Multidimensional Motif Discovery in Physiological and Biomedical Time Series Data

Arvind Balasubramanian
The University of Texas at Dallas
800 W Campbell Road
Richardson, TX 75080
arvind@utdallas.edu

Jun Wang
The University of Texas at Dallas
800 W Campbell Road
Richardson, TX 75080
wangjun@utdallas.edu

B. Prabhakaran
The University of Texas at Dallas
800 W Campbell Road
Richardson, TX 75080
bprabhakaran@utdallas.edu

ABSTRACT

Providing personalized diagnosis and therapy requires monitoring patient activity using various body sensors. Sensor data generated during personalized exercises or tasks may be too specific or inadequate to be reviewed and evaluated using supervised methods such as classification. We propose multidimensional time series motif discovery as a means for patient activity monitoring, since such motifs can capture repeating patterns across multiple dimensions of the data, and can serve as conformance indicators. Previous studies pertaining to mining multidimensional motifs have proposed offline algorithms and lack the capability of processing and mining motifs from multiple dimensions concurrently. In this paper, we propose an efficient approach to multidimensional motif discovery in body sensor generated time series data for monitoring performance of patients during therapy. We present two alternative models for multidimensional motifs based on motif co-occurrences and temporal ordering among motifs across multiple dimensions, with detailed formulation of the concepts proposed. The proposed method uses an efficient hashing based record to enable speedy update and retrieval of motif sets, and identification of multidimensional motifs. We also demonstrate the utility of this approach by applying it to synthetic as well as human motion data captured by on-body sensors in both unsupervised motif discovery and content based query resolution tasks. The approach is shown to be effective for (a) tracking repetitions during therapy sessions; (b) finding naturally occurring patterns; and (c) query resolution.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous; D.2.8 [Software Engineering]: Metrics—complexity measures, performance measures

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WOODSTOCK '97 El Paso, Texas USA
Copyright 20XX ACM X-XXXX-XX-X/XX/XX ...$15.00.

General Terms

Theory

Keywords

ACM proceedings, \LaTeX, text tagging

1. INTRODUCTION

The current state of healthcare requires monitoring patient performance while they undergo rehabilitative therapies to provide targeted and personalized care. Physicians and medical care providers usually assign tasks or activities to be carried out by patients either under direct supervision or under remote monitoring. Wearable body sensors along with data analytics have been envisaged for providing continued monitoring of patients and adding value to healthcare. With rapid advancement and confluence of miniaturization and sensing technologies, off-the-shelf wearable body sensors have been developed with capabilities to sense, record, and wirelessly transmit multiple physiological and kinematic data in real-time.

Based on their analysis of the sensor data, the physician can verify compliance, evaluate performance and provide personalized assistance and therapy to the patient for optimal efficacy. However, there are some challenges associated with the monitoring of patient activity using body sensor data – (i) Although the physician might have a clear understanding of normal and abnormal behavior in a patient’s activity, the captured sensor data might be too patient specific and inadequate for the task to be modeled as a classification problem. In other words, a generic classifier that attempts to differentiate between normal and abnormal behavior in patient’s physiological and kinematic data might be insufficient. (ii) Another important factor is that the natural characteristics exhibited in the patient’s data might evolve over the treatment period, i.e., as the patient makes progress, the physician may accordingly need to modify the diagnostic paradigm. Therefore, we need a personalized and context aware approach that (a) analyzes patient activity in order to detect and differentiate between normal and aberrant behavior in an unsupervised or semi-supervised manner; and (b) is scalable across multiple heterogeneous sensors capturing different data (e.g., ECG, blood pressure, motion, etc.). (see Figure 1).

Motif Discovery in sensor generated multi-dimensional time series data is an intuitive and effective solution for mon-
itoring and tracking structural similarities in patient data. Motifs are approximately repeated subsequences or patterns in any time series data. Multidimensional motifs (MDMs) extend the concept to co-occurrences of such motifs across multiple dimensions. MDMs are well suited for our requirements such as personalization as well as context awareness, since they can identify the patterns generated during a therapy session, without the need for classification, and can thus serve as indicators of conformance. Additionally, content based querying over recurring patterns and MDMs for retrieval of patterns of interest is beneficial in cases where a particular medical event or episode and associated symptoms or reasons need to be analyzed.

An ideal MDM discovery approach should – (i) provide for temporal lags between dimensional components of an MDM; (ii) provide for the instances of the same motif to be of different lengths, as well as allow for the individual dimensional components of an MDM to vary in length, eg. unequal strides of walk motion having similar gait characteristics; (iii) enable concurrent processing of multiple dimensions of real-time streaming sensor information, maintain parallel records of motifs in each dimension as well as track co-occurrences between motifs across multiple dimensions; (iv) provide content based querying functionality that can allow the retrieval of specific sections of multidimensional data that consist motifs of interest; (v) be flexible and noise tolerant to induce robustness against variation, and ensure detection of all potential instances of a motif.

In this paper, we present a MDM discovery approach that is compatible with real-time processing of time series streams. The contributions of this approach are:

(a) two alternative models for multidimensional motifs – an unordered “bag of motifs” based model for representing motif co-occurrences having common constituent motifs, and a temporal ordering based model for representing motif groups having common set of temporal relations among constituent motifs (with an optional constraint of having common constituent motifs as well).
(b) concurrent online processing of all dimensions in the multidimensional time series for mining MDMs in real time using an efficient hashing based records to enable speedy update and retrieval of motif sets;
(c) unsupervised identification of variable length MDMs, along with support for content-based querying on motifs.

1.1 Motivation for two alternative models for MDMs

In coordinated biomedical sensor information, such as motion activity, co-occurrence information is useful in associating patterns from different dimensions of the sensor stream based on temporal proximity, which is addressed by our “bag of motifs” model. However, in case of some specific studies involving heterogeneous sensor information, it might also be desirable to capture the exact temporal sequence in which the co-occurring motifs are ordered. For example, when captured concurrently with 3D motion activity data, useful temporal relations can be observed in muscle activity data captured using Electromygram (EMG) sensors. For example, in a jump activity, the onset of EMG signal (contraction of tibialis anterior in this example) occurs prior to onset of the toe movement as the subject prepares to jump [26]. (See Figure 2). Thus, a recurring temporal ordering of motifs can be an indicator of an MDM as well, in contrast with simple co-occurrence of motifs. In fact, similarities between MDMs having different constituent motifs can observed based only on the temporal relations between the constituent motifs.

In the following sections, we briefly review previous work relevant to our study. We present some fundamental concepts and definitions as the basis formulation for our approach, followed by a detailed discussion of the method with supporting examples. Finally, we discuss the implementational aspects of the method, and evaluate the utility of the proposed approach in identifying recurring multidimensional patterns using both synthetic as well as real body sensor datasets. The proposed approach is shown to be effective for (a) tracking repetitions during therapy sessions; (b) finding naturally occurring patterns; and (c) query resolution.

2. RELATED WORK
In recent years, motif discovery in time series has been one of the most important data mining tasks, irrespective of domain and nature. There have been many studies [1, 3, 4, 5, 6, 7, 10, 12, 18, 33] that address the problem of offline as well as online motif discovery despite factors such as noise, scaling, translation etc. in unidimensional data. While unidimensional time series motifs can be mined and interpreted individually, co-occurring motifs discovered in multidimensional time series need to be associated with each other over multiple dimensions of the data captured.

A number of recent studies have proposed solutions for MDM discovery. Tanaka et al [28] reduced multidimensional time series data to a single dimension using Principal Component Analysis (PCA), followed by motif discovery based on minimum description length principle. McGovern et al [15] suggested finding the salient dimensions and then identifying the key temporal motifs in those dimensions, with applications to severe weather prediction. Muenen et al [18] presented an online approach for unidimensional motif discovery and maintenance, and discussed scaling up to multiple dimensions. Chiu et al [7] addressed the issue of motif distortion in the presence of noise, and proposed a probabilistic approach based on random projection, to find similar SAX words (see Appendix Section B) even when they differed by one or more symbols. Minnen et al [16] and Vahdatpour et al [29] extended this approach for MDMs, combining collision matrices resulting from random projection and partial similarity matching for each dimension, by identifying relevant dimensions and by graph clustering, respectively. Ordonez et al. [20, 21] proposed an approach for physiological data classification based on vector-space model for words and documents, presenting a Multivariate Bag of Patterns and Stacked Bag of Patterns representations of multivariate time series. More recently, Moazeni et al. [17] proposed a query based multidimensional signal search, with provision for missing or switched dimensions, noise, scaling and asynchrony.

However, the approaches presented in [16, 29, 20, 21] would not identify motifs of variable length beyond the SAX (see Appendix section B) word segmentation range. Also, a common limitation among these approaches, including the one presented in [17], is that they involve processing the entire information in one dimension before processing the next dimension, or prior knowledge of the entire set of unidimensional patterns that can occur in the multidimensional data, thus implicitly requiring all the data to be available on disk. These requirements pose a challenge in case of applications involving multidimensional time series streams that require real-time tagging of motif co-occurrences, restricting the approaches to mining known patterns or query-based content after complete data from all dimensions is acquired and processed. Unknown patterns and evolving pattern groups cannot be addressed by them in real time. It is necessary to develop a concurrent synchronous motif mining approach that is applicable for real-time purposes, such as sensor data streams where all data is not available before processing. Our approach attempts to overcome this limitation by using a concurrent processing methodology to mine and maintain motifs across multiple dimensions in an online manner. In a previous work [4], we extended the unidimensional variable length motif discovery technique proposed by Li et al [12] to enhance robustness and noise tolerance in motif discovery. In this paper, we present a parallelized motif co-occurrence record to mine MDMs synchronously across multiple dimensions.

With regard to mining the temporal relations between patterns in multivariate data, [9] proposed induction of temporal grammatical rules for identifying the most significant temporal pattern and a rule-based description of primitive patterns, with application to sleep apnea. Allen’s interval algebra [2] is particularly useful in modeling sequence information. This has been used in interval based event mining [23], frequent pattern mining from multiple data domains [24] for classification and similarity evaluation of event sequences [11]. A technique for mining temporal ordering of motifs in time series is presented in [22], but the approach is restricted to lag patterns and does not allow any lag variation within instances of the same lag pattern. The proposed approach uses the interval relations to formulate an alternative definition of an MDM based on the temporal ordering among unidimensional motifs.

3. PRELIMINARIES

3.1 Background

The following definitions of time series and time series motif pertain to a unidimensional/univariate time series.

Definition: A time series $T = (o_1, o_2, o_3, ..., o_n)$ is a temporally ordered sequence of $n$ real value observations.

The sampling rate of the time series, and consequently, the interval between adjacent observations may vary. However, for sake of simplicity of formulation, we assume the sampling rate to be uniform. Varying sampling rates can always be managed by processing the time series in terms of time durations rather than number of samples.

Definition: A time series subsequence $subT_i$ of a time series $T$ is a temporally ordered subset of consecutive real value observations $(o_i, o_{i+1}, o_{i+2}, ..., o_j)$ from $T$ ($1 \leq i < j \leq n$), where $n$ is the length of $T$. The length of the subsequence is $j - i + 1 \leq n$.

Definition: The begin timestamp $begTs(subT_i)$ and end timestamp $endTs(subT_i)$ of a time series subsequence $subT_i = (o_i, o_{i+1}, o_{i+2}, ..., o_j)$ are the time instants $t_i$ and $t_j$ at which the first observation $o_i$ and last observation $o_j$ in $subT_i$ occur, respectively.

In our work, for sake of simplicity, the term timestamp $ts$ when used by itself denotes the begin timestamp $begTs$.

Definition: A unidimensional time series motif $M$ is a set $\{subT_i\}$ of two or more time series subsequences that are evaluated to be approximately similar to each other based on a given distance measure/metric.

Thus, $dist(subTs, subTs) \leq D, \forall\{subTs, subTs\} \in M$, where $dist$ is a distance measure/metric, and $D$ is a threshold value set on $dist$.

Also, subsequences belonging to a motif should be non-overlapping, i.e. $\forall\{subTs, subTs\} \in M, (endTs(subTs) < begTs(subTs)) \lor (endTs(subTs) < begTs(subTs))$.

Definition: An instance $m$ of a unidimensional motif $M$ is any $subT_i \in M$.

Thus, $M$ can be considered to be a distinct motif type, and the motif instance $m$ is one of the occurrences that belong to the motif type indicated by $M$.

In Figure 3, $M_{11}, M_{22}, M_{31}, M_{41}$ are all time series motifs having two instances each, designated by $m_{11}, m_{22}, m_{31}, m_{41}$.
3.2 Formulation of Multidimensional Motifs

The following definitions pertain to multidimensional time series.

**Definition:** A multidimensional time series MDT = \{T_1, T_2, T_3, ..., T_d\} is a set of d synchronous time series. \( T_i \) is the \( i \)th time series in an MDT, and is therefore referred to as the \( i \)th dimension of the MDT.

The synchronous nature of the multidimensional time series MDT is by virtue of the temporal correspondence between information from the time series from each dimension of MDT. The two implications of processing such synchronous time series information are – (i) The information from all dimensions of the time series is available simultaneously; and (ii) the processing of MDT is done in units of time, represented by the width of temporal window, that is equal across all dimensions of MDT. The second implication normalizes any variation of sampling rate between dimensions, since the temporal window for processing is quantized in units of time and not samples.

**Definition:** An instance \( m_a \) of a motif \( M_a \) and an instance \( m_b \) of a motif \( M_b \) occurring in the same or different dimensions of an MDT are said to be neighbors, if \( |ts(m_a) - ts(m_b)| \leq r \), where \( ts(m_a) \), \( ts(m_b) \) are the timestamps of \( m_a \) and \( m_b \), respectively, and \( r \) is a prespecified integer value called the neighbor distance.

Motif instances that are neighbors, are considered to be co-occurring.

**Definition:** A motif neighborhood is a consecutive set of time instants \((t_p, t_{p+r}, t_{p+2r}, ..., t_{p+k})\), where \( r \) is the prespecified neighbor distance, so that, any two motif instances \( m_a \) and \( m_b \) would be neighbors if \( t_p \leq ts(m_a) \leq t_{p+r} \) and \( t_p \leq ts(m_b) \leq t_{p+r} \).

Thus, any motif instances occurring within the same motif neighborhood are neighbors.

In Figure 3, if the neighborhood distance be 4, then the instance of \( M_{21} \) at 22 is a neighbor of the instance of \( M_{31} \) at 26, but not of the instance of \( M_{22} \) at 29.

**Definition:** A MotifBag MBag = \{\( M_1, M_2, ..., M_k \)\}(k \geq 2) is an unordered set of motifs such that, there exists at least one set bag of motif instances \{\( m_1, m_2, ..., m_k \)\} (\( m_i \in M_i, i = 1, 2, ..., k \)) where, \( \forall m_a, m_b \in \text{bag}\{1 \leq a, b \leq k\}, m_a \) and \( m_b \) are neighbors.

The set of motif instances bag is referred to as an instance of the MotifBag MBag.

The motifs in a MotifBag are referred to as its members and can span two or more dimensions.

For example, if \( m_1 \in M_1, m_3 \in M_2 \) and \( m_4 \in M_4 \) are motif instances that are neighbors, then the MotifBags formed are \{\( M_1, M_2 \)\}, \{\( M_1, M_3 \)\} and \{\( M_1, M_4 \)\} respectively.

**Definition:** If \( \text{bag} \) is an instance of MotifBag MBag, and a motif instance \( m \) is a member of \( \text{bag} \), then index(\( \text{bag} \), \( m \)) = (\( \text{dim}, \text{seq} \)) where \( \text{dim} \) is the dimension of the MDT in which \( m \) occurs, and \( \text{seq} \) is the the sequence number of \( m \) in its dimension within the range of \( \text{bag} \). For example, in Figure 4, for the first instance of the MotifBag formed by \{\( M_{22}, M_{41} \)\}, index(\( M_{22} \)) = (2, 1) and index(\( M_{41} \)) = (4, 1).

**Definition:** An unordered Multidimensional Motif UMDM is a MotifBag having two or more non-overlapping instances. Thus, for any two instances umd\( m_a \) and umd\( m_b \) of UMDM, \( \text{end}_a(\text{umd}_a) < \text{beg}_a(\text{umd}_a) \) or \( \text{end}_b(\text{umd}_b) < \text{beg}_b(\text{umd}_b) \).

Figure 4 uses our previous example to illustrate the concept of a UMDM. The motif instances of \( M_{11}, M_{21} \) and \( M_{31} \) together constitute a UMDM, since they form a MotifBag that has two instances. The same is true for \( M_{22} \) and \( M_{41} \).

**Definition:** A temporal relation \( R(m_a, m_b) \) between two motif instances \( m_a \in M_a \) and \( m_b \in M_b \) is the representation of the temporal occurrence of \( m_a \) with respect to \( m_b \).

As dictated by Allen's Interval Algebra [2], the various definitions of \( R(m_a, m_b) \) are listed in Table 1.

**Definition:** A MotifOrdering is a set of pairwise temporal relations among instances of motifs belonging to a MotifBag. Thus, a MotifOrdering \( \text{MOrd} \) is a set of temporal relations \( \{R_i(m_a, m_b)\}, \forall m_a \in M_a, m_b \in M_b \) where
Table 1: Temporal Relations set by Allen’s Interval Algebra

<table>
<thead>
<tr>
<th>$R$</th>
<th>Temporal Relation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_1$</td>
<td>Precedes $(m_a, m_b)$</td>
<td>$R_1\text{end}_t(m_a) &lt; \text{begin}_t(m_b)$ and $\text{end}_t(m_a) = \text{begin}_t(m_b)$</td>
</tr>
<tr>
<td>$R_2$</td>
<td>Meets $(m_a, m_b)$</td>
<td>$\text{begin}_t(m_a) = \text{begin}_t(m_b)$</td>
</tr>
<tr>
<td>$R_3$</td>
<td>Overlaps $(m_a, m_b)$</td>
<td>$\text{begin}_t(m_a) &lt; \text{begin}_t(m_b)$ and $\text{end}_t(m_a) &lt; \text{end}_t(m_b)$ and $\text{end}_t(m_a) &gt; \text{begin}_t(m_b)$ and $\text{begin}_t(m_a) &gt; \text{end}_t(m_b)$</td>
</tr>
<tr>
<td>$R_4$</td>
<td>Starts $(m_a, m_b)$</td>
<td>$\text{begin}_t(m_a) = \text{begin}_t(m_b)$ and $\text{end}_t(m_a) &lt; \text{end}_t(m_b)$</td>
</tr>
<tr>
<td>$R_5$</td>
<td>Contains $(m_a, m_b)$</td>
<td>$\text{begin}_t(m_a) &lt; \text{begin}_t(m_b)$ and $\text{end}_t(m_a) &gt; \text{end}_t(m_b)$</td>
</tr>
<tr>
<td>$R_6$</td>
<td>Finishes $(m_a, m_b)$</td>
<td>$\text{begin}_t(m_a) &gt; \text{begin}_t(m_b)$ and $\text{end}_t(m_a) = \text{end}_t(m_b)$</td>
</tr>
<tr>
<td>$R_7$</td>
<td>Equals $(m_a, m_b)$</td>
<td>$\text{begin}_t(m_a) = \text{begin}_t(m_b)$ and $\text{end}_t(m_a) = \text{end}_t(m_b)$</td>
</tr>
</tbody>
</table>

$M_a, M_b \in \text{bag}$ and $\text{bag}$ is a MotifBag instance, and $R_i \in \{R_1, R_2, ..., R_7\}$ (see Table 1).

An instance of a MotifOrdering is a single occurrence of a MotifOrdering in an MDT. Two instances of a MotifOrdering have the same number of participating motifs and constituent temporal relations with a dimension-wise one-to-one mapping among the temporal relations, but the need not have the same member motifs.

$\text{ord}_1 = \{R(m_a, m_b)\}$ and $\text{ord}_2 = \{R(m_c, m_d)\}$ are instances of the same MotifOrdering $\text{MOrd}$ if and only if, $\forall R(m_a, m_b) \in \text{ord}_1, \exists R'(m_c, m_d)$ such that $R = R'$, $\text{index}(a, m_a) = \text{index}(a, m_c)$ and $\text{index}(b, m_b) = \text{index}(b, m_d)$, where $a$ and $b$ are MotifBag instances with $m_a, m_b \in \text{bag}_1$ and $m_c, m_d \in \text{bag}_2$.

A MotifOrdering can have multiple instances throughout an MDT.

In Figure 5, $M_{11}$ (at 2), $M_{21}$ (at 3) and $M_{31}$ (at 2) are neighbors with temporal relations $\text{AAS}$: Overlaps($M_{11}, M_{21}$), Equals($M_{11}, M_{31}$), Overlaps($M_{31}, M_{21}$). Again, $M_{11}$ (at 2), $M_{21}$ (at 3) and $M_{31}$ (at 2) are neighbors with temporal relations $\text{AAS}$: Overlaps($M_{11}, M_{22}$), Equals($M_{11}, M_{32}$), Overlaps($M_{32}, M_{22}$). These two form instances of the same MotifOrdering even though the member motifs are not the same, since both sets are governed by the same temporal relations.

Definition: An Unconstrained Temporal Ordering based Multidimensional Motif TMDM is a MotifOrdering having two or more non-overlapping instances. Thus, for any two instances tmdm$_a$ and tmdm$_b$ of TMDM, $(\text{end}_t(tmdm_a) < \text{beg}_t(tmdm_b)) \lor (\text{end}_t(tmdm_a) < \text{beg}_t(tmdm_b))$.

In Figure 5, the MotifOrdering with instances $M_{11}, M_{21}, M_{11}$ and $M_{11}, M_{22}, M_{31}$ is a TMDM since it has two instances.

Note 1: Two instances of the same MotifOrdering can be instances of two different MotifBags. Similarly, two instances of the same MotifBag can be instances of two different MotifOrderings.

Note 2: By definition, two instances of the same UMDM should have the exact same member motifs, irrespective of the temporal ordering among the motifs in the instance. In contrast, two instances of the same TMDM should have the exact same temporal ordering, while the actual member motifs can be different.

Definition: A Constrained Temporal Ordering based Multidimensional Motif CTMDM is a TMDM which has the same set of member motifs across all instances. Thus, for any two instances ctmdm$_a$ and ctmdm$_b$ of CTMDM, $(\text{end}_t(ctmdm_a) < \text{beg}_t(ctmdm_b)) \lor (\text{end}_t(ctmdm_a) < \text{beg}_t(ctmdm_b))$.

CTMDM$_1 = \{R(m_a, m_b)\}$ and CTMDM$_2 = \{R'(m_c, m_d)\}$ are two instances of CTMDM if and only if, $\forall R(m_a, m_b) \in \text{ctmdm}_1, \exists R'(m_c, m_d) \in \text{ctmdm}_2$ such that, (i) $R = R'$, (ii) $\text{index}(a, m_a) = \text{index}(a, m_c)$ and $\text{index}(b, m_b) = \text{index}(b, m_d)$, where $a$ and $b$ are MotifBag instances with $m_a, m_b \in \text{bag}_1$, $m_c, m_d \in \text{bag}_2$. (iii) $m_a$ and $m_b$ are instances of the same motif, and $m_c$ and $m_d$ are instances of the same motif. The third point represents the additional constraint on the existing definition of TMDM.

3.3 Problem Statement

We now formulate the formal statements of the problems being addressed by the proposed approach.

Problem Statement 1: Unordered Multidimensional Motif Discovery Given a d-dimensional time series MDT, identify and retrieve every distinct Unordered Multidimensional Motif UMDM and its instances.

Problem Statement 2: Temporally ordered Multidimensional Motif Discovery Given a d-dimensional time series MDT, identify and retrieve every distinct Temporally ordered Multidimensional Motif TMDM and its instances.

4. PROPOSED METHOD

4.1 Multidimensional Motif Discovery

The proposed approach provides the functionality of concurrently processing all dimensions of the multidimensional time series data, while indexing unidimensional motifs and their co-occurrences as and when they are encountered.

4.1.1 Identifying Co-occurrences of Unidimensional Motif Instances

Motif instances from different dimensions can be grouped together based on their timestamps as co-occurring motif instances. To achieve this, we execute concurrent motif discovery on all dimensions of the multidimensional time series data (see Section 5.1), while maintaining a record of the occurrences of unidimensional motif instances in terms of their timestamps. We use an inverted timestamp index, which maintains each timestamp mapped to the unidimensional motifs that have an instance occurring at that particular timestamp, i.e. $\{ts \rightarrow \{M_i\}\}$ (see Figure 6). A particular timestamp value can have one or more motif instances from different dimensions mapped against it. This inverted index is maintained and updated throughout the course of unidimensional motif discovery, with motif instances identified from all dimensions of the time series recorded in this index. The grouping of such motif instances, or neighbors, is based on a predefined temporal proximity, referred to as the neighborhood distance, and can be achieved using a sliding window having a width equal to the neighborhood distance on the timestamp index keys. Table 2 shows the different
sets of neighbors obtained from the example time series using a neighborhood distance of 4.

### 4.1.2 Indexing Co-occurring Motif Groups

Multidimensional motifs are quite simply recurring groups of co-occurring unidimensional motif instances spanning multiple dimensions. Once the neighbor sets belonging to a particular duration of time have been identified, we need to record each combination of such neighboring motif instances, referred to as MotifBags (See Section 3.2), as a potential UMDM. Each MotifBag would thus be a distinct combination of two or more motif instances from two or more dimensions.

On the entry of each new motif instance in the timestamp index, every motif neighborhood that the newly identified motif instance belongs to is searched. From each such neighborhood, sets of all possible combinations of neighboring motif instances are computed and designated as MotifBags (see Algorithm 1).

**Reducing Time complexity**: Generation of the exhaustive set of neighbor combinations involving a particular motif instance can take \( r \times O(2^k) \) time in the worst case scenario, where \( r \) is the neighborhood distance and \( k \) is the number of neighbors in each neighborhood with respect to the motif instance. However, a simple optimization of this step can reduce this time cost. Instead of executing on the entry of every new motif instance, the procedure can wait till unidimensional motif discovery processing is completed for that timestamp, and all currently discovered motif instances having this timestamp are indexed.

Since recurring MotifBag instances are potential UMDMs, we need to track the occurrences of previously encountered MotifBags while simultaneously recording new motifBags. For this purpose, we use a frequency index MBIndex to keep track of each distinct MotifBag and its frequency (number of instances) (see Figure 7). Each distinct identified MotifBag is recorded in the index, with the timestamp information of its current instance. Over time, when another instance of an existing MotifBag is encountered, the timestamp information of the instance is added to the timestamp list corresponding to the MotifBag, thus updating the frequency of the MotifBag. Thus, the MBIndex is of the form \( \{MBag \rightarrow \{ts(bag1), ts(bag2), \ldots\}\} \), where \( MBag \) is a uniquely identified MotifBag with instances \( bag1, bag2 \) etc.

In the interest of efficiency, the frequency index MBIndex is implemented using a hashtable, where the MotifBags serve as keys, whereas the list of each MotifBag’s instances along with timestamp information is stored as the value corresponding to the appropriate MotifBag key. The implementation of the MBIndex has the following features:

(i) Using a hashtable enables constant time addition and updation of MotifBag entries

(ii) MBIndex records co-occurrence associations in an incremental manner such that any subset of the cooccurrence is also recognized. Even after the creation of a MotifBag \( \{M_{ix}, M_{jy}, M_{kz}\} \), the motifBags \( \{M_{ix}, M_{jy}\}, \{M_{jy}, M_{kz}\} \) and \( \{M_{ix}, M_{kz}\} \) are not discarded and continue to be tracked. This provision makes the approach more robust in identifying the individual importance of each lower order MDM and their contributions to any higher order MDM they might constitute.

(iii) Since the MBIndex is updated on-demand, the number of updates is kept to a minimum.

(iv) MBIndex can be queried on in real time to retrieve the status of a specific MotifBag at any instant of time, such as its frequency, number of members, timestamps of instances etc.

### 4.1.3 Identifying UMDMs

Conceptually, an Unordered Multidimensional Motif is a recurring group of co-occurring motif instances. As explained in Section 3, this concept translates to a UMDM being a MotifBag that has two or more instances. Thus, whenever the frequency of a MotifBag as recorded in the MBIndex exceeds the desired threshold, the algorithm flags the MotifBag as UMDM, with the instances of the MotifBag becoming instances of the UMDM.

### 4.2 Temporally ordered Multidimensional Motif Discovery

We propose an alternative model for representing MDMs based on tracking recurring groups of temporal relations among co-occurring motif instances.

#### 4.2.1 Representing temporal relations among motifs

We propose a tempOrd tuple format to store pairwise temporal relations between motif instances. Each tempOrd tuple contains three pieces of information – \([a \rightarrow p, b \rightarrow q, R]\)
\[ a - p \text{ and } b - q \text{ designate the participating motif instances and } R \text{ is the temporal relation as given by Allen's interval algebra [2]. The motif indicated by } a - p \text{ is the } p\text{th motif in the } a\text{th dimension while the motif indicated by } b - q \text{ is the } q\text{th motif in the } b\text{th dimension, where } p \text{ and } q \text{ are orderings with respect to the particular combination under consideration. For example, if a particular combination consisted of a single motif instance in the first dimension and two motif instances in the second dimension, then the designations for the motif instances would be 1-1, 2-1 and 2-2. There are two main advantages of this representation format:}

(i) Since only the local ordering and temporal relations are retained, and not the actual motif identifiers, the temporal ordering information becomes generic and can be compared with any other combination of motifs for similarity in temporal ordering. (ii) Use of Allen’s interval algebra reduces redundancy by not requiring the inverse temporal relations to be recorded. For instance, if \([a - p, b - q, R]\) is a tuple formed, then Allen’s formulation eliminates the need for the inverse tuple \([b - q, a - p, R']\) to be stored (if \(R\) is Before, then \(R'\) would be After).

4.2.2 Indexing Groups of Temporal relations

From each motif neighborhood, sets of all possible combinations of neighboring motif instances are computed. For each of these combinations the pairwise temporal relations are captured and encoded into tempOrd tuples. Thus, each combination of neighbor motifs yields a set of tempOrd tuples, which we refer to as a MotifOrdering (see Algorithm 1).

Note: MotifOrderings can be formed by extracting temporal relations among motif instances that are members of a MotifBag (see Section 4.1.2), thus building upon existing functionality and making our method more efficient.

The process of indexing MotifOrderings is the exact same as that in case of MotifBags. We simply use a different frequency index, called the MOIndex, to track and record each distinct MotifOrdering and its frequency (number of instances) (see Figure 8). Each distinctly identified MotifOrdering is recorded in the index, with the timestamp information of its current instance. Over time, when another instance of an existing MotifOrdering is encountered, the timestamp information of the instance is added to list corresponding to the MotifOrdering, thus updating the frequency of the MotifOrdering.

As in case of MBIndex, the MOIndex is also implemented using a hashtable to benefit from the advantages as described before.

4.2.3 Identification of TMDMs

By definition, a Temporally Ordered Multidimensional Motif is a MotifOrdering that has two or more instances. Thus, whenever the frequency of a MotifOrdering as recorded in the MOIndex exceeds the desired threshold which is two, the algorithm flags the MotifOrdering as a TMDM, with the instances of the motifBag becoming instances of the TMDM.

Constrained TMDMs require constituent motifs as well as temporal relations to be identical. This can be achieved by cross-referencing entries in MBIndex and MOIndex and tracking frequencies of such joint entries.

Algorithm 1 provides the pseudocode for the procedure to extract MotifBags and MotifOrderings to be executed at the end of each timestamp.

Algorithm 1 \(\text{GenerateMBMO}(\cdot)\)

Input: Timestamp \(ts\), Neighborhood distance \(r\)
Output: set \(MB\) of MotifBags, set \(MO\) of MotifOrderings
\[ MB \leftarrow \{\}, \quad MO \leftarrow \{\}\] for valid timestamps \(p \leftarrow ts\) to \(ts + r\) do
\[
\text{currentNeighborhood} \leftarrow p - r \text{ to } p
\]
\[
\text{currentNeighborSet} \leftarrow \{\}
\]
for all timestamp \(s\) in \(\text{currentNeighborhood}\) do
\[
\text{motifSet} \leftarrow \text{tsIndex.getMotifsAt}(s)
\]
\[
\text{currentNeighborSet} \leftarrow \text{currentNeighborSet} \cup \text{motifSet}
\]
end for
\[
\text{pwSet} \leftarrow \text{generatePowerSet(currentNeighborSet)}
\]
\[
\text{pwSet} \leftarrow \text{pwSet} \leftarrow \text{nullset}()\]
for all set \(bag\) in \(\text{pwSet}\) do
if \(bag\) is clean (No overlapping motifs) then
\[
MB \leftarrow MB \cup bag
\]
\[
\text{mo} \leftarrow \{\}
\]
for all pair of motif instances \(m_a, m_b\) in \(bag\) do
Form \(\text{tempOrd}\) tuple using relation \(R\) between \(m_a\) and \(m_b\)
\[
\text{mo} \leftarrow m_a \cup \text{tempOrd}
\]
end for
\[
MO \leftarrow MO \cup mo\]
end if
end for
end for

5. IMPLEMENTATION AND EXPERIMENTS

The proposed approach addresses two important issues in MDM discovery – detection of variable length motifs and compatibility with online/real-time/streaming data analysis. The various aspects of our implementation are as follows:

(i) Discretization of individual dimensions using SAX
– For efficient processing, the sensor data streams are transformed to a string of discrete symbols using the SAX symbolic representation (See Appendix Section B).

Figure 8: MOIndex: Hashtable recording the MotifOrderings as keys and the corresponding time stamp lists.

\[\begin{align*}
\text{[1-1, 2-1, Overlaps]} & \quad 2 & \quad 26 \\
\text{[1-1, 3-1, Overlaps]} & \quad 2 & \quad 26 \\
\text{[1-1, 2-1, Overlaps]} & \quad 2 & \quad 26 \\
\text{[1-1, 3-1, Overlaps]} & \quad 2 & \quad 26 \\
\text{[4-1, 2-1, Overlaps]} & \quad 2 & \quad 26 \\
\text{[2-1, 1-1, Overlaps]} & \quad 22 & \\
\text{[2-1, 3-1, Overlaps]} & \quad 22 & \\
\text{[2-1, 1-1, Overlaps]} & \quad 22 & \\
\text{[1-1, 1-1, Overlaps]} & \quad 29 & \\
\end{align*}\]
can be referred to as symbol steps. Moves across all dimensions one symbol at a time, in what word) at a time. This can be visualized as a window that processes each of the strings in parallel, one symbol (SAX and retrieval of motif sets, and identification of MDMs.

An efficient hashing based record to enable speedy update for representing motif co-occurrence information, as well as the index of the symbol step in which it occurs. The greatest merits of using Sequitur for motif discovery are its efficiency in finding naturally occurring repetitive patterns of variable lengths, and its identification of a hierarchy among the motifs found.

(iii) Incorporating Flexibility for Spatial Variance

In a previous work [4], we proposed an enhancement to Sequitur for noise tolerant motif discovery, which relaxes the match threshold to allow two dissimilar strings to be matched using a partial match criteria. The application of this enhancement has been shown to effectively retrieve instances of a motif which are similar to each other except for local variations due to noise and distortion. This partial match criteria has been incorporated in the proposed approach to enable better identification of genuine MotifBag instances, and subsequently, MDMs.

5.1 Concurrent Processing of Multidimensional Time Series Data

The proposed approach introduces two alternative models for representing motif co-occurrence information, as well as an efficient hashing based record to enable speedy update and retrieval of motif sets, and identification of MDMs.

The enhanced rule mining approach derived from Sequitur processes each of the strings in parallel, one symbol (SAX word) at a time. This can be visualized as a window that moves across all dimensions one symbol at a time, in what can be referred to as symbol steps. The timestamp of a symbol is the index of the symbol step in which it occurs. The timestamp of a rule instance is the timestamp of the first component symbol of the rule. At each symbol step, the window encounters the next symbol in each dimension. The bigrams (constituted by the previous and current symbols) from each dimension are then recorded (added/updated). Whenever a motif instance is identified in any of the dimensions, it is indexed in the inverted timestamp index (see Section 4.1.1) and the MBindex (see Section 4.1.2).

5.2 Experiments

In order to evaluate the performance of the motif discovery techniques presented in the previous sections, we conduct a set of experiments involving – (i) unsupervised motif discovery to retrieve all MDMs identified; and (ii) content based query resolution to identify and retrieve approximate matches found in multidimensional time series with respect to a given multidimensional query.

5.2.1 Analysis A: Synthetic Data (Planted Motifs)

For this part of our evaluation of the proposed method, we artificially created a synthetic multidimensional time series with random walk data, and embedded or “planted” a number of motifs in every dimension. The resulting synthetic dataset provides examples for unordered MDMs. Each planted motif has instances that are approximately similar to each other. We then executed unguided MDM discovery on this dataset using the proposed approach. A sliding window of 50 samples with a shift of 10 samples was used to scan the individual dimensions concurrently. Figure 9 illustrates the synthetic dataset used along with the planted motifs, and also provides some examples of the results obtained from motif discovery. As can be seen from the figure, the approach is able to identify every unique MDM resulting from the co-occurrences of the planted motifs, with all of their instances. The recall of the method on this dataset is thus 100%. However, the unguided method also retrieves a number of MDMs using the interval data between planted motifs, which are not relevant and can be discarded. Precision therefore is not a suitable performance measure for unsupervised motif discovery, unlike classification.

There are several noteworthy observations in these results – (i)the MDMs discovered span different number of dimensions, and are also hierarchically recorded. For instance, even though the MDM in Figure 9(d) is a formed by a subset of the motifs forming the MDM in Figure 9(b), maintaining records for hierarchical MDMs proves beneficial in finding instances that are not common to both MDMs; (ii) identified MDMs well exceed the length of the basis sliding window, while individual unidimensional motifs forming an MDM need not be of the same length either; (iii) given an intuitive idea of the expected patterns, the approach can use

Figure 9: Results of Unordered Multidimensional Motif Discovery on Synthetic Dataset with Planted Motifs. (a) The synthetic dataset with planted motif types indicated by the dotted boxes, with length of 60, 60, 90 and 50 samples, respectively. (b), (c) and (d) The three UMDMs in the dataset with their instances (highlighted in red, better viewable on electronic version). Please note that variable length motifs are captured, with motif lengths sometimes exceeding the base sliding window length, such as the one in (b).

Figure 10: Examples of (a) Unconstrained TMDM and (b) Constrained TMDM (highlighted in red, better viewable on electronic version).
a suitable basis sliding window to track and “grow” patterns of lengths longer than the sliding window length.

Figure 10(a) presents an example of unconstrained TMDMs identified from the synthetic data. In the given example, the TMDM has three instances spanning two dimensions. We can see that two instances have common member motifs while the third has different member motifs. However, in each of these instances, the motif in the first dimension has a “During” relationship with the motif in the second dimension. In contrast, under the constrained TMDM, the instance with different member motifs is not indexed, as shown in Figure 10(b).

5.2.2 Analysis B: Real Data (Body Sensor Data)

Electromagnetic Articulography – As a step towards validating the proposed approach and demonstrating its utility in identifying similar patterns in multidimensional time series data generated biomedical sensors, we tested the approach on an Electromagnetic Articulography (EMA) dataset collected for research in silent speech interfaces [32, 31]. The dataset used in this experiment is the articulatory data in sequences of 25 words repeated by a single speaker. All sequences were manually annotated with the onset and offset of those words. The electromagnetic articulograph (EMA) (Carstens Inc., Germany) [34] records tongue movements by establishing a calibrated electromagnetic field in a cube that induces electric current into tiny sensor coils that are attached to the surface of the articulators. The 3-D movements of the tiny sensor coils were tracked and recorded when a subject was talking. The spatial accuracy of AG500 is 0.5 mm [34]. Four sensors - T1 (Tongue Tip), T2 (Tongue Blade), T3 (Tongue Body Front) and T4 (Tongue Body Back) - were attached approximately 10 mm from each other at the midline of the tongue. Two sensors - T1 (Tongue Tip) and T4 (Tongue Body Back) - were attached approximately 40 mm from each other at the midline of the tongue [30]. A low-pass filter (20 Hz) was applied to remove noise. Finally, only the y and z dimensions of the tongue sensors were used for analysis because the lateral movement (along the x dimension) of the tongue is not significant in normal speech production [30].

The objective of this experiment was to see if the proposed multidimensional motif discovery approach could identify repeating patterns across dimensions, and to evaluate the consistency of the MDMs identified with the documented word annotations. Figure 8 illustrates a few examples of the MDMs identified by the proposed approach in the motion trajectory of the tongue for the same sequence of 25 words repeated twice. At first glance, it is evident that the MDMs identified pertain to sections that represent trajectories for the same sequence of words. Such MDMs can be exhibited in two or more dimensions as shown.

It is possible that in addition to relevant sections, a few false positives may also be returned, as shown in Figure. However, while these superfuous motif instances may not pertain to the correct sequence, their identification indicates that the motion trajectories of a different sequence of words may be similar. In case of silent speech interfaces, a high level of accuracy is required in classifying articulatory data, in order to differentiate between words pronounced. While MDMs may not be relied upon for classification purposes, they can serve as indicators of sets of words that have similar motion trajectories. The identification of such sets can assist in fine tuning a classifier, or making the differentiation stricter for those mix of words.

Physical Activity Monitoring – The proposed approach was tested on human physical activity data. The dataset used is from the Physical Activity Monitoring for Aging People (PAMAP) [27]. The entire dataset contains time series data which are collected from multiple inertial sensors (e.g., accelerometers, gyroscopes, magnetometers, etc.) mounted on three different locations on the human body – hand, chest and shoe (foot). We executed MDM discovery on the three dimensional accelerometer information from the inertial sensor attached to the subject’s shoe for the activity of “Nordic walk” followed by “Running”. The sampling rate is 100 Hz.

Figure 13 shows a sample of the MDMs identified using a sliding window of 50 samples with a shift of 2 samples. From the data used, the approach showed a recall of 94% for “Nordic Walking” and and 57% for “Running”. The precision usually is low since the unsupervised motif discovery identifies many similar segments of the data along with the motifs of interest. This demonstrates that, while MDMs are able to capture most of the “Nordic walk” iterations, the “Running” activity poses a harder challenge owing to its erratic and inconsistent nature. It could also be due to the fact that the data was collected from aged individuals who exhibit a lack of stability that is usually observed in younger subjects. Nevertheless, this example clearly shows that while MDMs can be trusted to highlight repeating patterns based on similarity, they are still not reliable enough for classification purposes.

Motion Capture and Muscle Activity – In order to demonstrate the mining of TMDMs in real datasets, we use Mocap-EMG data collected for a “Raise Arm” activity [25]. The example uses the sensor data generated from motion capture sensors on the wrist joint (3 dimensions) captured synchronously with corresponding muscle activity data for the tibialis anterior muscle (1 dimension), captured at a sampling rate of 120 Hz. Examples of TMDMs from a section of the Mocap-EMG data for 10 repetitions of the activity using a sliding window and neighborhood distance 500 samples each are shown in Figure 14. Figure 14(a) shows a TMDM, where the onset times of muscle activity shown by EMG data and the wrist joint displacement are the same. On the other hand, Figure 14(b) shows a TMDM, where the muscle activity shown by EMG data has an onset time be-
Figure 12: Multidimensional motifs identified in Articulatory data (Sampling rate 200 Hz) for a sequence of 25 words repeated twice. In each of the above cases, the instances of the detected MDM correctly correspond to repetitions of the same word sequence. (a) ‘mute, them, fate, tang’ (b) ‘rap, vast, dab, ways’ (c) ‘start, moose, trash, awe’ (d) ‘log, nut, frog, gloss’. In (d), a spurious MDM is also observed (encircled), which is identified due to similarities in the motion trajectories.

Figure 13: Example from the PAMAP dataset for a sequence of motions of “Nordic Walking” followed by “Running”. UMDMs identified and their instances (highlighted in red, better viewable on electronic version).

Figure 14: Examples of TMDMs from a section of the Mocap-EMG data for 10 repetitions of the “Raise Arm” activity (sampling rate 120 Hz) – (a) onset times for wrist join displacement and tibialis anterior muscle activity are the same; (b) onset time of tibialis anterior muscle precedes that of wrist joint displacement. (TMDMs are highlighted in red, better viewable on electronic version).
fore the onset time of the wrist joint displacement. Please note that in both cases, the constituent motifs are not identical across instances, while the temporal relations among the motifs are. This shows the utility of the proposed approach in capturing temporal relations among unidimensional motifs and repetitions of such relationship groups.

5.2.3 Analysis C: Runtime Performance and Scalability

In order to explore the scalability of the proposed method, we studied the variation of the execution times of the method with increasing number of dimensions as well as increasing lengths of time series.

The first example is taken from the PhysioNet/Computers in Cardiology Challenge 2009 Test Set B dataset [8], that provides the body sensor data information for patients who experienced acute hypotensive episodes (AHE), including heart rate, systolic blood pressure, diastolic blood pressure, mean arterial pressure, and respiratory rate. The section of data used in our analysis has 6 dimensions of synchronous time series data having 7500 samples each, with a sampling rate of 125 Hz. The results obtained using a sliding window of 100 samples is shown in Figure 15 (a).

The second example is from the PAMAP dataset, introduced in Section 5.2.2. The section of data used has 6 dimensions of synchronous time series data having 30000 samples each. The results obtained using a sliding window of 100 samples is shown in Figure 15 (b).

From these examples, it can be seen that from time series data captured over 200 seconds, MDMs can be mined within 5 seconds. This trade-off is very favorable for real-time applications, since the concurrent mining of the unidimensional and multidimensional motifs can be achieved within a very small fraction of the actual duration of recording. It should be noted that this runtime would vary with the concentration of unidimensional motifs mined, since a larger number of motifs results in a “pattern explosion”, thus requiring much more time to generate and index pattern combinations. With increasing dimensions, execution times may be handled better using distributed processing or multi-threading (not within the scope of this study).

5.2.4 Content based Querying on Motifs

In a query based approach, a multidimensional query is submitted to the motif discovery routine, which then should retrieve all instances of the given query. Consider a multidimensional query \( Q = \{ q_i \} \), on a \( d \)-dimensional time series, where \( i \in \{ 1, 2, 3, ..., d \} \). The objective would be to search for instances of the unidimensional query \( q_i \) in the \( i \)th dimension of the multidimensional time series, and then to identify non-trivial co-occurrences of these unidimensional query matches that form valid instances of the query \( Q \).

An online query resolution task can be used to find MDMs matching the query in time series streams. This approach would be a truncated version of the proposed approach. In the course of the unidimensional motif discovery routine, the inverted timestamp index simply records the motif instances from each dimension \( i \) that are identified as valid matches to the corresponding unidimensional query \( q_i \). Thereafter, the existing MotifBag creation and updating can be optimized by not considering any neighbor combinations set that has a cardinality not equal to the cardinality of the query \( Q \).

Thus, the only MotifBags formed are combinations of co-occurrences instances of the unidimensional queries.

On the other hand, our approach can also support multidimensional queries in an offline scenario, where all MDMs identified are already recorded and available. In this scenario, it might be enough to search the identified set of MDMs to find a match with the given multidimensional query \( Q \), which would spare the effort of processing the actual data again to find the individual unidimensional queries \( q_i \).

Figure 16 demonstrates the query resolution functionality of the proposed approach, using two different queries on the synthetic dataset used in Analysis A (Section 5.2.1). The similarity between the queries and the identified motifs are computed using Dynamic Time Warping.

6. CONCLUSION

Monitoring patient conformance during personalized rehabilitative activities requires tracking repeating patterns in multidimensional time series data from body sensor networks with heterogeneous sensors. Multidimensional Motifs (MDMs) provide an intuitive and effective solution for this purpose without requiring prior evidence unlike classification techniques. This paper presents an enhanced and efficient method for MDM discovery, with ability to concurrently process multidimensional time series data, and identify and index MDMs and their instances using an efficient hashing based indexing approach. The unordered and temporally ordered models proposed for MDMs are beneficial in tracking variable length motif co-occurrences and temporal orderings among motifs across multiple dimensions. The approach is also found to be well suited for content based querying applications. The scalability and efficiency of the approach holds promise for online analysis and monitoring in healthcare applications. The utility of the proposed approach in (a) tracking repetitions during therapy sessions; (b) finding naturally occurring patterns; and (c) query resolution makes it a valuable assistive tool in analysis and interpretation of body sensor data characteristics.

Our future endeavors in this direction would involve extending the MDM mining approach to associated applications such as anomaly detection, association rule mining be-
tween MDMs, and patient grouping based on similarities based on characteristics extracted from their data in the interest of patient profiling and personalized treatment and rehabilitation.

7. REFERENCES


APPENDIX

A. UNIDIMENSIONAL MOTIF DISCOVERY

The proposed multidimensional motif discovery approach assumes the availability of a unidimensional motif discovery technique, that identifies and retrieves a set of time series motifs from each dimension of the multidimensional time series. Given a d-dimensional time series \( MDT = \{T_1, T_2, T_3, ..., T_d\} \), the motif discovery algorithm would return each unique unidimensional motif \( M_{ij} \) and its \( k \) instances \( \{m_{ijk} \} \), where unidimensional motifs instances \( \{m_{ijk} \} \) such that \( M_{ij}^{t} \) is the \( t \)th motif occurring in the \( i \)th dimension \((t = 1, 2, ..., d)\) \((j = 1, 2, ..., n_i)\), where \( d \) is the number of dimensions in the time series data, \( n_i \) is the number of motifs identified in the \( i \)th dimension. Each motif instances \( m_{ijk} \) has a timestamp \( ts(m_{ijk}) \) associated with it.

In the interest of efficiency, the unidimensional motif discovery on the individual dimensions is parallelized. The incremental steps in which time series subsequences are chosen for evaluation is standardized in terms of timestamps. This timestamp value can be the index of the sliding window that is used to extract subsequences for evaluation. In motif discovery methods where the individual time series are discretized into a string of symbols, a constant sliding window would be used across all dimensions, with the sequence number of the symbols as timestamps.

B. SYMBOLIC AGGREGATE APPROXIMATION (SAX)

SAX (Symbolic Aggregate approximation) representation [13] is a widely popular symbolic representation for time series data that provides (i) efficient dimensionality reduction while retaining essential features; and (ii) lower bounding of the distance measure. Its basic principle is that a time series \( C \) of length \( n \) can be represented in a \( w \)-dimensional space by a vector \( \bar{C} = c_1, c_2, ..., c_w \), where the \( w \)th element of \( \bar{C} \) is given by

\[
\bar{c}_j = \frac{1}{n} \sum_{i=1}^{n} \alpha (c_i - \mu_j, \sigma_j) + c_j.
\]

In SAX, each time subsequence is \( z \)-normalized (mean = 0 and \( SD = 1 \)), and split into \( w \) equal segments. For each segment, the mean is calculated and a symbol is assigned based on a set of breakpoints that divide the distribution space into \( \alpha \) equiprobable regions, where \( \alpha \) is the alphabet size. Thus, each time subsequence is converted into a string or word of length \( w \), formed by symbols from an alphabet of size \( \alpha \). Figure 17 shows the usage of SAX to represent time series subsequences as strings of symbols, with \( w = 4 \) and \( \alpha = 31,2,3 \). For the definition of breakpoints, please refer to [13].
Figure 17: Segmentation of original time series into subsequences using overlapping sliding windows; (b) Chart showing sample of breakpoints that divide the distribution space into equiprobable regions for each alphabet size $\alpha$; (c) Derivation of SAX representation for each subsequence.

Figure 18: Motif mining from a SAX symbol string using Sequitur.

C. SEQUITUR STRING COMPRESSION ALGORITHM

Sequitur string compression algorithm [19] uses a context free grammar based rule building approach to index repeating bigrams in a string of symbols. Every bigram (two consecutive symbols) in the string of symbols is recorded. Whenever a bigram is repeated, all occurrences of the bigram are replaced by a non-terminal symbol. If a rule for such a substitution does not already exist, a new rule is created and added to the existing rule base (see Figure 18). Also, rules that are not used more than once are discarded, emphasizing the meaningfulness of the rules that are retained.

Since Sequitur processes strings one symbol at a time from beginning to end, it has been shown to be a useful online technique for finding repetitive patterns. A pattern or motif corresponds to any of the different rules in the rule base, since every rule corresponds to a sequence that occurs at least twice in the symbol string. Sequitur is therefore used to mine rules and identify motifs from a given string of SAX words, where each SAX word is a symbol. The approach provides scope for identifying truly variable length motifs, as is obvious from the possibility of a non terminal substitution for a bigram of non-terminals, leading to identification of motifs that can be a few SAX words in length.