n_p + n_p log_2 s_p \tag{7}$$

In addition, we need to define the description length $DL(\tilde{C}|SC)$. This is the description length of $\tilde{C}$ where a subsequence SC is replaced with one symbol. The length of such a sequence is $\acute{n}_a$ and the frequency of appearance SC in $\tilde{C}$ is q. The description length $DL(C|SC)$ is calculated as follows:

$$DL(\tilde{C}|SC) = log_2 \acute{n}_a + \acute{n}_a log_2(s_a + q) \tag{8}$$

Where, $log_2 \acute{n}_a$ is the number of bits required to encode the number of symbols of $\tilde{C}$. $\acute{n}_a log_2(s_a + q)$ is the number of bits required to encode the labels of all $\acute{n}_a$ symbols. Finally, MDL estimation function $MDL(\tilde{C}|SC)$ of $\tilde{C}$ to SC is defined as follows:

$$\begin{aligned} MDL(\tilde{C}|SC) &= DL(\tilde{C}|SC) + DL(SC) \\ &= log_2 \acute{n}_a + \acute{n}_a log_2(s_a + q) \\ &\quad + log_2 n_p + n_p log_2 s_p \end{aligned} \tag{9}$$

We consider that the subsequence SC which has the minimum value of the MDL estimation function is the optimum pattern of $\tilde{C}$.

3.3 The Optimum Pattern Extracting Algorithm

In this section, we illustrate the optimum pattern detection algorithm using the definition in section 3.2. Fig. 4 shows the visualization of the algorithm. We obtain subsequences of symbols from the sequence $\tilde{C}$ by shifting analysis window with certain lengths. For instance, in Fig.4(a), we obtain the subsequences, such as “ABC”, “BCB”, “CBB” etc. Then, we regard the pattern which appears frequently as the best pattern of the current symbol sequence (in Fig. 4(a), it is “BCB”). We calculate MDL estimation function M_1 and length of the pattern L_1. In addition, we calculate the location of the pointer P_1 which shows the beginning of the pattern. For instance, in Fig. 4(a), the obtained pointers are located at $2, 5, 10$ etc. Then, we replace these patterns with another symbol such as “$\acute{A}$” in Fig. 4(b). The above analysis is repeated until there is no pattern that appears more than twice in the sequence of symbols. (such as the sequence in Fig. 4(c)).

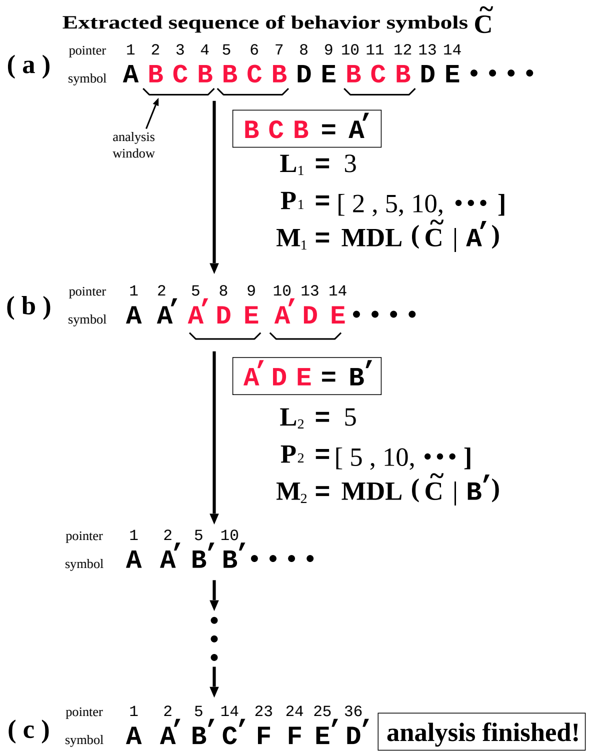
When the analysis is finished, the pattern with the smallest value of MDL estimation function is regarded as the best pattern in the sequence of $\tilde{C}$. Using the length of the pattern L_p, the optimum period length of the motif T_{opt} is calculated as follows:

$$T_{opt} = T_{min} + L_p \tag{10}$$

We consider this pattern as the candidate of motif. Here, we focus on the fact that the sequence of the symbols represents the time-series data. We can guess that the sequence $\tilde{C}$ is obtained by shifting the analysis window. Hence, we can simply consider that a subsequence of original time-series which begins at the pointer P is regarded as the candidates of motif.

4 The Motif Discovery Algorithm

In this section, we describe the motif discovery algorithm from multi dimensional time-series. First, we transform multi dimensional time-series into one dimensional time


Extracted sequence of behavior symbols C̃
(a)
pointer 1 2 3 4 5 6 7 8 9 10 11 12 13 14
symbol A B C B B C B D E B C B D E • • • •
analysis window
B C B = A′
L1 = 3
P1 = [2 , 5, 10, • • •]
M1 = MDL (C̃ | A′)
(b)
pointer 1 2 5 8 9 10 13 14
symbol A A′ A′ D E A′ D E • • • •
A′ D E = B′
L2 = 5
P2 = [5 , 10, • • •]
M2 = MDL (C̃ | B′)
pointer 1 2 5 10
symbol A A′ B′ B′ • • • •
(c)
pointer 1 2 5 14 23 24 25 36
symbol A A′ B′ C′ F F E′ D′
analysis finished!

IF $\min(M_1, M_2, M_3, \cdots) = M_p$

THEN period length of motif $T_{opt} = T_{min} + L_p$

pointers of motif candidates $P = P_p$

Fig. 4. Discovery of optimum motif period length and motif candidate from symbol sequence.

series based on PCA described in Section 2. Then, we need to normalize one dimensional time-series. Because we need to reduce the influence of user-defined threshold of the distance of the ADM algorithm [5]. The ADM algorithm is a part of the EMMA algorithm. Second, an optimum period length is calculated based on MDL principle. Finally, we discover a motif from the candidates by simplified EMMA algorithm.

In EMMA algorithm, each subsequence of a time-series is stored in a hash table to group similar subsequence together. The address of the hash table with the largest number of subsequences is called MPC (Most Promising Candidate). Every subsequence in MPC is regarded as a candidate for a motif. However, the motifs may not be discovered from these candidates as they are shown in Fig. 1(b). It is due to the invalidity of these candidates extracted only on the basis that they appear frequently in the time-series data.

On the other hand, in our algorithm, we use the candidates of motif which is extracted based on the MDL principle as the MPC. From the point of validity of the MDL principle, these subsequences are the best candidates of the motif to the time-series data. We can say that they are the optimum candidates in a broad sense, because they are extracted from a sequence of “behavior symbols”. That is, these subsequences are appeared frequently in a time-series data.

On the contrast, in the EMMA algorithm, the motif is discovered from the candidates by the ADM algorithm [5]. Fig. 5 shows the visualization of the algorithm. The ADM algorithm returns the best motifs from the original MPC subset. These subsequences have the feature that the distance between each two counterparts of the motifs is smaller than the threshold of the distance. However, if the number of motifs is smaller than a certain value, it is considered that the extracted motif does not qualify as a “motif” as defined in Section 1.

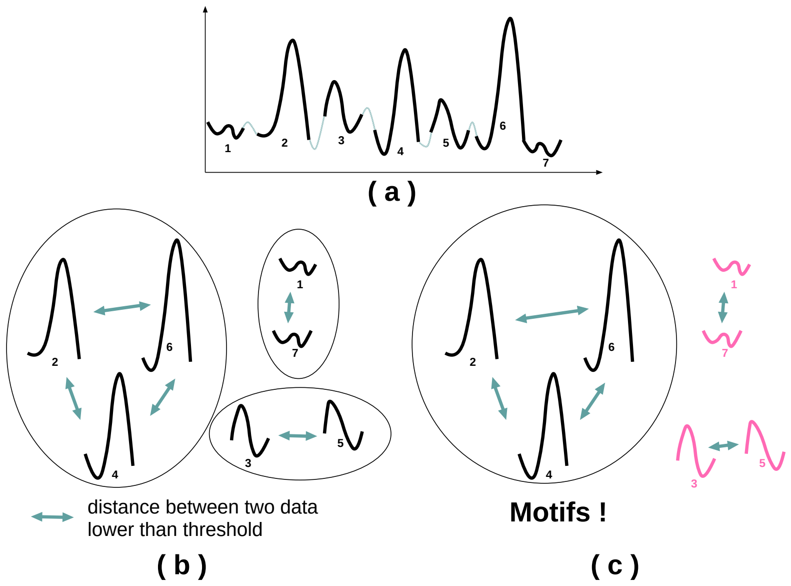


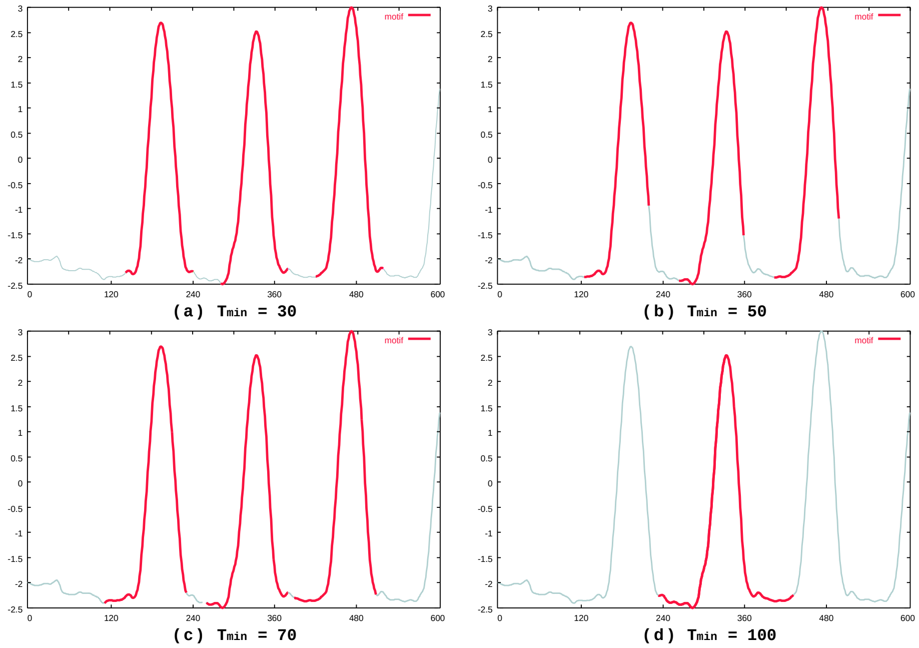
Fig. 5. Visualization of the ADM algorithm. (a) Discovering candidates of motifs using our algorithm. In this example, the number of candidates is 7. (b) Calculate distances between each two candidates and cluster candidates whose distances are lower than the user defined threshold of distance between each of the two candidates . (c) A cluster that has the most number of candidates is regarded as the best motifs.

Then MPC is recalculated and the algorithm retries to discover the motif with the new candidates in new MPC. This increases the computation time because of the invalidity of Lin’s MPC determination method described above. On the other hand, our algorithm, needs no iteration for using the ADM algorithm. Because most subsequences of our candidates can turn into "true" motifs due to the validity of extracted candidates. Hence, we can say that our algorithm is better than the EMMA algorithm from the view point of computation time.

5 Experimental Evaluation

In this section, to show the efficiency of our motif discovery algorithm, we extract motif from multi dimensional time-series data set. We use 3 dimensional time-series data set of human motion obtained from the motion capture system.

First, we show the effects of using different T_{min} value. We assume that it is possible to discover motifs with various period length using small T_{min} value. We also assume that motifs discoverd using large T_{min} value may not satisfy the definition of motif. In Fig. 6 and Fig. 7, the examples of motifs discoverd using various T_{min} value are shown.



(a) T_{min} = 30

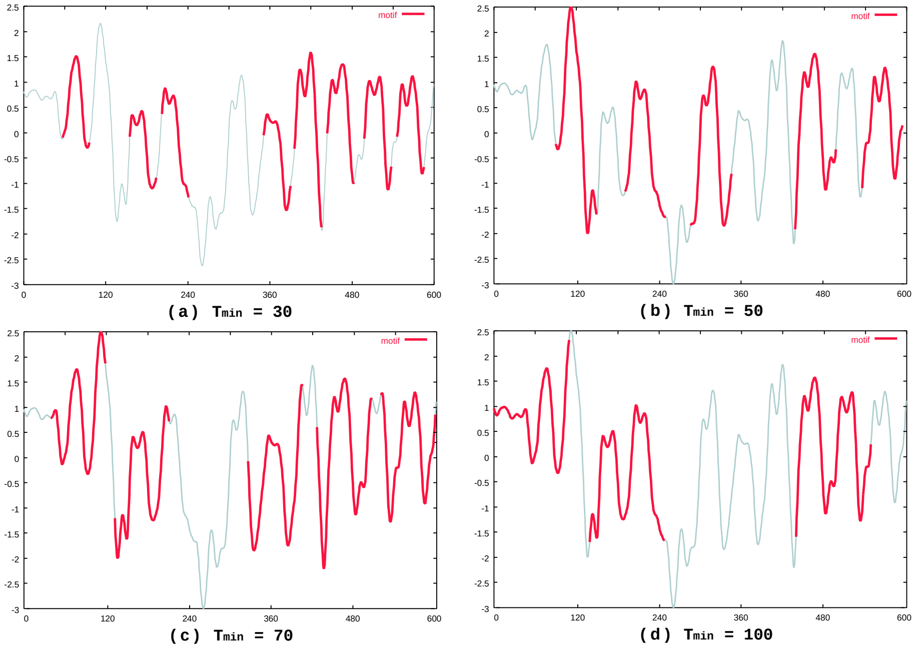
(b) T_{min} = 50

(c) T_{min} = 70

(d) T_{min} = 100

Fig. 6. Examples of discoverd motif from the time series data "Feet movement while Walking" using various value of T_{min}.

We can intuitively recognize the optimum period lengths of the time series data which is about 90 in Fig. 6, and about 40 in Fig. 7. In Fig. 6 (a)(b)(c), it seems that extracted motifs using the T_{min} values which are less than 90 have similar behavior and satisfy the definition of motif. On the other hand, in Fig. 7, it seems that all extracted



(a) T_{min} = 30

(b) T_{min} = 50

(c) T_{min} = 70

(d) T_{min} = 100

Fig. 7. Examples of discoverd motif from the time series data “Head movement while Running” using various value of T_{min}.

motifs satisfy the definition of motif. However, motifs extracted using the T_{min} values which are more than 40 (Fig. 7 (b)(c)(d)) are not the most frequently occuring pattern in the time series. From this result, we can prove that it is possible to extract the optimum motifs that have various optimum period lengths with small T_{min} value. So, we use $T_{min} = 30$ for the following experiments.

Fig. 8 shows the motifs extracted from time series data using our motif discovery algorithm. From the result, it is observed that each motif is discovered using different period lengths. It is also observed that every extracted motif satisfies the definition of motif.

Next, we direct our attention toward considering the motif in terms of the validity of multi dimensionality. Fig. 9 shows an example of each coordinate of the motif. As seen from these results, the motifs of coordinate x and y satisfy the feature of motif. Because, we can intuituvely find that these motifs have the same behavior. On the other hand, the motif of coordinate z is far from the characteristic of a motif. It occurs due to our method of dimensionality reduction. In this process, the PCA regards coordinate x and y as significant coordinates, but coodinate z as an insignificant coordinate. So, the algorithm mainly extracts information based on the former two coordinates.

However, it has a validity from the viewpoint of the human motion. We can recognize intuitively the feature of motion in case of walking. That is, the coodinate “y” (expressing the movement towards the upper and lower sides), the coodinate “x” (ex-

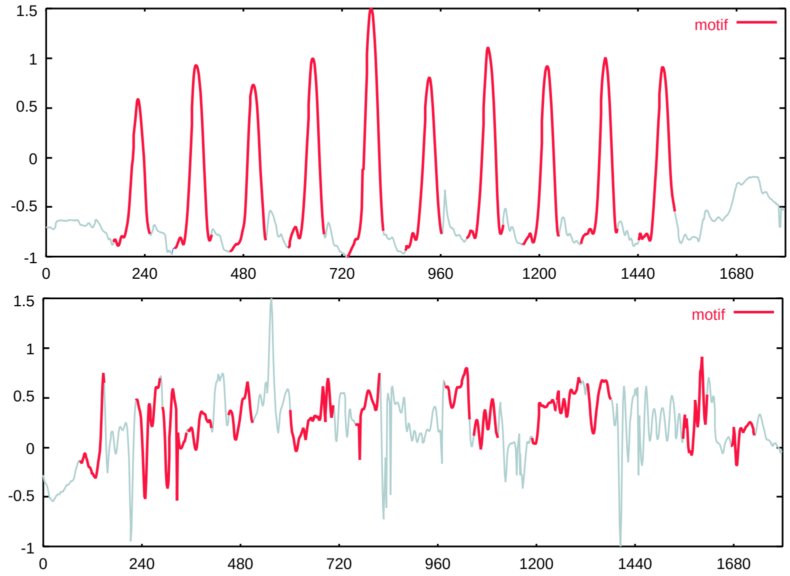


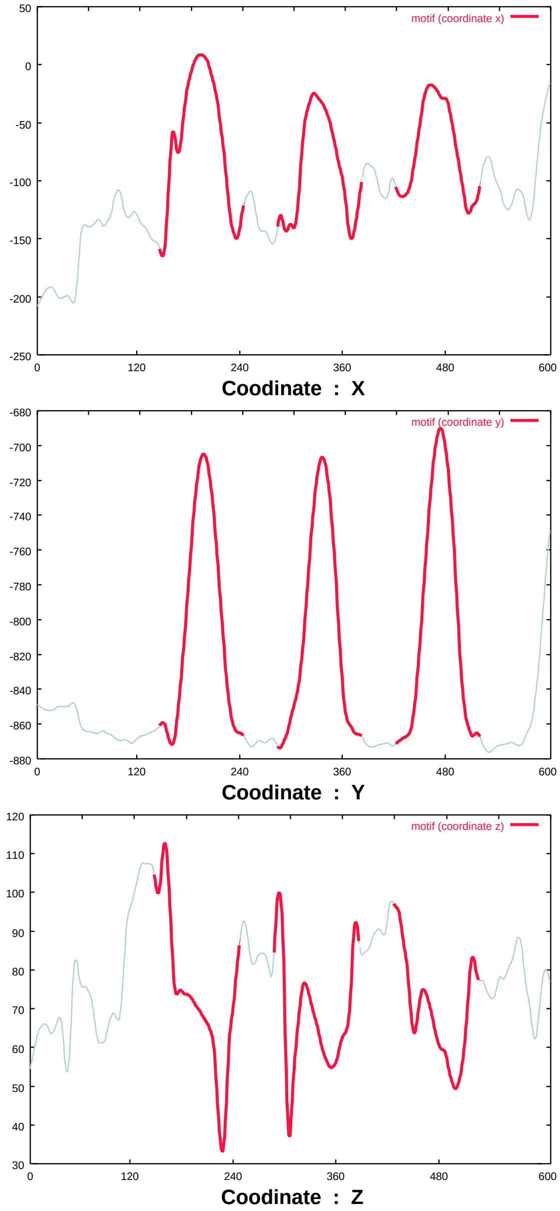
Fig. 8. The motif discoverd in various 1-dimension time series transformed from 3-dimension time series. From top to bottom, the figure represents “Feet movement while Walking”, “Neck movement while running”.

pressing the movement towards left and right), and the coodinate “z” (expressing movement towards backward and forward). For this reason, our motif discovery algorithm is useful for analysing various multi dimensional time-series data.

6 Conclusions and Further Work

In this paper we presented an algorithm for efficiently discovering a motif from multi dimensional time-series data by dynamically detecting an optimum period length of motif. We proved our algorithm’s advantage that, it can extract a motif that human can recognize intuitively. From the result of our experimentation, our algorithm is effective to mine the various unexpected periodicities or extracting rules from time-series, etc. There are several directions to extend this work:

- Although all data with the same behavior are transformed into sequence of symbols, it may be possible that all sequences of symbols are not necessarily be the same at all. It is due to the lack of removing “noise” from time series data completely in the process of generating the symbol sequence. For this problem, we will use the technique of pattern matching of symbol sequence that is not affected by “noise”. For example, this method may be widely used in the genome analysis etc.
- A threshold of distance is used in the ADM algorithm that slightly influence the extraction of motifs. Thus, we hope to determine it dynamically.



Coodinate : X

Coodinate : Y

Coodinate : Z

Fig. 9. An example of discoverd motif from the original 3-dimension time series , “Feet movement while Walking”. The figure represents coordinate “x”, “y”,“z”, respectively.

References

1. Gautam, D., King-lp, L., Heikki, M., Gopal, R., Padhraic, S.: Rule Discovery from Time Series. Proc. of the 4th Int'l Conference on Knowledge Discovery and Data Mining (1998) 16–22
2. Cyril, G., Peter, T., Egill, R., Arup, N.F., Kai, H.L.: On Clustering fMRI Time Series. NeuroImage **9** (1999) 298–310
3. Yu, J.X., K.Ng, M., Huang, J.Z.: Patterns Discovery Based on Time-Series Decomposition. Proc. of PAKDD'2001 (2001) 336–347
4. Berberidis, C., Vlahavas, I., Aref, W.G., Atallah, M., Elmagarmid, A.K.: On The Discovery of Weak Periodicities in Large Time Series. Proc. of PAKDD'2002 (2002) 51–61
5. Lin, J., Keogh, E., Lonardi, S., Patel, P.: Finding Motifs in Time Series. Proc. of the 2nd Workshop on Temporal Data Mining (2002) 53–68
6. Mori, T., Uehara, K.: Extraction of Primitive Motion And Discovery of Association Rules from Human Motion. Proc. of the 10th IEEE International Workshop on Robot and Human Communication (2001) 200–206
7. Rissanen, J.: Stochastic Complexity in Statistical Inquiry. Volume 15. World Scientific (1989)
8. Heras, D.B., Cabaleiro, J.C., Perez, V.B., Costas, P., Rivera, F.F.: Principal Component Analysis on Vector Computers. Proc. of VECPAR (1996) 416–428
9. Kalpakis, K., Gada, D., Puttagunta, V.: Distance Measures for Effective Clustering of ARIMA Time-Series. Proc. of the 2001 IEEE International Conference on Data Mining (2001) 273–280